Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог» (2013)

Выпуск 12

В двух томах

Том 1. Основная программа конференции

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2013)

Issue 12

Volume 1 of 2. Main conference program

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает
искреннюю благодарность Российскому фонду фундаментальных
исследований за финансовую поддержку,
грант № 13-06-06047

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19): В 2 т.

Т. 1: Основная программа конференции. — М.: Изд-во РГГУ, 2013.

Сборник включает 84 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2013», 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

12-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит материалы 19-й Международной конференции «Диалог». В результате работы 54 рецензентов для сборника было отобрано 84 доклада, охватывающих различные направления исследований в области компьютерного моделирования и анализа естественного языка. В настоящем сборнике представлены:

- Лингвистическая семантика и семантический анализ;
- Формальные модели языка и их применение;
- Теоретическая и компьютерная лексикография;
- Методы оценки (evaluation) систем анализа текстов и машинного перевода;
- Корпусная лингвистика. Создание, применение, оценка корпусов;
- Новые лингвистические ресурсы;
- Интернет как лингвистический ресурс. Лингвистические технологии в Интернете;
- Онтологии. Извлечение знаний из текстов;
- Компьютерный анализ документов: реферирование, классификация, поиск;
- Автоматический анализ тональности текстов;
- Машинный перевод;
- Модели общения. Коммуникация, диалог и речевой акт;
- Анализ и синтез речи.

«Диалог» является ведущей российской конференцией по компьютерной лингвистике и, видимо, единственным в мире форумом, посвященным прежде всего проблемам компьютерного анализа русского языка. Принципиальной особенностью конференции, ее основополагающей традицией является особое внимание к технологиям автоматического анализа текста, основанным на лингвистических моделях. Именно этим объясняется и состав участников, и программа конференции, в которой соседствуют теоретические и прикладные исследования. В «Диалоге» представлены также и работы, сделанные в рамках статистических подходов, что позволяет, в частности, сравнивать полученные результаты.

«Диалог» является не только местом обмена опытом и представления новых достижений. Он является также и форумом для разработки и апробирования методик верификации и оценки как результатов лингвистических исследований, так и эффективности работы различных видов систем анализа текстов на русском языке. Целью этой работы являются единые для авторов и рецензентов принципы доказательства и оценки объективности, эффективности и научной новизны предлагаемых решений и методики проведения сравнительного тестирования, на которых могли бы основываться такие оценки.

Схожие проблемы решает в области информационного поиска семинар РОМИП: не случайно, что вот уже второй год «Диалог» и РОМИП проводят совместные дорожки тестирования, результаты участников которых докладываются на «Диалоге» и публикуются в этом сборнике.

В этом году проводилось два соревнования: по анализу тональности (продолжение тестирования 2012 года) и по оценке систем Машинного Перевода (для англо-русской языковой пары).

Особая роль русского языка обусловливает наличие в программе работ по адаптации к нему известных алгоритмов и методов, разработанных для других языков. Доказанные положительные или отрицательные результаты такого применения рассматриваются рецензентами как новые.

За год, прошедший после последней конференции, «Диалог» понес невосполнимую потерю: ушел из жизни выдающийся лингвист и один из отцов-основателей «Диалога» Александр Евгеньевич Кибрик. Трудно переоценить его роль в создании особой концепции и самой атмосферы конференции, которая сохраняется вот уже почти 40 лет, начиная с первых семинаров середины 70-х годов, из которых и вырос «Диалог». Основными чертами этой концепции всегда оставались широта взгляда, междисциплинарность, сочетание конструктивности и теоретической значимости обсуждаемых проблем. В этом году А. Е. Кибрику посвящено специальное заседание, материалы которого также вошли в сборник.

Несмотря на традиционную широту тематики докладов одного года, они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет.

Программный комитет «Диалога»
Редколлегия ежегодника «Компьютерная лингвистика
и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится под патронажем Российского Фонда Фундаментальных Исследований при организационной поддержке компании ABBYY.

Учредителями конференции являются:

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

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

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

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Байтин Алексей Владимирович	Компания Yandex
Беликов Владимир Иванович	Институт русского языка им. В.В. Виноградова РАН
Браславский Павел Исаакович	Kontur Labs; Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Ляшевская Ольга Николаевна	Universitet i Tromsø
Сердюков Павел Викторович	Компания Yandex
Соколова Елена Григорьевна	РосНИИ искусственного интеллекта
Толдова Светлана Юрьевна	Филологический факультет МГУ
Шаров Сергей Александрович	University of Leeds, UK

Секретариат

Белкина Александра Андреевна, *секретарь оргкомитета*	Компания ABBYY
Мытникова Татьяна Александровна, *координатор*	Компания ABBYY

Рецензенты

Августинова Таня
Азарова Ирина Владимировна
Апресян Валентина Юрьевна
Байтин Алексей Владимирович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданов Алексей Владимирович
Богданова Наталья Викторовна
Богуславский Игорь Михайлович
Бонч-Осмоловская Анастасия Александровна
Браславский Павел Исаакович
Гельбух Александр Феликсович
Горностай Татьяна Александровна
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрынин Владимир Юрьевич
Дружкин Константин Юрьевич
Захаров Леонид Михайлович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кибрик Андрей Александрович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Кэрролл Джон
Лахути Делир Гасемович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Маккарти Диана
Падучева Елена Викторовна
Пазельская Анна Германовна
Подлесская Вера Исааковна
Савельев Василий Евгеньевич
Селегей Владимир Павлович
Семенова-Флюр Вера Эммануиловна
Сердюков Павел Викторович
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Старостин Анатолий Сергеевич
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Федорова Ольга Викторовна
Филиппова Екатерина Александровна
Хорошевский Владимир Федорович
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Юдина Мария Владимировна
Янко Татьяна Евгеньевна

Contents*

Раздел I.
Основная программа конференции

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

Раздел I.

Основная программа конференции

ВЛИЯНИЕ СИНТАКСИЧЕСКОЙ СТРУКТУРЫ НА ИЗВЛЕЧЕНИЕ КОЛЛОКАЦИЙ-СУЩЕСТВИТЕЛЬНЫХ ПРИ ГЛАГОЛАХ

Акинина Ю. С. (jakinina@hse.ru),
Кузнецов И. О. (iokuznetsov@hse.ru)
Центр семантических технологий НИУ ВШЭ, Москва, Россия

Толдова С. Ю. (toldova@yandex.ru)
МГУ имени М. В. Ломоносова; Центр семантических технологий НИУ ВШЭ, Москва, Россия

Ключевые слова: коллокации, глагольная сочетаемость, автоматический синтаксический анализ, корпусные методы

THE IMPACT OF SYNTACTIC STRUCTURE ON VERB-NOUN COLLOCATION EXTRACTION

Akinina Y. S. (jakinina@hse.ru),
Kuznetsov I. O. (iokuznetsov@hse.ru)
The Center for Semantic Technologies, Higher School of Economics, Moscow, Russia

Toldova S. Y. (toldova@yandex.ru)
Lomonosov Moscow State University, The Center for Semantic Technologies, Higher School of Economics, Moscow, Russia

Automatic verb-noun collocation extraction is an important natural language processing task. The results obtained in this area of research can be used in a variety of applications including language modeling, thesaurus building, semantic role labeling, and machine translation. Our paper describes an experiment aimed at comparing the verb-noun collocation lists extracted from a large corpus using a raw word order-based and a syntax-based approach. The hypothesis was that the latter method would result in less noisy and more exhaustive collocation sets. The experiment has shown that the collocation sets obtained using the two methods have a surprisingly low degree of correspondence. Moreover, the collocate lists extracted by means of the window-based method are often more complete than the ones obtained by means of the syntax-based algorithm, despite

its ability to filter out adjacent collocates and reach the distant ones. In order to interpret these differences, we provide a qualitative analysis of some common mismatch cases.

Keywords: collocations, verb compatibility, parsing, corpus methods

1. Introduction

The identification of semantically related words is an actual NLP task. Most of the special lexicographic research in that area is focused on identifying synonymy, multiword expressions or hyperonymy-hyponymy relations. One special case of lexical relation extraction is modeling verb semantics using the information about the verb-noun compatibility. The information obtained from the verb-noun distribution model is then used in a wide range of NLP tasks such as semantic-role labeling ([Gildea, Jurafsky 2002]), word sense disambiguation ([Kustova, Toldova 2009], fact extraction, thesaurus building ([Lin 98], [Pado, Lapata 2007]), machine translation ([Orliac, Dillinger 2003]) and others. Modeling lexical relations between words can be done automatically by methods of collocation extraction.

Generally, the collocation extraction is a two-step process which includes candidate extraction and candidate ranking. A variety of methods are proposed for executing each of these steps. In particular, the candidate selection can be performed using either linear or syntactic text representation. In the first case, the words in the source texts are treated as consequent units, while the second model relies on syntactic representation in which units are connected non-linearly. Being more complex from a technical point of view, the latter method should result in noise reduction and higher recall due to the additional syntactic filtering. Our aim is to compare the collocation sets obtained by using these two candidate extraction methods in the same setting, and to analyze the differences between the results.

2. Background

2.1. Notion of collocation

The definition of "collocation" differs across linguistic traditions. From the theoretical point of view, collocation can be considered to be some kind of a "fixed phrase", in a scale where fixed phrases are opposed to "free phrases" (see [Khokhlova 2008] for a review). However, when it gets to practice, retrieving a particular theoretically predetermined class of phrases can become problematic. A more practical definition of collocation within the corpus linguistics paradigm will be "two or more lexical units that co-occur more often than would be expected by chance" [Manning, Schütze 1999]. Statistically-based methods of collocation extraction ranks word pairs according to a certain measure of association, which evaluates

the chance of their occurrence. As a consequence a high rank can be obtained not only by fixed expressions such as *сломать голову* “rack one’/s brains”, but by free word combinations such as *сломать руку* “to break smb’s hand”. While the former is an idiom listed in a dictionary the meaning of the latter is compositional. However the noun *рука* in the latter is a “typical” argument for the verb сломать. This type of word-combinations should be also taken into consideration for verb semantics modeling. Thus we use the term collocation for both types of word combinations discussed above.

2.2. Collocation candidate selection methods

Choosing the method of compiling lists of possible collocates is a crucial step in collocation extraction. The variety of methods can be roughly divided in two groups. The methods from the first group, which we refer to as **window-based methods** (e. g. [Church, Hanks 1990], [Breidt 1993], [Todirascu et al. 2008], [Todirascu, Gledhill 2008]), rely on linear word order model, in which the collocation candidates are extracted from a fixed-size window, and the distances correspond to the raw distance between two (or more) words as presented in the source document. Analyzing adjacent bigrams can be regarded as a particular case with window size of 1. Applying POS or pattern filters can be implemented as an approximation to syntactic structures ([Klyshinskij et al. 2010], [Todirascu et al. 2008]). The second group can be referred to as **syntax-based methods** ([Lin 1998], [Kilgariff, Tugwell 2002], [Khokhlova 2009]). The methods from this group rely on syntactic structure instead of using the linear representation. The candidate list is generated based on syntactic relations. Both candidate selection methods have advantages and shortcomings. Window-based methods tend to extract additional noisy data and ignore the long-distance syntactical links ([Kilgariff, Tugwell 2002]), but are easy to implement. Using the syntax-based methods makes it possible to filter out spurious examples in the nearest context and also access the distant collocates, which are invisible in the window-based linear representation ([Kilgariff, Tugwell 2002]). The increase in precision comes at the cost of carefully describing all the syntactical constructions, in which two collocation candidates can occur.

2.3. Verb-noun collocation extraction

When it comes to verb-noun (V-N) collocations, the researchers’ aim is usually to extract some specific, theoretically predetermined types of constructions ([Breidt 1993], [Todirascu et al. 2008], [Todirascu, Gledhill 2008]). There is a particular interest among researchers for the task of V-N collocation extraction. The majority of the works are based on the combination of morphological pattern-based and statistically based methods (see [Todirascu et al. 2008] for French and Romanian, [Breidt 1993] for German, [Todirascu, Gledhill 2008] for English and Romanian, [Todirascu et al. 2008] for German). For this method a high level of noise is reported (c. f. [Todirascu, Gledhill 2008]). On the one hand, a certain amount of totally irrelevant V-N pairs were extracted by means of the method.

On the other hand, the authors were looking for some particular types of collocations. For instance, subject+predicate collocations or combinations with circumstantial adjunct ([Todirascu et al. 2008], [Todirascu, Gledhill 2008]) were out of the author's interest. Therefore, the conclusion was that the syntactic information was required to detect such combinations ([Todirascu et al. 2008]). The authors of [Breidt 1993] also claim that syntactic parsing is necessary to distinguish subject-verb from object-verb combinations. Indeed, applying syntactic filters over a parsed corpus in English allows getting statistical information from Subject-Verb-Object triples to accurately answer the questions about typical arguments, e. g. "*What can you drink?*" [Church, Hanks 1990].

There are some works that focus on collocation extraction in Russian, but so far the main method has consisted of applying/comparing different word association measures over the lists of adjacent word units (e. g. [Khokhlova 2008], [Jagunova, Pivovarova 2010]). The experience of using syntax-based Word Sketches methodology presented in [Khokhlova 2009] claims viability of this method for collocation search in Russian, but does not analyze Verb-Noun collocations in particular. The work presented in [Klyshinskij et al. 2010] concerns extracting verb lexical compatibility information in general (that is, not just fixed phrases, but typical free phrases as well), but it relies on a huge text corpus to bypass the syntactic parsing using a number of assumptions (such as "the next noun phrase after a single verb most probably depends on it"). The authors of [Klyshinskij et al. 2010] report a good correlation between the high-frequency part of the list and the lists obtained with collocation methods.

To sum up, the majority of works on Verb-Noun collocations investigate the nouns that are involved in some particular types of verb-argument relations or Verb-Noun fixed Expressions (excluding Klyshinskij et al. 2010). In our research all the syntactically related to a verb noun are taken into consideration irrespective of their syntactic role (direct object, circumstantial NP etc.).

3. Experiment

3.1. Setup

The goal of the experiment described below is to compare the verb-noun collocations extracted from a large corpus with and without use of syntactic information. The corpus was preprocessed with a tokenizer, a POS-tagger, a morphology analyzer and a syntactic parser. We use the **Pointwise mutual information** (**PMI**) as a statistic measure for verb-noun collocation extraction. The two methods for collocation candidates extraction are used: the first one is a window-based bag of words method, the second one is based on the results of syntactic parsing.

The resulting collocation sets were grouped by verb and compared in order to evaluate the degree of correspondence between the extracted sets. The results of this comparison were then analyzed in detail, and several conclusions about the mismatching cases were made.

3.2. Corpus

In order to obtain sufficient amounts of initial data, a roughly 9 million word corpus of Russian newspaper texts was used[1]. The corpus consists of planarized random sentences sampled from various news articles published in the period from April 2011 to April 2012. This results in a certain lexical skewness, as the vocabulary, describing the events which have taken place in that period of time, influences the word distribution over the corpus. However, we believe that this factor can be disregarded, taking into account that the corpus was used to compare two automatic methods on the same dataset without use of any external data.

3.3. Preprocessing

The corpus has been preprocessed using a set of tools developed by S. Sharoff and J. Nivre ([Sharoff, Nivre 2011]), which includes a tokenizer, a TreeTagger-based ([Schmid 1994]) part-of-speech tagger, a lemmatizer based on CSTLemma ([Jongejan, Dalianis 2009]), and a Russian dependency parser model for the MaltParser ([Nivre et al. 2006]) trained on SynTagRus syntactic corpus ([Boguslavsky et al. 2000]). The preprocessed texts have been allocated in a relational database and indexed. The resulting dataset consists of tokens which are mapped to words; each word is assigned a set of morphological features and a lemma, and for each sentence a set of labeled dependency relations is fixed. S. Sharoff and J. Nivre report 95–97 % POS-tagger accuracy and an unlabeled attachment score of 88 ([Sharoff, Nivre 2011]), which seems sufficient for our type of analysis, taking into account that we aim to extract statistically dependent word pairs from a large corpus, and the impact of accidental errors should be smoothed by the dataset size. Although the relation labels are available, they weren't used during the experiment, so all the dependency links were treated as unlabeled ones.

3.4. Collocation extraction

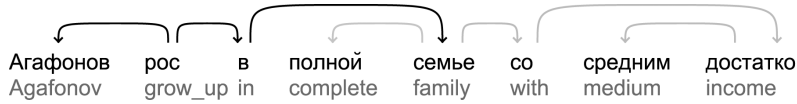
We have extracted collocations from the syntactically parsed corpus using two different strategies for obtaining the initial collocation candidate lists. Only finite verb forms were analyzed due to the fact, that the non-finite forms in Russian often lack some of the overtly expressed arguments, so taking these forms into account would require additional transformation and preprocessing steps.

The first candidate extraction strategy is to build potential collocate pairs by extracting unlabeled verb-noun dependency relations. In case of prepositional objects where the dependency relation points at a preposition, the preposition was skipped (see the collocation candidates *расти* ('grow up') — *Агафонов* ('Agafonov')

[1] The corpus was collected by H. Christensen and is available on http://corpora.heliohost.org/

и расти (‘grow up’) — *семья* (‘family’) in the example (1)). The total of 358,915 verb-noun pairs was obtained. We will refer to the collocations resulting from these pairs as syntax-based or dependency-based collocations.

(1)



(‘Agafonov grew up in a complete family with medium income’)

The second strategy is to use a window-based approach. In order to estimate the appropriate window size, the distribution of distances between verbs and their dependent nouns was analyzed (see ***Fig. 1***) and the window size of [−5; 5] words was selected as a result.

For every finite verb form in the corpus, we have extracted all the nouns found in the same sentence in the context of [−5; 5] words. The non-word tokens such as punctuation and numbers were ignored. From all the extracted pairings, a collocation candidate list containing verb lemma, noun lemma and the collocation frequency was formed. 708,131 collocations were extracted using the bag-of-words strategy. We refer to the collocations obtained using this method as **window-based collocations**.

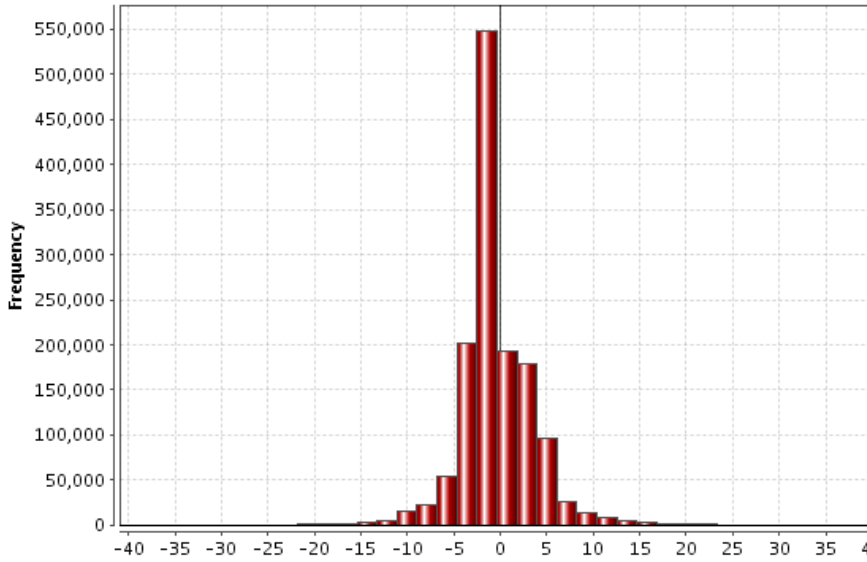


Fig. 1. Verb-argument distance distribution

The collocation candidates were ranked using the PMI metric, which is defined as

$$pmi(x,y) = \log \frac{P(x,y)}{P(x) * P(y)}$$

PMI as a word association measure has several drawbacks, among them an overrating of infrequent combinations and a poor accordance with expert collocation lists evaluation ([Evert, Krenn 2001]). The first is usually handled by establishing frequency cutoff thresholds. In our experiment only the combinations containing verbs with total raw frequency more than 100 and nouns with total raw frequency more than 10 were analyzed. As for the combination frequency threshold itself, we have found out that the cutoff value of 10 filters out too many combinations, resulting in very short final sets as compared to the sets obtained using lower cutoff threshold. Lower cutoff thresholds introduce some noisy data but also increase the recall, e. g.:

(2) ***сломать*** *('break')*
c10wc10 syntax, **window**: *рука ('arm'), нога ('leg')*
c5wc5 syntax, **window**: *нога ('leg'), нос ('nose'), ребро ('rib'), рука ('arm')*
c2wc2 syntax: *нога ('leg'), нос ('nose'), ребро ('rib'), результат ('result'), рука ('arm'), челюсть ('jaw')*
c2wc2 window: *андрей ('Andrej'), бедро ('hip'), год ('year'), женщина ('woman'), камера ('camera'), лицо ('face'), мальчик ('boy'), матч ('match'), нога ('leg'), нос ('nose'), падение ('fall'), палец ('finger'), побои ('beating'), раз('once'), ребро ('rib'), результат ('result'), рука ('arm'), челюсть ('jaw'), шея ('neck')*

The Cn notation is used to denote the cutoff threshold of n for syntactic model while the WCn denotes the cutoff threshold n for window-based model. For example, c10wc5 means that thresholds of 10 and 5 were applied to syntax- and window-based models respectively.

We have varied the frequency cutoff thresholds to examine the changes in correspondence between collocate sets built by using dependency-based and window-based approach. We have also compared the lists obtained using unequal thresholds.

3.5. Evaluation

The candidate sets extracted for each verb were ranked by PMI, and only the top 20 collocates were selected. In order to evaluate the degree of correspondence between the lists obtained using syntax- and window-based methods, for each verb we have calculated two weighted intersection measures using the formula:

$$WI(A;B) = \frac{|x \in (A \cap B)|}{|x \in A|}$$

Let window be the set of collocations for a given verb extracted using the window-based technique. Let syntax be the one extracted using the syntax-based method. The measure aims to describe how good the window-based list of nouns fits into the one extracted using the syntax-based representation. The measure is inversely the ratio of words from the syntax-based list, which are also presented in the list of words obtained by applying the window method. These measures can be thought of as Precision and Recall with syntax-based set treated as Key. As with Precision and Recall, the harmonic mean of two measures (F-measure) was also computed using the standard formula:

$$F_1 = \frac{2 * WI(window, syntax) * WI(syntax, window)}{WI(window, syntax) + WI(syntax, window)}$$

The comparison results are presented in Table 1.

Table 1. Comparison of window-based and syntax-based collocations

WI(window,syntax)				WI(syntax,window)			
	wc10	wc5	wc2		wc10	wc5	wc2
c10	0,62109	0,27844	0,11761	c10	0,93730	0,84038	0,60494
c5	0,79878	0,55038	0,20042	c5	0,60583	0,88075	0,64143
c2	0,67867	0,66600	0,49630	c2	0,22842	0,44996	0,69669
average=0,48974				average=0,65396			

F1-measure			
	wc10	wc5	wc2
c10	0,71847	0,38428	0,17429
c5	0,66367	0,64752	0,26928
c2	0,30926	0,51122	0,55542

3.6. Results

The evaluation shows a moderate level of correspondence between the results obtained by comparing the two methods discussed. As WI measures for different combinations of minimal combination frequency threshold shows (*Table 1*) using distant threshold values (e.g. *c10-wc2*) leads to the worst results. According to the table, the best F_1 is achieved using the threshold of 10 for both syntax- and window-based algorithms, though the small size of resulting sets must be taken into account. The best *WI(syntax,window)* is achieved by using cutoff threshold of 5 on syntax- and the one of 10 on window-based candidates.

The value of *WI(syntax,window)* averaged on all threshold combinations is significantly higher than the one of *WI(syntax,window)*. This reflects the fact that in many cases the majority of the words from syntax-based lists are included into the window-based lists, while the opposite is false.

Taking into account the starting hypothesis and presuming that the syntactic relations should give less noisy data, one could suggest that the collocation sets, obtained using window-based candidate list, would lack precision and introduce too much noisy data. However, the expert analysis shows that in many cases the collocates extracted by the window-based model are perfectly relevant to the task and should be treated as correct. Consider the following examples from sets obtained using frequency threshold of 5 in both algorithms. Matching words are shown in bold, and relevant words are underlined.

(3) ***забить*** *('kick, score')*c5wc5
syntax: ***ворота ('goal')***, ***год ('year')***, ***гол ('goal')***, ***голова ('head')***, ***матч ('match')***, ***минута ('minute')***, ***мяч ('ball')***, ***сезон ('season')***, ***тревога ('alarm')***, *форвард ('forward')*, ***шайба ('puck')***
window: ***ворота ('goal')***, ***год ('year')***, ***гол ('goal')***, ***голова ('head')***, *игра ('game')*, *команда ('team')*, ***матч ('match')***, ***минута ('minute')***, *момент ('moment')*, ***мяч ('ball')***, *пенальти ('penalty')*, *полузащитник ('halfback')*, ***сезон ('season')***, *смерть ('death')*, *состав ('members')*, *счет ('score')*, *тайм ('time')*, ***тревога ('alarm')***, *чемпионат ('championship')*, ***шайба('puck')***

4. Discussion

The analysis of the results shows that both methods share some common advantages and disadvantages, and the particular disadvantages of each method can be both due to experimental setting drawbacks and linguistic features of the texts. It turns out that, contrary to the expectations, the window-based method tends to extract some relevant verb-noun collocations, which are absent in the sets obtained by the syntax-based method. While the window-based approach also results in a higher level of noise, the syntax-based method suffers from narrowness of syntactic patterns used to extract collocation candidates. Our results show that using simple syntactic patterns is insufficient to model the semantic relations between predicate verbs and their arguments, which results in lower recall.

4.1. Common shortcomings

4.1.1. Corpus skewness

A common shortcoming of the lists obtained using both candidate extraction techniques is collocation specificity, which is related to the skewness of the source data. The texts in our corpus were obtained from news articles released in a one-year

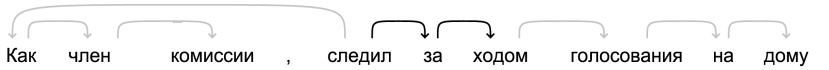
period, so the names of objects which were often mentioned in the media in that time span influence the statistics obtained from the whole corpus. That problem could be partially solved by using a larger and more representative corpus or recognizing and filtering out named entities.

(4) ***возглавить*** *('be head of')*c10wc10
syntax: *год ('year'), рейтинг ('rating'), совет ('council'), список ('list')*
window: *александр ('Alexander'), владимир ('Vladimir'), год ('year'), группа ('group'), дмитрий ('Dmitry'), комитет ('committee'), медведев ('Medvedev'), отделение ('department'), партия ('party'), правительство ('government'), президент ('president'), путин ('Putin'), рейтинг ('rating'), руководитель ('leader'), сергей ('Sergej'), совет ('council'), список ('list'), управление ('board'), человек ('man')*

4.1.2. Capturing the parts of other constructions

In some cases, the verb is syntactically related to a head of a fixed expression. In this case the collocations extracted by both methods will be invalid (see the examples below):

(5) ***следить*** *('follow')* c5wc5
syntax: *ход ('progress')*
window: *ход ('progress'), голосование ('voting')*



(6) ('As a member of the commission, he followed the progress of the voting at home')

4.2. Window method disadvantages

4.2.1. Capturing the dependant of a valid collocate

These cases are similar to the example (7), but here the collocation extracted by the syntactic method may be considered valid. At the same time, the window-based approach erroneously extracts its dependant:

(7) ***отклонить*** *('decline')*:
syntax: *жалоба ('complaint'), иск ('suit'), предложение ('proposition'), суд ('court')*
window: *жалоба ('complaint'), иск ('suit'), москва ('Moscow'), предложение ('proposition'), суд ('court')*

(8) ('Yesterday the Moscow Arbitrage Court declined the suit of Rosimushestvo to JSC…')

The collocations extracted this way reflect the skewness of the corpus.

4.2.2. Frequent uninformative noise

The occurrence of high-frequency non-informative words like *человек* (*'man'*), *год* (*'year'*), *Россия* (*'Russia'*) is a prominent feature of the collocation lists extracted by window-based method. They may be linguistically unrelated, as *человек* in (9):

(9) ***отпустить*** *('let out')*c5wc5
syntax: *залог ('bail'), игрок ('player'), свобода ('freedom'), суд ('court')*
window: *залог ('bail'), игрок ('player'), свобода ('freedom'), суд ('court'), человек ('man')*

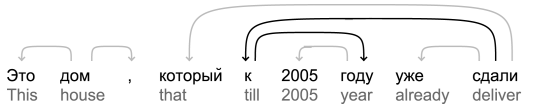
(10)

('Many of them were let out, but several people spent all night at the police station')

It may also be a member of a regular circumstantial construction as *год in* (11):

(11) ***сдать*** *('pass')*c5wc5
syntax: *экзамен ('exam')*
window: *год ('year'), экзамен ('exam')*

(12)



('Is this the house that has already been commissioned by 2005?')

This type of nouns in the top of the window lists is due to the general frequency of expressions of an event time in a clause (it's also true for some other semantic relations). For instance, the collocation *год ('year')* is found only by window-based method in 137 verbs out of 548 within the cutoff threshold of 5. We suppose that an additional procedure of filtering such cases could increase the degree of syntax-based and window-based lists overlapping.

4.3. Syntax-based method shortcomings

Although we have analyzed only finite verb forms in order to reduce syntactic complexity, there are still many issues related to describing the syntactic construction in which semantic relatedness can be expressed. In many cases, a related noun was not captured by the syntax-based candidate extraction algorithm due to the absence of the direct syntactic relationship to the verb in a sentence. Common cases include relative clauses, argument coordination and object pronominalization.

4.3.1. Relative clauses

One common case which is not taken into account by our syntax-based model is the one when the verb is located in a relative clause as in the following example:

(13) ('After the lecture that he read us at school…')

The window-based model was able to extract the candidate *лекции* (*lectures*) + *прочитать* (*read*) while the power of our syntax pattern-based method was insufficient to capture the semantic relatedness between these two words. In cases when the amount of such constructions is high, this issue can influence the overall corpus statistics, e. g.:

(14) ***прочитать*** *('read')* c10wc10
syntax: *интернет ('Internet'), книга ('book')*
window: *интернет ('Internet'), книга ('book'), лекция ('lecture')*

4.3.2. Argument coordination

Another syntactic relation type that should be taken into account is coordination. The parser that we used is based on the framework where the dependency relation between a verb and its coordinated arguments is drawn to the first of these arguments, followed by a chain dependency through a conjunction. See Figure 16, where the verb "выехали" (go, leave) is connected only to the first argument *полицейские* (*policemen*). That first argument is then connected to the second one, *сотрудники* (*officials*) with the conjunction *и* (*and*).

(15) ***выехать*** *('drive off)*c5wc5
syntax: *автомобиль ('car'), группа ('group'), место ('place'), полоса ('lane'), раз ('once')*
window: *автомобиль ('car'), глава ("head), год ('year'), группа ('group'), движение ('traffic'), дом ('house'), машина ('car'), место ('place'), область ('region'), полиция ('police'), полоса ('lane'), происшествие ('accident'), сотрудник ('official'), управление ('board'), человек ('man')*

(16) ('The police and the officials of the recruiting office drove off on their track')

Note that the window-based method succeeds to extract collocations in some of these cases.

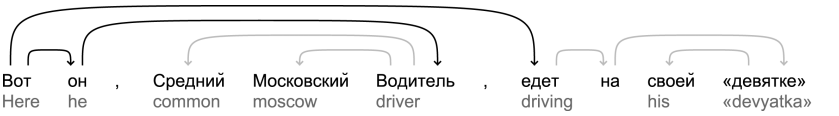
4.3.3. Argument pronominalization

The final drawback of the syntactic method which is worth mentioning is that our model lacks co-reference information. In many cases, the core arguments of a verb

(especially, the subject and object) are substituted by a pronoun. When the antecedent is in the same sentence, it still can be located by the window-based approach, but the syntax-based candidate extractor fails to identify the candidate due to the lack of coreference information, as in the following example:

(17) ***ехать*** *(‘go, travel’)*c5wc5
syntax: *вагон (‘carriage’), машина (‘car’)*
window: *автобус (‘bus’), вагон (‘carriage’), водитель (‘driver’), год (‘year’), машина (‘car’), минута (‘minute’), человек (‘man’)*

(18)



(‘Here he is, the Average Moscow Driver, traveling in his “devyatka” (car model)’)

However, the antecedent is not always located in the same sentence. In this case, both methods fail to identify collocation candidates. Improving the preprocessor by adding a co-reference resolution engine should increase the overall numbers of collocation candidates and soften the consequences of the fact that some collocate types tend to be pronominalized more often than the others.

4.4. Typical argument extraction

Although some researches use (or assume the need of use of) syntactic information to extract typical arguments from the collocation lists or to filter them out, our study shows that both methods are suitable for the extraction of such verb-noun constructions. Both typical subjects and typical objects can be retrieved by either method, see examples below:

Typical subjects
(19) ***арестовать*** *(‘arrest’)*c10wc10
syntax: *полиция (‘police’), суд (‘court’)*
window: *год (‘year’), полиция (‘police’), суд(‘court’)*

Typical objects
(20) ***сломать*** *(‘break’)*c10wc10
syntax: *рука (‘arm’), нога (‘leg’)*
window*: рука (‘arm’), нога (‘leg’)*

The possible way to take into consideration the particular type of arguments in the window-based method is to use the more granulated noun morphological features such as cases. However the distinguishing between these two cases or, in more complex cases, between subject and object collocates within the list of one verb was beyond the scope of our research.

5. Conclusion

In our study we have compared two methods of building collocation candidate lists within the framework of verb-noun collocation extraction. We have conducted an experiment on extracting and ranking collocation candidates from a large preprocessed corpus of news data using two different candidate extraction methods. An automatic comparison of collocation lists obtained using window-based and syntax-based candidate extractors has shown only a moderate level of correspondence. The detailed analysis of the comparison results makes it possible to identify common advantages and disadvantages of both methods.

In general, the window-based extractor seems to outperform the one based on a syntax-driven approach in terms of recall. Our results show that the simple syntax collocation model which only takes direct and prepositional verb-noun dependencies into account is not powerful enough. It is due to two basic phenomena. The first one is that there are a sufficient number of cases when the semantically related nouns are not immediate dependant of a verb. Moreover they can occur close to a verb but in another clause. The second one is the anaphora phenomena. The arguments of a verb can be pronominalized or omitted in real discourse especially as far as subject NPs is concerned. Adding special modules for syntax-based collocation extraction for treating these phenomena might improve the quality of the syntax-based method.

References

1. *Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N.* (2000), Dependency treebank for Russian: concept, tools, types of information. In Proceedings of the 18th conference on Computational linguistics — Volume 2, COLING '00, pp. 987–991.
2. *Breidt E.* (1993), Extraction of V-N-collocations from text corpora: A feasibility study for German, Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Columbus, USA.
3. *Church K. W., Hanks P.* (1990), Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), pp. 22–29.
4. *Evert S., Krenn B.* (2001), Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, pp. 188–195.
5. *Gildea D., Jurafsky D.* (2002), Automatic Labeling of Semantic Roles , Computational Linguistics, 28(3).
6. *Jagunova E., Pivovarova L.* (2010), The nature of collocations in Russian. The experience of automatic extraction and classification in news text. ['Priroda kollokatsiy v russkom jazyke. Opyt avtomaticheskogo izvlechenia I klassifikatsii na material novostnykh tekstov'], NTI, 2, №6. M.
7. *Jongejan B., Dalianis H.* (2009), Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore : Association for Computational Linguistics, 2009. pp. 145–153.

8. *Khokhlova M.* (2008), Extracting Collocations in Russian : Statistics vs . Dictionary, JADT 2008: Proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12–14, 2008, pp. 613–624.
9. *Khokhlova M.* (2008), The experimental verification of collocation extraction ['Eksperimentalnaja proverka metodov vydelenija kollokatsij'], Slavica Helsingiensia, 34. Helsniki. pp. 343–357.
10. *Khokhlova M.* (2009), Applying Word Sketches to Russian. In: Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, pp. 91–99.
11. *Kilgarriff A., Tugwell D.* (2002), Sketching words, Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins. Marie-Hélène Corréard (Ed.) EURALEX, pp. 125–137.
12. *Klyshinskij E., Kochetkova N., Litvinov M., Maksimov V.* (2010), Automatic construction of word combination database using a huge text corpus ['Avtomaticheskoje formirovanije bazy sochetaemosti slov na osnove ochen bolshogo korpusa tekstov'], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010"], Bekasovo, pp. 181–185.
13. *Kustova G., Toldova S.* (2009), RNC: Semantic filters for the verb disambiguation ['NKRJA: semanticheskije filtry dlja razreshenija mnogoznachnosti glagolov']. Russian national corpus: 2006–2008. New results and perspectives. ['Natsionalnyi korpus russkogo jazyka: 2006–2008. Novye rezultaty i perspektivy']. Saint-Petersburg: Nestor-Istorija.
14. *Lin D.* (1998), Automatic Retrieval and Clustering of Similar Words, COLING-ACL98, Montreal, Canada.
15. *Nivre J., Hall J., Nilsson J.* (2006), Maltparser: A data-driven parser-generator for dependency parsing. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 2216–2219.
16. *Orliac B., Dillinger M.* (2003). Collocation extraction for machine translation, Proceedings of Machine Translation Summit IX, New Orleans, LA, USA, pp. 292–298.
17. *Pado S., Lapata M.* (2007), Dependency-based Construction of Semantic Space Models, Computational Linguistics 33(2), pp. 161–199.
18. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
19. *Sharoff S., Nivre J.* (2011), The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Proc. Dialogue 2011, Russian Conference on Computational Linguistics.
20. *Todirascu A., Gledhill C.* (2008), Extracting Collocations in Context: The case of Verb-Noun Constructions in English and Romanian. Recherches Anglaises et Nord-Américaines (RANAM), Université Marc Bloch, Strasbourg.
21. *Todirascu A., Tufis D., Heid U., Gledhill C., Stefanescu D., Weller M., Rousselot F.* (2008), A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions, Proceedings of LREC'2008, Marrakesh, Morocco.

АЛЕКСАНДР ЕВГЕНЬЕВИЧ КИБРИК: ОТ СТРУКТУРАЛИЗМА К НОВЫМ ИДЕЯМ

Алпатов В. М. (v-alpatov@yandex.ru)

Институт языкознания РАН, Москва, Россия

Деятельность А. Е. Кибрика отразила процесс движения от структурной к функциональной лингвистике; этот процесс характерен для всё современной лингвистики, но А. Е. Кибрик очень чётко сформулировал принципы новой парадигмы, особенно в статье «Лингвистические постулаты» (1983/1992). Он указал на ограниченность структурной лингвистики, изучающей лишь структуру языка. Он писал, что необходимо изучать лингвистические явления вместе с мыслительной деятельностью говорящего человека, призывал выявлять языковые процессы так, как они происходят «на самом деле», подчёркивал центральную роль семантики в языке. Теперь развитие лингвистики происходит в направлении, указанном А. Е. Кибриком тридцать лет назад.

ALEXANDR EVGENJEVICH KIBRIK: FROM STRUCTURALISM TO NEW IDEAS

Alpatov V. M. (v-alpatov@yandex.ru)

The Institute of Linguistics RAS, Moscow, Russia

The activities of A.E. Kibrik reflected the process of movement from the structural linguistic to the functional one; this process is characteristic for the modern linguistics but Kibrik formulated the main principles of a new paradigm very precisely, especially in his article “Linguistic postulates” (1983–1992). He pointed to the narrowness of the structural linguistics studying only the structure of the language. He wrote that it is necessary to study linguistic phenomena with the mental activity of speaking persons, called to reveal linguistic processes as a matter of fact, emphasized the central role of semantics in the language. Now linguistics develops in the direction that was determined by A. E. Kibrik thirty years ago.

Александр Евгеньевич Кибрик (1939—2012) внёс вклад в самые разные области науки о языке: синтаксис и семантику, типологию и полевую лингвистику, кавказоведение и прикладное языкознание. Его всегда влекли теоретические проблемы, но он умел анализировать и самые конкретные языковые факты. И прикладные исследования занимали в его научном творчестве

большое место. Кстати, и многолетнее его преподавание на ОСИПЛ началось (если не считать недолгого ведения курса латыни) с чисто прикладного курса, где он рассказывал студентам проводимую им на кафедре хоздоговорную работу. И в, пожалуй, главной теоретической книге А. Е. Кибрика [Кибрик 1992], которой в основном посвящён доклад (далее в ссылках на неё будет указываться лишь номер страницы), одна из её четырёх частей (261–323) называется «Приложения языкознания».

Но я, прежде всего, хочу рассмотреть развитие в работах А. Е. Кибрика самых общих и вечных проблем лингвистики. Представляется, что он, как никто другой в отечественном языкознании, отразил в своих публикациях происшедшую во второй половине XX в. смену научной парадигмы, переход от изучения языка «в себе и для себя» к исследованию языкового функционирования.

Годы, когда Александр Евгеньевич начал заниматься лингвистикой, в советской науке о языке характеризовались противостоянием структуралистов и так называемых традиционалистов (последние в отличие от структуралистов не составляли какого-либо единства). А. Е. Кибрик с самого начала принадлежал к структурному лагерю, одним из центров которого была кафедра МГУ, тогда возглавлявшаяся В. А. Звегинцевым, на которой Александр Евгеньевич проработает всю жизнь.

Тогда на кафедре и вокруг неё господствовали идеи позднего структурализма, в этом духе воспитывали студентов. Грубо говоря, принято было сочетать постулаты, шедшие от Ф. де Соссюра (изучать язык, а не речь; соблюдать чисто лингвистический подход), и появившееся позже стремление всё математизировать. Идеалом казалось максимальное приближение лингвистики к математике, и в подходах, и в языке. Начиная от Ф. де Соссюра, считалось нужным ограничиваться развитием «внутренней лингвистики», исследующей устройство языка.

К автономии лингвистики стремились представители структурной лингвистики разных направлений. Согласно дескриптивистам (З. Харрис), лингвист изучает лишь повторяющиеся свойства речевых отрезков, а всем прочим из связанного с языком должны заниматься антропологи, психологи, литературоведы и пр. Л. Ельмслев считал, что лингвистическая теория строится чисто произвольно как исчисление, независимо от конкретных языков. Даже более умеренный Н. С. Трубецкой, занимавшийся не только лингвистикой, писал Р. О. Якобсону в 1929 г.: «К сожалению, надо признаться, что такой «культурно-исторический» уклон в современной русской лингвистике существует, и что наша «структуральная» лингвистика вовсе не является господствующей. Яковлев напр. весь в культурно-исторических соблазнах» [Трубецкой 2004: 131]. То есть, с его точки зрения, структурный и «культурно-исторический» подходы были несовместимы.

Лингвисты структурного лагеря тех лет (всё же с разной степенью последовательности) обрывали связи своей науки с другими гуманитарными дисциплинами, оставляя эти связи на долю «традиционалистов». Характерно и игнорирование в те годы социолингвистики научным окружением А. Е. Кибрика и большинством выпускников ОСИПЛ (В. И. Беликов — одно из редких исключений). Это отчасти происходило по экстралингвистическим причинам,

но и собственно научная ситуация способствовала тому, что и социолингвистика оставлялась «традиционалистам».

Уровень развития той или иной науки тогда (не только среди лингвистов) измерялся по степени использования в ней математики. А математика — всё-таки совершенно особая наука, в том числе и в том отношении, что она стремится к сложности. Ряд лингвистов использовал крайне сложный математический аппарат для построения моделей очень простых, в сущности, явлений, которые давно изучали без каких-либо усложнений. Здесь нельзя не назвать Ю. К. Лекомцева, настоящего подвижника науки, он продолжал трудиться, несмотря на тяжёлую болезнь. Его деятельность вызывала уважение, но вставал вопрос: нужно ли строить математическую модель, скажем, понятия фонемы, если и пражцы, и москвичи добивались в фонологии результатов и без этого.

Математический идеал сложности влиял и на собственно лингвистические подходы. И. А. Мельчук, оставшийся верным этому идеалу и в другую эпоху, пишет по проблеме слова: «Лингвисты не случайно веками искали определение этому понятию: как и все базовые понятия, «слово» — это очень сложный объект. Его описание требует углубленного анализа многочисленных фактов, большого числа промежуточных понятий, а также исследований в смежных областях, которые сами по себе не менее сложны» [Мельчук 1997: 95–96]. Замечу, что многократные попытки такого «углубленного анализа», предпринимавшиеся многими учёными, мало что дали.

В целом можно сказать, что испокон веков перед лингвистами стояли три вопроса: «Как устроен язык?», «Как изменяется язык?» и «Как функционирует язык?». Первый из них исторически начал изучаться первым (если говорить о европейской традиции, то в Древней Греции). В XIX в. наука сосредоточилась на изучении изменений. С начала XX в. структурный подход снова, на более высоком уровне занялся устройством языка. Вопросы же функционирования языка до середины XX в. ставили лишь немногие учёные (В. фон Гумбольдт, позднее Э. Сепир), при замечательных догадках ещё не имевшие метода.

А. Е. Кибрик, начинавший работать в рамках, установленных структурализмом, и позже, в главном изменив точку зрения, отдавал должное науке тех лет. «Лингвистика… занималась… инвентаризациями и классификациями наблюдаемых единиц различных типов…. Такой этап был необходимым звеном в эволюции лингвистики и проделанный ею труд был поистине титаническим» (26). «Метод «черного ящика» действительно оказался полезным на начальных этапах моделирования» (19). «Алгоритмическое мышление на определенном этапе имело революционное значение для лингвистики, дав метод уточнения и формализации многих грамматических процессов» (36).

И всё же к 1960-м гг. были учёные, шедшие дальше. Первым был Н. Хомский. Его значение тогда у нас только-только начинали осознавать; многие лингвисты, расположенные к новым методам, долго считали его дескриптивистом. Первым, кто в СССР в полной мере оценил его новаторство, был В. А. Звегинцев, который постоянно выделял у Н. Хомского призыв к тому, чтобы лингвистика изучала язык вместе с говорящим на нём человеком. Но Звегинцев подчёркивал, что Хомский, хотя и включил лингвистику в состав психологии

познания, но слишком увлёкся формальным аппаратом и больше декларировал обращение к говорящему человеку, чем реально это делал. «Компетенция идеального говорящего — слушающего» также вводила лингвистику в жёсткие рамки, пусть несколько иные, чем это было у Ф. де Соссюра.

У нас последовательных хомскианцев долго не было, хотя примерно до середины 70-х гг. за развитием генеративных идей следили. Альтернативой Хомскому у нас тогда считалась модель «Смысл ⇔ Текст». Она также была одним из этапов движения в сторону расширения научных горизонтов. Помню такие два устных высказывания И. А. Мельчука 60–70-х гг. (передаю по памяти): «У Хомского лучше, чем у нас, разработан формальный аппарат, но это и сдерживает его развитие, содержательная сторона за формализмами пропадает». «Мне по существу не важен машинный перевод. Важно понять, каким образом человек говорит». То есть у разработчиков модели были и интерес к содержательным процессам, и стремление понять, что происходит «на самом деле». Отмечу и гораздо большую по сравнению с хомскианством роль семантики в модели. Несомненно, что А. Е. Кибрик, пройдя через этап следования данной модели, многое из её принципов отразит и в будущих постулатах. Но не всё. Для Мельчука сложность формального аппарата, если и могла становиться недостатком, то только таким, который является продолжением достоинств; показательна вышеприведённая цитата о слове. И модель находилась в пределах языка, по Соссюру; мыслительные процессы оставались внутри «чёрного ящика».

В ранних публикациях А. Е. Кибрик ещё находился под влиянием модели «Смысл ⇔ Текст», но к началу 1980-х гг. он пришёл к выводу о том, что и ей свойственно ограничение подходов. «Практически все существующие модели языка, как статические (классическая традиционная грамматика языка, концепция Соссюра, Ельмслева и др.), так и динамические (генеративная грамматика, модель «Смысл ⇔ Текст» и др.), страдают недоучетом функциональной предопределенности языка, производности его от речевой деятельности и прагматических условий его использования» (16).

По-новому учёный подошёл к общим проблемам языка в статьях и докладах, которые появлялись, начиная с середины 1970-х годов, и в итоге составили книгу [Кибрик 1992]. Среди них особенно важна статья «Лингвистические постулаты» (17–27), впервые изданная в 1983 г. и переработанная при включении в книгу. Автор поставил в ней цель «по возможности эксплицитно сформулировать те коренные сдвиги в лингвистической идеологии, которые происходят в ней в последние годы и, как мне думается, должны предопределить ее развитие в ближайшем будущем» (17).

Я сначала перечислю постулаты, а затем прокомментирую большинство из них. «Адекватная модель языка должна объяснять, как он устроен «на самом деле» (19). «Всё, что имеет отношение к существованию и функционированию языка, входит в компетенцию лингвистики» (20). «Как содержательные, так и формальные свойства синтаксиса в значительной степени предопределены семантическим уровнем» (21). «К области семантики (в широком смысле) относится вся информация, которую имеет в виду говорящий при развертывании высказывания и которую необходимо восстановить адресату для правильной интерпретации

этого высказывания» (22). «Необходима разработка лингвистических моделей класса «МЫСЛЬ-СООБЩЕНИЕ» (23). «Исходными объектами лингвистического описания следует считать значения» (24). «Устройство грамматической формы отражает тем или иным образом суть смысла» (25). «Сложны лингвистические представления о языке вследствие их неадекватности, а язык устроен просто» (25).

Первый постулат направлен против метода «черного ящика», где «задачей моделирования является не изучение ненаблюдаемой «сущности» оригинала, а построение некоторого конструкта произвольной внутренней природы, внешние проявления которого идентичны внешним, наблюдаемым проявлениям оригинала» (18). Этот метод основан «на допущении, что различные по своей природе объекты могут иметь идентичные «входы» и «выходы» (19). Однако «адекватная модель языка должна объяснять, как он устроен «на самом деле». Что такое «язык на самом деле»? Это совокупность тех знаний, которыми располагает человек, осуществляя языковую деятельность на соответствующем языке. В отличие от метода «черного ящика» «естественное» моделирование языка должно осуществляться с учетом того, как человек реально пользуется языком, то есть, как он овладевает языком, как хранит в своей памяти знания о языке, как использует эти знания в процессе говорения, слушания, познавательной деятельности, и т.д. … Предполагается, что различные по своему устройству объекты такого класса сложности, к которому относится естественный язык, не могут иметь идентичных «входов» и «выходов» (19).

Очевидно, что «естественное» моделирование языка невозможно в рамках внутренней лингвистики, по Ф. де Соссюру. Для него нужно изучать речь и психологические основы речевой деятельности. Такое изучение, разумеется, происходило и раньше, но многое только начинало исследоваться, а то, что уже было сделано, например, работы А. Р. Лурия и др. по афазии, часто воспринималось как что-то далёкое от лингвистики.

С этим связан и следующий постулат: о границах лингвистики. «То, что считалось «не лингвистикой» на одном этапе, включается в него на следующем. Этот процесс лингвистической экспансии нельзя считать законченным…. И каждый раз снятие очередных ограничений дает новый толчок лингвистической теории, конкретным лингвистическим исследованиям» (20).

В истории науки о языке чередуются периоды расширения и сужения её сюжетов. Двумя эпохами сужения тематики стали период господства сравнительно-исторического языкознания в XIX в. и времена структурализма. Но к 1980-м годам ситуация стала меняться. И если хомскианцы лишь заменили прежние рамки новыми, то лингвисты иных направлений вообще отказались устанавливать жесткие границы лингвистики.

Следующий постулат связан с соотношением синтаксиса и семантики. К 1980-м гг. преимущественное внимание раннего структурализма к фонологии и морфологии уже ушло в прошлое, центр теоретических интересов переместился в сторону до того плохо изученных синтаксиса и семантики. Но приоритеты лингвисты расставляли по-разному. Центральное место синтаксиса в языке — один из важнейших пунктов генеративной теории, и постулат А. Е. Кибрика о предопределении синтаксиса семантикой явно полемичен

против генеративизма. Полемизируя также с «самоограничением, доведённым до абсурда», у некоторых дескриптивистов, изгонявших из лингвистики значения, он пишет: «Можно было бы, нарочито утрируя, сказать прямо противоположное: в лингвистике ничего (или почти ничего) нет, кроме проблемы значения» (20). А. Е. Кибрик отмечает, что «традиционное» языкознание могло понять проблему адекватнее: о примате семантики над синтаксисом говорил еще А. А. Шахматов (22). К данному постулату примыкает и постулат о границах семантики, утверждающий её максимально широкое понимание; в том числе и то, что «прагматические компоненты могут рассматриваться как частные сферы семантического представления» (22).

Следующий постулат прямо полемичен по отношению к модели «Смысл ⇔ Текст», отказывавшейся от рассмотрения мышления. Семантика имеет приоритет перед синтаксисом и тем более перед морфологией, однако в свою очередь «естественно считать, что мыслительные процессы имеют приоритет перед семантикой, поскольку мысль предшествует смыслу и создает его» (24). Но мыслительные процессы изучает целый ряд дисциплин, и «имманентность» лингвистики, тем самым, иллюзорна.

Ещё один постулат касается выбора одного из двух главных направлений изучения языка от смысла к форме или от формы к смыслу. Веками почти всё языкознание шло по пути от формы к смыслу (пассивной грамматики, по Л. В. Щербе), моделирующему деятельность слушающего. Иной путь (активной грамматики, по Л. В. Щербе), моделирующий деятельность говорящего, всерьёз начал избираться лишь недавно. Причины этого понятны. Значения в отличие от форм нам не даны, и критерии их выделения разрабатывались с большим трудом. Однако, исходя из принципа приоритета семантики, А. Е. Кибрик осуждает традиционный подход от формы к значению: «Чем более скрупулезно проводятся такого рода исследования, тем более сложными оказываются соответствия между изучаемыми единицами (формами) и приписываемыми им значениями. Всякий раз оказывается, что некоторая форма имеет много значений (причем распределение значений крайне сложно и все глубже утопает в стихии контекста), а каждое из значений в свою очередь имеет много способов выражения» (24). Он признаёт, что «процесс познавательной деятельности лингвиста должен быть циклическим» (19), совмещающим оба пути исследования, но, прежде всего, нужно идти от значений. Пожалуй, это самый спорный из постулатов: два пути исследования, приближающиеся с разных сторон к своему объекту, отражают две стороны речевой деятельности человека и равно важны. Другое дело — то, что исследования, идущие от значения, развиты хуже, и здесь можно ожидать значительного продвижения.

Соотношение между смыслом и формой в ином аспекте рассматривается и в постулате о мотивированности. Форма, в конечном итоге, всегда мотивируется смыслом, хотя в истории языков бывает, «что эта связь стерта, демотивирована», и тогда надо «искать исходное мотивированное состояние» (25). См. также в других главах книги: «Многие аномалии синхронных состояний парадигматических систем могут быть устранены при более широком объяснительном подходе, учитывающем, в частности, закономерности исторических

изменений грамматических систем» (99). «С исторической точки зрения сомнительно наличие в языке немотивированных связей между значением и формой; кажущееся отсутствие мотивации следует объяснять тем, что эта связь стерта, демотивирована, и необходимо найти исходное мотивированное состояние…. Такой подход означал бы формирование содержательной, а не формальной исторической лингвистики» (130).

Постулат о неразрывной связи грамматических форм и смысла был противопоставлен сразу нескольким положениям структурной лингвистики (или хотя бы части его направлений). Во-первых, неадекватным признано преобладавшее в структурализме представление об описании языка как конечной цели работы лингвиста. А. Е. Кибрик пишет: «Если лингвистика недавнего прошлого допускала лишь вопросы типа «как?» и избегала вопросов «почему?», то теперь ситуация должна коренным образом измениться: нужны именно ответы на вопросы типа «почему?», поскольку только они могут что-либо объяснить» (25). См. также о типологии в другой статье книги: «На смену безраздельного господства… КАК-типологии приходит объяснительная ПОЧЕМУ-типология, призванная ответить не только на вопросы о существовании, но и о причинах существования / несуществования тех или иных явлений» (29). Объяснительный подход к языку, предложенный ещё модистами в XIII–XIV вв., редко встречался у структуралистов. Но к моменту появления статьи его уже провозгласил Н. Хомский, однако концепция А. Е. Кибрика существенно другая.

Во-вторых, из презумпции исходной мотивации следует ограниченное количество грамматических возможностей, которыми обладают языки. Если дескриптивисты полагали, что в каждом новом языке мы можем обнаружить то, что нам никогда не встречалось, то из постулата А. Е. Кибрика следует, что категории грамматики не могут варьироваться до бесконечности. Как сказано во включённой в книгу статье «Язык» в ЛЭС, «степень покрытия универсума значений и принципы его членения не произвольны и не беспредельно разнообразны» (13). Кстати, отмечу, что после сенсационного открытия языка дирбал в 1970е гг., в лингвистике больше, кажется, не было столь повлиявших на развитие теорий фактических открытий.

В-третьих, тезис о поисках «исходного мотивированного состояния» снимает противопоставление синхронии и диахронии в той жёсткой форме, которую придал ей Ф. де Соссюр. Это больше похоже на идеи И. А. Бодуэна де Куртенэ: статический анализ необходим, но полностью язык быть познан только в динамике. А «стёртая, демотивированная связь» неожиданно напоминает «технизацию» В. И. Абаева. Этот языковед в годы молодости А. Е. Кибрика считался крайним «традиционалистом», но в некоторых случаях мог высказывать интересные идеи.

Последний постулат касается мнимой, по мнению А. Е. Кибрика, сложности языка. «До сих пор… сложность… описаний приписывалась языку как его сущностное свойство» (25). Но «модели такого рода, развиваясь, имеют тенденцию экспоненциально усложняться и становиться не упрощенными копиями оригинала, а объектами, по сложности не только не уступающими оригиналу, но скорее всего значительно его превосходящими» (19). Выше упоминалось,

что такая осуждаемая А. Е. Кибриком сложность естественна при построении лингвистики по образцу математики. Однако, как сказано в статье из ЛЭС, «Чем проще объект устроен (то есть чем непосредственнее его структура отражает его функции), тем сложнее его познать (в особенности при недоучете функционального аспекта)» (15). Представляется, что слово как раз является таким понятием. Как писал А. Е. Кибрик, «словарные единицы хранятся как готовые к употреблению, автоматически воспроизводимые двусторонние сущности, в то время как единицы, в образовании которых участвуют грамматические правила, в готовом виде в памяти отсутствуют и специально строятся в соответствии с некоторым коммуникативным заданием» (13–14). Слово — базовая единица словаря в этом смысле, а её лингвистические свойства могут быть разными и не совпадать в разных языках.

Вопрос о простоте и сложности в науке о языке тесно связан с проблемой формализации лингвистических исследований. На этой проблеме А. Е. Кибрик остановился в двух главах данной книги, посвящённых типологии. Вспоминая ряд попыток 1960-х годов, он пишет: «Не стоит говорить о такой формализации, которая создается исключительно ради самой себя и не обеспечивает нового, более глубокого понимания предмета, лишь переписывая другим способом давно известные неформализованные истины» (42). В связи с этим он согласен с высказыванием Дж. Лакова: «В настоящий исторический момент любое описание языка, которое тесно связано с некоторой формальной теорией, не в состоянии описать большинство явлений, имеющихся в языке» (42–43). Из этого «не следует делать вывод о неприемлемости формального аппарата для описания редких языков. Однако использование его должно быть очень осторожным и осмотрительным» (43). Формализация важна, прежде всего, для лингвиста, толкая его к системному анализу, но не для читателя: «Всякое хорошее формальное описание может быть изложено и неформально» (43).

В другом месте автор высказывается о причинах ограниченности формальных моделей: «Далеко не все языковые явления поддаются описанию с помощью правил-предписаний… Все это заставляет усомниться в универсальности алгоритмического способа мышления и строить деятельностную модель языка на принципе неполной детерминированности» (33). Но как математизировать неполную детерминированность? Опять вспоминается непримиримый враг А. Е. Кибрика и его друзей В. И. Абаев, который, исходя из других теоретических позиций и используя иные термины, пришёл к довольно сходному выводу. Этот языковед не отрицал использования точных методов, в частности, в прикладной лингвистике. Тем не менее, он писал о «математических операциях»: «Во-первых, полезная отдача этих операций как в теоретико-познавательном плане, так и в прикладном в большинстве случаев слишком незначительна по сравнению с затраченным временем и трудом. Во-вторых, — и это главное, — в языкознании… количественные показатели неспособны выявить самое главное — качественное своеобразие явлений. Самое тонкое, самое глубокое, самое человеческое, а потому самое важное в языке остается за пределами применения чисто математических приемов» [Абаев 2006: 121].

Заканчивается статья о постулатах общей оценкой современной лингвистики, которая «еще не осознала своего предназначения», занимаясь «инвентаризациями и классификациями наблюдаемых единиц различных типов» (26). «Такой этап был необходимым звеном в эволюции лингвистики, и проделанный ею труд был поистине титаническим, как титаничен успех младенца, впервые принимающего сидячее положение» (26–27). Мне кажется, что в этих словах Александр Евгеньевич выразил расставание с иллюзиями своей молодости, когда казалось, что цели ясны, а построение полной формализованной модели языка — дело близкого будущего.

Идеи статьи, разумеется, не были только личным достижением А. Е. Кибрика, одновременно к сходным взглядам приходили и его коллеги. Он указывает в начале статьи на ряд отечественных и зарубежных учёных. Но если говорить о предшественниках, то первым должен быть назван Э. Сепир.

Вот, например, выделение функций языка в статье из ЛЭС: «Язык является основной общественно значимой (опосредованной мышлением) формой отражения окружающей человека действительности и самого себя, т. е. формой хранения знаний о действительности (эпистемическая функция), а также средством получения нового знания о действительности (познавательная или когнитивная функция). Эпистемическая функция связывает язык с действительностью…, а познавательная — с мыслительной деятельностью человека… В то же время язык является основным средством человеческого общения (коммуникативная функция), средством передачи информации от говорящего к слушающему (адресату)» (9–10). Несомненно, это классификация Э. Сепира, уточнённая и дополненная в связи с дальнейшим развитием науки. Отмечу также и то, что в эпоху борьбы лингвистов за «имманентность» Э. Сепир подчёркивал связь лингвистики с гуманитарными науками. Необходимо отметить, что под редакцией А. Е. Кибрика вышел том трудов американского учёного [Сепир 1993]. А уже во второй половине 2000 гг. он сказал мне: «Соссюр уже не актуален, его время прошло. А вот Сепир очень современен».

А в целом в нашей науке в эти годы шёл, хотя и по-разному, общий процесс движения в сторону того, что сформулировал А. Е. Кибрик. Могли играть роль и внешние факторы вроде эмиграции того или иного учёного, но и независимо от них движение стало явным. Одни лингвисты заметно эволюционировали в эту сторону (Ю. Д. Апресян, Е. В. Падучева и, безусловно, сам А. Е. Кибрик), другие и раньше придерживались идей, близких к идеям рассматриваемой статьи, но теперь их место в нашей науке стало более значительным (Н. Д. Арутюнова, В. Г. Гак). Прагматика в начале 1970-х гг. нередко вызывала к себе настороженное отношение, однако уже к 1980-м гг. прагматика (как и теория речевых актов, дискурсный анализ, лингвистика текста и др.) стала одним из ведущих направлений. Показательна серия «Новое в зарубежной лингвистике», где поздние выпуски в основном включали в себя переводы работ именно по этой тематике, а во вступительных статьях ведущие советские специалисты анализировали процессы, происходящие в мировой науке. Отход от структурализма при одновременном отказе от следования постулатам Н. Хомского проявился в те годы очень явно. Что касается генеративизма, то после 1972 г.

у нас Хомского долго не переводили, а за генеративной литературой многие просто перестали следить; это проявилось в том, что порождающую семантику (идейно наиболее близкое ответвление генеративизма) долго считали ведущим направлением и после того как она в США потеряла популярность.

Новые подходы к науке о языке и резкое расширение её границ в 1980-е гг. стали достаточно очевидны. Я, например, прочтя её в книжном варианте 1992 г., воспринял её не как откровение или нечто новое, но как концентрированное выражение тех идей, которые мне уже казались несомненными и очевидными. Однако прямо о чертах нового этапа в лингвистике говорили редко, отчасти по экстралингвистическим причинам. Александр Евгеньевич, надо отдать ему должное, едва ли не первым это сделал, причём нашёл удачную форму для этого. Не вступая ни с кем в прямую полемику, не называя имён, он сумел высказать главное. В начале 1980-х гг. не все лингвисты того лагеря, к которому он принадлежал, поддержали его постулаты, но время подтвердило их правильность.

Теперь же структурная лингвистика в духе 1960-х гг. продолжает существовать, но отошла на периферию науки. В мире преобладают два направления: так и не завоевавший у нас господства генеративизм и так называемый функционализм. Последнее направление не образует единства, но его представители в разных странах объединяются некоторыми общими принципами, многие из которых, как мне представляется, были удачно два-три десятка лет назад сформулированы А. Е. Кибриком.

Литература

1. *Абаев* 2006 — Абаев В. И. Статьи по теории и истории языкознания. М., 2006 (первое издание — 1965 г.).
2. *Кибрик* 1992 — Кибрик А. Е. Очерки по общим и прикладным вопросам языкознания. М., МГУ, 1992.
3. *Мельчук* 1997 — Мельчук И. А. Курс общей морфологии. Т. I. Слово. М., 1997.
4. *Сепир* 1993 — Сепир Э. Избранные работы по языкознанию и культурологии. М., 1993.
5. *Трубецкой* 2004 — Письма и заметки Н. С. Трубецкого. М., 2004.

ИСПОЛЬЗОВАНИЕ МЕТОДА УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ ДЛЯ ОБРАБОТКИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Антонова А. Ю. (a-antonova@list.ru)
НИУ Высшая школа экономики, Москва, Россия

Соловьев А. Н. (a.solovyev@i-teco.ru)
ЗАО «Ай-Теко», Москва, Россия

Работа посвящена исследованию метода условных случайных полей (Conditional Random Fields — CRF) на русскоязычных текстах. В частности, продемонстрированы результаты использования CRF в задачах распознавания именованных сущностей, определения частей речи и сентимент-анализа сообщений относительно объекта тональности. Результаты CRF сравниваются с результатами, полученными другими методами.

Ключевые слова: Марковские поля, метод условных случайных полей (CRF), машинное обучение, распознавание именованных сущностей, анализ тональности сообщений, определение частей речи

CONDITIONAL RANDOM FIELD MODELS FOR THE PROCESSING OF RUSSIAN

Antonova A. Y. (a-antonova@list.ru)
Higher School of Economics, Moscow, Russia

Soloviev A. N. (a.solovyev@i-teco.ru)
CJSC I-Teco, Moscow, Russia

The paper aims to illustrate the applicability of conditional random field (CRF) models to Russian texts. Introduced in 2001, CRF method has been successfully exploited and proved its efficiency for a variety of NLP tasks. Its main advantage over HMM is the possibility to model the dependencies and interdependencies in sequential data. Yet this approach has not been widely used for Russian.

Since CRF operates with language-independent features, its initial adaptation for Russian can be minimalistic. We show how CRF models produce state-of-the-art quality for several basic NLP tasks, including named entity recognition, part-of-speech tagging and object-oriented sentiment analysis. We exploited CRF-Suite tool to train and evaluate our models. We used a corpus of news texts for NER and POS-tagging tasks and a subcorpus from Russian Twitter for SA. The results of the evaluation were compared to other existing methods for each of the three tasks.

Key words: Markov fields, conditional random fields (CRF), machine learning, named entity recognition, object oriented sentiment analysis of short messages, part-of-speech tagging

1. Введение

Обработка больших массивов данных поставила перед разработчиками инструментальных систем задачу обеспечения высокой скорости получения результата без существенной потери качества. Так, например, только в русскоязычном сегменте Twitter ежедневно публикуются около 8-10 млн. твитов. Увеличение объема потока текстовых данных привело к тому, что статистические методы стали неотъемлемой частью области text mining.

К числу методов, успешно применяемых в обработке текста, относится метод случайных Марковских полей и его модификация — метод условных случайных полей (CRF — Conditional Random Fields), который нашел широкое применение в лингвистических приложениях, требующих разметки больших объемов текста на основе некоторых параметров. Чаще всего этот метод применяют в задачах распознавания специальных терминов [Finkel at al. 2004, Dingare at al. 2004], именных групп [McCallum 2003, Ratinov 2009], поверхностного синтаксиса (pos-taggers и shallow parsing) [Sha 2003, Sutton 2004] и т. п. Также данный метод находит свое применение в задачах разрешения лексической омонимии [Sutton 2004], анафорических ссылок [McCallum 2005], сентимент-анализе [Choi 2005, Sadamitsu 2008, Mao 2006], машинном переводе [Lavergne 2011]. (Этот ряд можно продолжить набором задач из других предметных областей: биоинформатики, компьютерной графики и пр.)

Метод CRF хорошо исследован для английского, немецкого, арабского, китайского и некоторых других широко распространенных языков. К сожалению, для русского языка этот метод еще не нашел столь широкого применения.

Целью нашей статьи является апробировать возможности CRF применительно к русскому языку на примере определения частей речи, выделения именованных сущностей и сентимент-анализа текста относительно объекта тональности. Важной особенностью нашего исследования являлось отсутствие лингвистической предобработки текста, т. е. анализ проводился на плоском тексте.

2. Описание CRF

Метод CRF (Conditional Random Fields) [Lafferty 2001, Klinger 2007] относится к статистическим лингвистическим методам. Данный метод является одной из возможных реализаций Марковских случайных полей.

Марковским случайным полем или Марковской сетью (Markov random field, Markov network) называют графовую модель, которая используется для представления совместных распределений набора нескольких случайных переменных. Формально Марковское случайное поле состоит из следующих компонентов:

- неориентированный граф или фактор-граф $G = (V, E)$,где каждая вершина $v \in V$ является случайной переменной X и каждое ребро $\{u,v\} \in E$ представляет собой зависимость между случайными величинами u и v.
- набор потенциальных функций (potential function) или факторов $\{\varphi_k\}$, одна для каждой клики в графе G (полный подграф). Функция φ_k ставит

каждому возможному состоянию элементов клики в соответствие некоторое неотрицательное вещественнозначное число.

Вершины, не являющиеся смежными, должны соответствовать условно независимым случайным величинам. Группа смежных вершин образует клику, набор состояний вершин является аргументом соответствующей потенциальной функции.

Совместное распределение набора случайных величин $X=\{x_k\}$ в Марковском случайном поле вычисляется по формуле:

$$P(x) = \frac{1}{z}\prod_k \varphi_k(x_{\{D\}}) \quad (1)$$

где $\varphi_k(x_{\{k\}})$ — потенциальная функция, описывающая состояние случайных величин в k-ой клике; Z — коэффициент нормализации вычисляется по формуле:

$$Z = \sum_{x \in X} \prod_k \varphi_k(x_{\{k\}}) \quad (2)$$

Множество входных лексем $X=\{x_t\}$ и множество соответствующих им типов $Y=\{y_t\}$ в совокупности образуют множество случайных переменных $V=X \cup Y$. Для решения задачи извлечения информации из текста достаточно определить условную вероятность $P(Y \mid X)$. Потенциальная функция имеет вид:

$$\varphi_k(x_{\{k\}}) = \exp\left(\sum_k \lambda_k f_k(y_y, y_{t-1}, x_t)\right) \quad (3)$$

где $\Sigma\{\lambda_k\}$ вещественнозначный параметрический вектор, и $\Sigma\{f_k(y_t, y_{t-1}, x_t)\}$ — набор признаковых функций. Тогда линейным условным случайным полем называется распределение вероятности вида:

$$p(y \mid x) = \frac{1}{z(x)}\prod_k \exp\left(\sum_k \lambda_k f_k(y_y, y_{t-1}, x_t)\right) \quad (4)$$

Коэффициент нормализации тогда Z(x) вычисляется по формуле:

$$Z(x) = \sum_y \prod_k \exp\left(\sum_k \lambda_k f_k(y_y, y_{t-1}, x_t)\right) \quad (5)$$

Метод CRF, как и метод MEMM (Maximum Entropy Markov Models), относится к дискриминативным вероятностным методам, в отличие от генеративных методов, таких как HMM (Hidden Markov Models) [Rabiner 1989, Николенко 2006] или метод «Наивного Баеса» (Naïve Bayes [McCallum 1998]).

По аналогии с MEMM [Bishop 2006, McCallum 2000], выбор факторов-признаков для задания вероятности перехода между состояниями при наличии наблюдаемого значения x_t зависит от специфики конкретных данных, но в отличии от того же MEMM, CRF может учитывать любые особенности и взаимозависимости в исходных данных. Вектор признаков $L=\{\lambda_k\}$ рассчитывается

на основе обучающей выборки и определяет вес каждой потенциальной функции. Для обучения и применения модели используются алгоритмы, аналогичные алгоритмам HMM: Витерби и его разновидность — алгоритм «вперед-назад» (forward-backward) [Sutton 2008].

Как показано в [Sutton 2006], скрытую Марковскую модель можно рассматривать как частный случай линейного условного случайного поля (CRF). В свою очередь, условное случайное поле можно рассматривать как разновидность Марковского случайного поля (см. рис. 1).

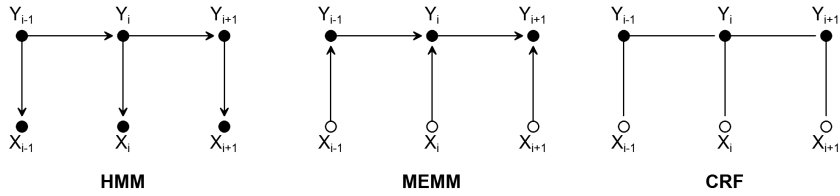


Рис. 1. *Изображение в виде графов для методов HMM, MEMM и CRF. Незакрашенные окружности обозначают, что распределение случайной величины не учитывается в модели. Стрелки указывают на зависимые узлы*

В условных случайных полях отсутствует т. н. label bias problem — ситуация, когда преимущество получают состояния с меньшим количеством переходов [Lafferty at al. 2001], так как строится единое распределение вероятностей и нормализация (коэффициент Z(x) из формулы (5)) производится в целом, а не в рамках отдельного состояния. Это, безусловно, является преимуществом метода: алгоритм не требует предположения независимости наблюдаемых переменных. Кроме того, использование произвольных факторов позволяет описать различные признаки определяемых объектов, что снижает требования к полноте обучающей выборки. Недостатком подхода CRF является вычислительная сложность анализа обучающей выборки, что затрудняет постоянное обновление модели при поступлении новых обучающих данных.

На сегодняшний день именно метод CRF является наиболее популярным и точным способом извлечения объектов из текста [Sarawagi 2008]. Например, он был реализован в проекте Стэндфордского университета Stanford Named Entity Recognizer [Stanford NER]. С той же целью этот метод с этого года успешно применяется в системах «X-Files» и «Аналитический Курьер» [Киселев 2007].

Как уже было отмечено, качество выделения сущностей методом, основанном на статистическом подходе, определяется полнотой обучающей выборки. Для этого необходимы размеченные корпуса, содержащие не только лексику, характеризующую данную предметную область, но и статистически значимые N-граммы, наиболее полно покрывающие произвольный текст. В отличии от других статистических методов, метод CRF, основанный на фактор-графах, требует гораздо меньшей обучающей выборки, поскольку статистически значимые сочетания могут быть определены как набор клик для исследуемого

объекта. В зависимости от решаемой задачи, на практике достаточно объема от нескольких сотен тысяч до миллионов термов. При этом точность будет определяться не только объемом выборки, но и выбранными факторами.

3. Эксперимент

Для экспериментов был выбран проект с открытым кодом http://www.chokkan.org/software/crfsuite/. Его преимущества относительно других открытых проектов (CRF++; HCRF; CRF package; набор утилит, основанных на алгоритме CRF Стенфордского Университета и др.) заключаются в его простоте, универсальности и скорости работы. Этот алгоритм универсален для любых видов классификационных задач, при этом включает в себя набор разных оптимизационных обучающих методов:

- Limited-memory Quasi-Newton method (LBFGS) [Andrew 2007] — квазиньютоновский метод с ограничением памяти BFGS [Nocedal 1980].
- Stochastic Gradient Descent (L2SGD) [Shalev-Shwartz 2007] — метод градиентного спуска.
- Averaged Perceptron (AP) [Collins 2002] — метод усредненного персептрона[1].
- Passive Aggressive (PA) [Crammer 2006] — алгоритм, основанный на бинарной классификации.
- Adaptive Regularization Of Weight Vector (AROW) [Mejer 2010] — метод адаптивной регуляризации весов вектора.

Все эксперименты проводились с различными методами оптимизации, из которых выбирался наилучший. Используемый алгоритм относится к линейным методам CRF.

3.1. Определение именованных сущностей (NER)

Метод CRF был использован в задаче распознавания именованных сущностей — NER (Name Entity Recognition). Для этого вручную был размечен небольшой корпус из новостных лент СМИ объемом более 71 тыс. предложений на самую разную тематику (более 1,5 млн. словоформ). Каждое предложение содержало хотя бы одну именную сущность. Для проведения тестов было выбрано пять типов сущностей (Табл. 1): физические лица (имена, фамилии, отчества), юридические лица (названия организаций, компаний и пр.), географические объекты (названия городов, улиц, рек и пр.), продукты (названия продуктов, включая марку и бренд) и события (названия форумов, съездов, мероприятий и пр.).

---

[1] Метод усредненного персептрона, как и алгоритм Passive Aggressive, часто используются в качестве самостоятельных методов в решении ряда лингвистических задач (например, классификации). В методе CRF эти алгоритмы используются только для оптимизации потенциальных функций (см. уравнение (3)).

Табл. 1. Частотное распределение именованных сущностей в размеченном корпусе. NAME — физ.лица, ORG — юр.лица, GEO — географические названия, PROD — продукты, EVENT — события

Тип сущности	Абс. частота в корпусе	Отн. частота в корпусе, %
NAME	37 729	24,99
ORG	43 646	28,91
GEO	52 889	35,04
PROD	6 425	4,26
EVENT	10 263	6,80

Для обучения и тестирования с кросс-валидацией полученный корпус был разделен на 4 части: три части использовались для обучения и одна для тестов. На одном и том же корпусе обучались модели с использованием каждого из перечисленных оптимизационных методов с различными экспертно задаваемыми параметрами.

В качестве факторов были выбраны только n-граммы (длины от двух до пяти) и графематические особенности написания именованных сущностей. Например, все заглавные буквы без точек (*МТС*), одна заглавная буква с точкой (*W.*), первые строчные с точкой, затем заглавная (*ул.Мира* — при слитном написании) и т. д. Всего было определено 14 параметров.

3.2. Сентимент-анализ

Современный сентимент-анализ текста включает в себя по крайней мере три вида задач [Liu 2010]:

1. Классификация тональных сообщений (позитив/негатив или более тонкая градация);
2. Определение сентимента относительно заданного объекта тональности (ОТ) (часто с последующей визуальной разметкой дерева зависимостей предложения, например, «***Правительство*** *одобрило новый указ президента, нарушив конституцию*», тут ОТ — «*Правительство*», у которого два противоположных сентимента);
3. Определение сентимента объекта тональности относительно его имплицитных и эксплицитных атрибутов (feature-based). Например, «У этого телефона большой аккумулятор, правда и вес немаленький», тут ОТ — *телефон*, а его атрибуты — *вес* и *аккумулятор* — имеют разную полярность.

В данном исследовании мы ограничимся второй областью задач сентимент-анализа, а именно: будем определять сентимент заданного объекта тональности в произвольном сообщении.

Для этого по заданным объектам (числом более 20, например, *МГУ, РЖД, радио, пальто, кино* и пр.) нами был собран корпус коротких твит-сообщений и экспертно размечен. Всего в 20 тысячах твитах были отмечены около 21 тысячи ОТ.

Для разметки слов использовались тональные словари, которые были собраны при разработке метода сентимент-анализа, основанного на правилах (описание и список словарей см. в [Pazelskaya 2011, Solovyev 2012]). К данным словарям были добавлены словари инверторов и шифтеров. Число всех словарей составило 34. Тональные словари были дополнены полными наборами словоформ (т. к. мы используем плоский текст без предобработки). Таким образом, общее число словарных форм получилось более 20 тыс., а полное число всех слов — более 400 тыс.

Вхождение слова или словосочетания в тональные словари определенного типа рассматривалось в качестве факторов CRF, сами слова не учитывались. Слово, не получавшее значения, исключалось из анализа, при этом знаки пунктуации сохранялись с нулевым весом. Ширина окна анализа составила 2–5 грамм. Классификация производилась на три класса: позитивный, негативный и нейтральный (Табл. 2).

Табл. 2. Частотное распределение в корпусе объектов тональности (ОТ) по классам тональности

Класс	Кол-во ОТ	%
Позитивный	6435	31,08
Негативный	6034	29,14
Нейтральный	8236	39,78

3.3. Определение частей речи (POS-tagger)

Цель данного раздела — показать, что метод CRF, примененный в задаче определения частей речи (частеречного тэгирования), дает результаты не ниже уровня, достигаемого в данный момент статистическими системами на материале русского языка. В этом случае мы можем говорить, что CRF метод не хуже или даже лучше других, ранее разработанных методов. Полученный результат мы будем сравнивать с результатами статистических систем. (Словарные системы представляют собой другой подход.)

Задачу определения части речи, традиционно рассматриваемую в прикладных системах как задачу классификации, решают с помощью 1) метода опорных векторов (SVM [Giménez 2004]), адаптированный для многозначной классификации,[2]) HMM [Brants 2000] и ряда других (AP, ME и т. п.) методов.

Для европейских языков (и прежде всего английского) статистические тэггеры уже давно [Brants 2000] приблизились и преодолели барьер в 97 % [Manning 2011]. В случае русского языка ситуация иная. Почти все участники соревнования морфологических парсеров, прошедшего в 2010-м году

[2] http://www.lsi.upc.edu/~nlp/SVMTool/

[Ляшевская и др. 2010], представляли инструменты, в которых работали алгоритмы, основанные на правилах. По дорожке PoS-tagging был достигнут результат, близкий к абсолютному: 99,4 % правильно определенных частей речи. Что касается, уровня качества статистических систем, в работе [Sharoff, Nivre 2011] описывается тэггер, обученный на корпусе SinTagRus, показавший результат 97 %.

В настоящей работе мы проводим сравнение между полученным CRF-классификатором, тэггером разработки Стенфордского университета и TreeTagger'ом[3], которые являются свободно распространяемыми как в виде приложений, так и в виде исходного кода.

Размер обучающего корпуса 2.2 млн. словоформ, тестового 670 тыс. словоформ. Как и в случае задачи выделения сущностей, основную часть обоих корпусов составили сообщения новостных лент на разную тематику. Разметка корпуса производилась с помощью морфологического модуля системы «Аналитический курьер» [Киселев 2007], наиболее частотные случаи омонимии размечались вручную. Список выделяемых частей речи был ограничен возможностями системы (так, например, совсем не выделялись частицы и междометия — и те, и другие попали в категорию «прочее» и были обозначены в разметке одинаковым тэгом).

4. Результаты

4.1. NER

Результаты тестирования NER представлены в Табл. 3–5. Лучшие результаты показывали методы оптимизации Averaged Perceptron и Passive Aggressive на триграммной модели. Измерения проводились как по пяти, так и по трем типам сущностей (Табл. 3 и 4).

Точность и полнота рассчитывались по следующим формулам:

(6) $$Precision = \frac{A}{A + C + D} * 100\%$$

(7) $$Recall = \frac{A}{A + B + C} * 100\%$$

Здесь A — количество верных срабатываний системы;
B — количество пропусков;
C — количество случаев типизации нетипизированной сущности;
D — тип сущности определен неверно.

Этот же принцип оценивания использовался и для задач, разбираемых ниже.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Табл. 3. Оценка качества определения пяти типов сущностей с разными методами оптимизации

	NAME	GEO	ORG	PROD	EVENT	Среднее
Точность						
AROW	91,36	89,57	83,49	75,08	77,61	83,43
L2SGD	91,19	89,52	83,88	70,14	75,69	82,09
LBFGS	91,39	89,53	84,25	69,95	76,05	82,23
PA	92,79	91,26	85,61	75,01	79,22	84,78
AP	92,58	91,22	85,23	76,09	79,43	**84,91**
Полнота						
AROW	93,85	94,05	86,48	84,02	80,65	87,81
L2SGD	93,61	94,22	87,23	84,28	83,27	88,61
LBFGS	93,57	94,34	87,10	83,39	83,32	88,34
PA	94,17	95,08	87,92	86,5	83,23	89,38
AP	94,23	95,01	88,02	87,14	83,84	**89,65**
F1						
AROW	93,60	91,75	84,96	79,30	79,10	85,54
L2SGD	92,38	91,81	85,53	76,71	79,34	85,15
LBFGS	92,47	91,87	85,65	76,08	79,52	85,12
PA	93,48	93,13	86,75	80,34	81,17	86,98
AP	93,39	93,08	86,60	81,24	81,57	**87,18**

Табл. 4. Оценка качества определения трех типов сущностей (результаты показаны только для двух методов оптимизации)

	NAME	GEO	ORG	Среднее
Точность, %				
AP	92,52	91,39	85,02	89,64
PA	93,05	91,47	85,5	90,01
Полнота, %				
AP	93,48	93,99	82,38	89,95
PA	93,25	93,96	82,23	89,81
F1, %				
AP	93,00	92,67	83,68	89,78
PA	93,15	92,7	83,83	89,89

Несколько меньшие значения полноты и точности для продуктов и событий объясняются недостаточной сбалансированностью исходного корпуса: продуктов и событий в текстах встречается гораздо меньше, чем геообъектов, физических и юридических лиц (см. Табл. 1).

Несмотря на то, что важным пунктом нашего исследования было какое-либо отсутствие предобработки текста, мы провели тест с нормализованным текстом, в котором все слова приведены к словарной форме. Нормализация в среднем не меняет результаты (см. Табл. 5).

Табл. 5. Средние значения полноты и точности при определении трех и пяти типов сущностей при нормализации слов

	Точность, %	**Полнота, %**	**F1, %**
Пять типов	84,76	89,67	87,06
Три типа	90,52	89,97	90,22

Результаты, полученные с помощью метода CRF, сравнивались с другими методами на том же корпусе (см. Табл. 6 и 7): MEMM и обычным словарным методом (обозначен Dict). MEMM был выбран, поскольку, как и CRF, принадлежит к дискриминативным методам, в отличие от генеративного HMM. Словарь для словарного метода был получен из обучающей выборки. Расчеты проводились как для трех, так и для пяти типов сущностей.

Табл. 6. Сравнительные результаты полноты и точности при определении пяти типов сущностей, полученные различными методами

	Dict	**MEMM**	**CRF**
Точность, %	90,35	89,08	84,91
Полнота, %	46,64	72,63	89,65
F1, %	59,39	79,89	**87,18**

Табл. 7. Сравнительные результаты полноты и точности при определении трех типов сущностей, полученные различными методами

	Dict	**MEMM**	**CRF**
Точность, %	94,04	93,26	90,01
Полнота, %	55,37	75,80	89,81
F1, %	69,60	83,59	**89,89**

Как и следовало ожидать, простая словарная разметка дает высокую точность (около 90%) при низкой полноте (47%). Следует заметить, что точность словарного метода не достигает 100%, т. к. одна и та же сущность может относиться к разным типам. Например, в предложении «*Кремль дал понять Киеву, что...*» *Кремль* и *Киев* выполняют функцию юридического лица, но не географического названия.

Наилучшие результаты для метода MEMM показал метод оптимизации Averaged Perceptron. Как видно из таблиц, результаты сравнительных методов уступают по F1-мере методу CRF.

4.2. Сентимент-анализ

Результаты тестирования показали, что наилучшую точность показывают, опять же, оптимизационные методы Averaged Perceptron и Passive Aggressive, причем для трех и четырех-граммных построений результаты получились примерно одинаковыми с небольшим перевесом у четырех-граммного окна. В Табл. 8 приведена точность полученной четырех-граммной модели для метода AP (Полнота всюду составила 100%, поскольку объекты тональности были заданы заранее.)

Табл. 8. Точность определения класса объекта тональности методом CRF

Класс	Точность, %
Негативный	84,44
Позитивный	84,89
Нейтральный	90,93
Среднее	**86,75**

Несмотря на то, что самым популярным статистическим методом сентимент-анализа текста является метод SVM [Liu 2010], мы для сравнения результатов тестирования использовали метод, основанный на правилах. На это у нас было две причины:

1. SVM классифицирует документы (предложения) относительно их полярности без какой-либо привязки к объекту тональности. В нашем алгоритме предполагается, что полярность жестко связана именно с объектом тональности (объект тональности был одним из факторов CRF[4]).
2. в методе, основанном на правилах, не только задается объект, относительно которого определяется тональность, но и используется тот же набор тональных словарей и шифтеров, *придающих лингвистический смысл результату.*

Сравнение с результатами ручного тестирования метода сентимент-анализа, основанного на правилах (Табл. 9) [Pazelskaya 2011], показывают более высокую степень точности CRF.

---

[4] В принципе, метод SVM позволяет сделать привязку к объекту мониторинга, но это скорее искусственный прием, связанный с особой компоновкой векторов, а не лексико-семантическими характеристиками тональности.

Табл. 9. Качество сентимент-анализа, основанного на правилах

	СМИ	Блогосфера
Точность, %	86,66	80,42
Полнота, %	82,37	68,88
F1, %	84,46	**74,20**

Это может быть вызвано разными причинами, одна из которых та, что обучающий корпус составлялся экспертами вручную, поэтому лишен разного рода «кривых» текстов с грубыми нарушениями грамматики и синтаксиса, в то время как модель сентимент-анализа на правилах тестируется на реальных текстах. Другой причиной повышения качества мог послужить удобный формат твитов: короткие сообщения не более 140 символов не имеющие сложной иерархической системы, как, например, блоги или форумы.

4.3. Part-of-Speech tagger

Для POS-tagger'а использовались те же оптимизационные алгоритмы, которые перечислены в разделе 3. Поскольку мы не ставили себе целью исследовать качество каждого из этих алгоритмов, то приведем только наилучший результат (Табл. 10). Он был получен с помощью алгоритма Passive Aggressive (вторым по качеству был метод Averaged Perceptron) на триграммах, с использованием «хвоста» слова длины три и графематических характеристик, упоминаемых в п. 3.1. Отметим, что упомянутый в [Manning 2011] барьер в 97% правильно классифицированных частей речи в нашем случае практически достигается: 96,7% (см. Табл. 12).

Табл. 10. Оценка результата частеречной классификации

	Точность, %	Полнота, %	F1, %
Биграммы, иных параметров нет	90,70	86,66	88,22
Триграммы, иных параметров нет	88,14	85,99	86,91
Триграммы + «хвосты» длины 3	93,79	92,47	93,10
«Хвосты» длины 3, без n-грамм	76,32	73,29	74,28
Триграммы + «хвосты» длины 3 + графематические характеристики	94,95	93,43	**94,14**

Для наилучшего случая приводится таблица (Табл. 11) с показателями качества классификации для каждого выделяемого частеречного класса.

Табл. 11. Результаты классификации для каждой выделяемой части речи

Часть речи	Относительная встречаемость частеречного класса, %	Точность, %	Полнота, %	F1, %
Существительное	30,42	96,03	96,98	96,50
Прилагательное	9,40	92,45	92,16	92,30
Глагол	9,12	98,32	98,86	98,59
Причастие	0,76	82,37	82,58	82,48
Деепричастие	0,24	94,80	90,11	92,40
Наречие	4,17	96,43	96,07	96,25
Предлог	9,83	99,39	99,61	99,50
Союз	5,92	99,40	99,54	99,47
Числительное	0,64	90,27	89,22	89,74
Числительное, записанное цифрами	1,56	92,80	94,78	93,78
Местоимение-существительное (личное)	1,20	99,31	99,84	99,57
Остальные местоимения	3,65	98,89	98,68	98,78
Сокращение	0,35	96,69	82,23	88,88
Знак препинания	17,54	99,97	99,88	99,93
«Остальное»	4,66	84,68	79,35	81,93

Чтобы сопоставить качество CRF-тэггера, были использованы два другие инструмента. Основу Стенфордского тэггера[5] составляет метод максимальной энтропии [Toutanova 2000, Manning 2011]. В свою очередь, инструмент TreeTagger (представленный еще в [Schmid 1994]) использует марковские модели и деревья решений для оценки вероятности перехода между состояниями. Обученная модель для русского языка была получена Сергеем Шаровым и находится в открытом доступе[6].

В таблице 12 приводится сравнения качества методов[7]. Под Accuracy в данном случае понимается процент правильно классифицированных слов от объема тестового корпуса.

---

5 http://nlp.stanford.edu/software/tagger.shtml

6 http://corpus.leeds.ac.uk/mocky/russian.par.gz

7 Трудность сопоставления результатов состояла в том, что для TreeTagger'а мы использовали заранее обученные модели с заданными частеречными классами, которые не полностью совпали с классами нашей разметки. Таким образом, в сравнительную таблицу попали только результаты по совпавшим классам. В случае Стенфордского тэггера проверялось пересечение по всем классам, поскольку Стенфордский тэггер обучался на том же корпусе, что и CRF-тэггер.

Табл. 12. Значение F1-меры для трех систем

	Accuracy, %
Stanford	79,39
TreeTagger	93,33
CRF	**96,75**

5. Выводы

В статье показана применимость CRF-метода в обработке текста на русском языке на примере задач выделения именнованных сущностей, сентимент-анализа коротких высказываний, частеречной классификации. Полученные результаты сравниваются с данными, полученными с помощью других подходов к рассматриваемым задачам. Как показывают результаты тестов, метод условных случайных полей (CRF) может составить существенную конкуренцию другим статистическим методам, используемым при лингвистической обработке текста.

Литература

1. *Bishop. Ch.* Pattern Recognition and Machine Learning. Springer. 2006.
2. *Brants, Th.* 2000. TnT — A Statistical Part-of-Speech Tagger. «6th Applied Natural Language Processing Conference»
3. *Choi Y., Cardie Cl., Riloff E., Patwardhan S.* Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. HLT ‹05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Pages 355–362. 2005.
4. *Collins M.* "Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). 1–8. 2002.
5. *Dingare Sh., Finkel J., Nissim M., Manning Ch., Grover C.* A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations. Comparative and Functional Genomics, Volume 6 (2005). Issue 1–2. Pages 77–85. 2004.
6. *Giménez J., Márquez L.* 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC›04). Lisbon, Portugal.
7. *Finkel J., Dingare Sh., Manning Ch.D., Nissim M., Alex B., Grover C.* Exploring deep knowledge resources in biomedical name recognition. JNLPBA 04 Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Pages 96–99. 2004.
8. *Galen A. and Jianfeng G.* "Scalable training of L1-regularized log-linear models". Proceedings of the 24th International Conference on Machine Learning (ICML 2007). 33–40. 2007.
9. *Klinger R., Tomanek K.* Classical Probabilistic Models and Conditional Random Fields. Algorithm Engineering Report TR07–2-013. Department of Computer Science. Dortmund University of Technology. December 2007. ISSN 1864–4503.
10. *Koby C., Ofer D., Joseph K., Sh. Shalev-Shwartz. and Singer.* "Yoram Online Passive-Aggressive Algorithms". Journal of Machine Learning Research. 7. Mar. Pages 551–585. 2006.
11. *Lafferty J., McCallum A., Pereira F.* "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". Proceedings of the 18th International Conference on Machine Learning. Pages 282–289. 2001.
12. *Lavergne T., Allauzen A., Yvon F.* "From n-gram based to CRF-based translation models" EMNLP'11, 6th workshop on statistical machine translation (WMT'11), Edinburgh, UK, July 2011
13. *Liu Bing.* "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing. Second Edition. (editors: N. Indurkhya and F. J. Damerau). 2010.
14. *Manning C. D.* 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander Gelbukh (ed.), Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608, pp. 171–189. Springer.

15. *Mao Yi., Lebanon G.* Isotonic Conditional Random Fields and Local Sentiment Flow. In proceeding of: Advances in Neural Information Processing Systems 19. Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems. Vancouver. British Columbia. Canada. December 4–7. 2006
16. *McCallum A., Li W.* Early results for named entity recognition with conditional random fields. feature induction and web-enhanced lexicons. In Seventh Conference on Natural Language Learning (CoNLL). 2003.
17. *McCallum A. and Nigam K.* "A Comparison of Event Models for Naive Bayes Text Classification". In AAAI/ICML-98 Workshop on Learning for Text Categorization. pp. 41–48. Technical Report WS-98-05. AAAI Press. 1998.
18. *McCallum A., Wellner B.* Conditional models of identity uncertainty with application to noun coreference. In Lawrence K. Saul. Y. Weiss. and L. Bottou. editors. Advances in Neural Information Processing Systems 17. Pages 905–912. MIT Press. Cambridge. MA. 2005.
19. *McCallum A., Freitag D. and Pereira F.* "Maximum entropy markov models for information extraction and segmentation." ICML-2000. pp. 591–599.
20. *Mejer A. and Crammer K.* "Confidence in Structured-Prediction using Confidence-Weighted Models". Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010). Pages 971–981. 2010.
21. *Nocedal.* "Updating Quasi-Newton Matrices with Limited Storage". Mathematics of Computation Jorge. 35. 151. Pages 773–782. 1980.
22. *Pazelskaya A., Solovyev A.* A Method of Sentiment analysis of Russian Text." Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011". Bekasovo. 2011. pp. 510–523.
23. *Rabiner, L. R.* A Tutorial on Hidden Markov Models andSelected Applications in Speech Recognition. In: Proceedings of the IEEE 77 (1989). No. 2. pp. 257–286.
24. *Ratinov L., Roth D.* Design Challenges and Misconceptions in Named Entity Recognition. CoNLL'09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Pages 147–155. 2009.
25. *Sadamitsu K., Sekine S., Yamamoto M.* Sentiment analysis based on probabilistic models using inter-sentence information. International Conference on Language Resources and Evaluation. 2008.
26. *Sarawagi S.* Information extraction. Foundations and Trends in Databases. 1(3). 2008.
27. *Schmid H.* Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.
28. *Sha F., Pereira F.* Shallow Parsing with Conditional Random Fields. NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology — Volume 1, pp. 134–141. 2003.
29. *Shalev-Shwartz Sh., Singer Y., and Srebro N.* "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM". Proceedings of the 24th International Conference on Machine Learning (ICML 2007). Pages 807–814. 2007.

30. *Serge Sharoff, Joakim Nivre,* (2011) The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Dialog 2011.
31. *Singla P., Domingos P.* Discriminative training of Markov logic networks. In Proceedings of the Twentieth National Conference on Artificial Intelligence, pp. 868–873. Pittsburgh. PA. 2005. AAAI Press.
32. *Solovyev A. N., Antonova A. Ju., Pazelskaya A. G.* Using Sentiment-Analysis for Text Information Extraction. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012". Bekasovo. 2012. pp. 616–627.
33. *Stanford Named Entity Recognizer* // http://www-nlp.stanford.edu/software/CRF-NER.shtml
34. *Sutton C.* Conditional probabilistic context-free grammars. Master's thesis. University of Massachusetts. 2004.
35. *Sutton C., McCallum A.* Introduction to Conditional Random Fields for Relational Learning. MIT Press. 2006.
36. *K. Toutanova and C. D. Manning. 2000.* Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63–70
37. *Kiselev S.* Sistemy «Analiticheskij kur'er» i X-Files — osnova tehnologii izvlechenija znanij tekstov iz proizvol'nyh istochnikov ["Analytical Courier" and X-Files Systems — mining data from various sources] // Biznes i bezopasnost' v Rossii [Business and Security in Russia]. 2007. — № 48. c. 102–106.
38. *Ljashevskaja O., Astaf'eva I., Bonch-Osmolovskaja A., Garejshina A., Grishina Ju., D'jachkov V., Ionov M., Koroleva A., Kudrinskij M., Litjagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., Koval' S.* NLP Evaluation: Russian Morphological Parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskie parsery russkogo jazyka] // Komp›juternaja lingvistika i intellektual'nye tehnologii [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"] (Bekasovo, May 26–30, 2010). Moscow, RSUH, 2010
39. *Nikolenko S.* Skrytye markovskie modeli [Hidden Markov Models]. ITMO, 2006.

СЕМАНТИКА ЭМОЦИОНАЛЬНЫХ КАУЗАТИВОВ: СТАТУС КАУЗАТИВНОГО КОМПОНЕНТА[1]

Апресян В. Ю. (vapresyan@hse.ru)

НИУ Высшая школа экономики, Москва, Россия

В статье рассматривается взаимодействие отрицания с эмоциональными каузативами. Определяется семантическая специфика эмоциональных каузативов по сравнению с прочими каузативами, в частности, их отношения с декаузативами. Устанавливаются факторы, влияющие на возможность отрицания каузативного компонента, а именно, несовершенный вид глагола, родовой референтный статус именной группы со значением причины, агентивность и конативность каузатива. Устанавливается и мотивируется связь между конативностью vs. ее отсутствием у агентивных каузативов и видом и статусом каузативного компонента в их значении (более близкий к ассертивному для конативов, более близкий к пресуппозитивному — для не-конативных). Устанавливается связь между типом эмоции и характерным для нее типом агентивных каузативов (конативные vs. не-конативные). Агентивные конативные каузативы характерны для базовых эмоций — страх, гнев, радость, грусть, удивление, стыд. Для разных типов эмоциональных каузативов постулируются разные каузативные компоненты, разные семантические структуры и разные аспектуальные свойства (отнесенность к третьему vs. четвертому типу видового противопоставления). Наибольшая степень конативности и, соответственно, ассертивный статус каузативного компонента свойственны трем эмоциональным каузативам — *злить*, *веселить* и *пугать*.

Ключевые слова: каузатив, декаузатив, агентивный, конативный, пресуппозиция, ассерция, семантическая структура, базовые эмоции

[1] Исследование осуществлено в рамках Программы «Научный фонд НИУ ВШЭ» в 2013–2014 гг., проект № 12-01-0102. Оно также поддерживалось Программой фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики», грантом РГНФ No. 10-04-00273а и грантом НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами РФ.

SEMANTICS OF EMOTION CAUSATIVES: THE STATUS OF THE CAUSATIVE COMPONENT

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

NRU Higher School of Economics, Moscow, Russia

The paper considers semantic structure of emotion causatives and their interaction with negation, namely, its narrow or wide scope. Emotion causatives are defined as a group of causatives with their specific semantic properties that distinguish them from other groups of causatives. One of those properties concerns their relation with corresponding decausatives, which, unlike causatives, do not license wide scope of negation. There are several factors that enable negation to have scope over the causative element in emotion causatives — their imperfective aspect, generic referential status of the causative NP phrase, agentivity and conativity of the causative. Non-agentive causatives never license the negation of the causative component. Agentive conative causatives license the negation of the causative component more frequently and easily than agentive non-conative causatives, prompting the assumption that in their semantic structures the causative component has different statuses (assertion in the former, presupposition in the latter). It also has different forms for conatives and non-conatives. Conativity vs. non-conativity of emotion causatives is related to the emotion type, with conative synthetic causatives being limited to basic emotions. The greatest degree of conativity and, hence, the assertive status of the causative component characterizes three emotion causatives — *zlit'* 'to make mad', *veselit'* 'to cheer up', and *pugat'* 'to frighten'.

Key words: causative, decausative, agentive, conative, intentional, presupposition, assertion, semantic structure, basic emotions

1. Введение

Статья посвящена семантическому взаимодействию отрицания и других семантически активных элементов с глаголами со значением каузации эмоции. Данные об этом взаимодействии позволяют уточнить представления об их семантической структуре, в частности, характере каузативного компонента, а также о пресуппозитивном vs. ассертивном статусе каузативного компонента у разных типов эмоциональных каузативов.

Возможность попадания пресуппозиции в сферу действия отрицания у разных типов предикатов рассматривались на примере агентивных каузативов физического воздействия [Карловска 1990]; эмоциональных каузативов [Кустова 1996]; сочетаний глаголов с некоторыми адвербиалами [Падучева 2005]; интерпретативов [Апресян 2006], [В. Апресян 1010]; глаголов-эмоциональных состояний в императиве [Апресян 2012].

Необходимость выделения эмоциональных каузативов в отдельный класс предикатов связана с тем, что они имеют ряд семантических особенностей, в том числе в отношении взаимодействия с отрицанием, которые отличают их от прочих типов каузативов. Одна из этих особенностей — соотношение каузативов и декаузативов на *-ся*. У глаголов со значением физических воздействий возможность разных сфер действия отрицания, имеющаяся у каузатива, часто наследуется декаузативом:

(1) а. *Представим, что князь Мышкин на вечере у Епанчиных не разбил вазу*
'Князь Мышкин не уронил вазу' — широкая СД отрицания
'Князь Мышкин уронил вазу, но не разбил ее' — узкая СД отрицания
б. *Китайская ваза не разбилась*
'Ваза не упала' — широкая СД отрицания
'Ваза упала, но не разбилась' — узкая СД отрицания

Эмоциональные декаузативы отличаются в этом отношении от декаузативов физического воздействия, а именно, первые, в отличие от последних, не наследуют от соответствующих каузативов возможности широкой сферы действия отрицания:

(2) *Он не обрадовал маму приездом*
'Он не приехал'
'Он приехал, но мама не обрадовалась'

(3) *Мама не обрадовалась его приезду*
*'Он не приехал ' — интерпретация с широкой СД отрицания невозможна
'Он приехал, но мама не обрадовалась'

Это связано с тем, что в семантическом классе глаголов со значением эмоций семантически исходными являются морфологически более сложные образования на *-ся (расстроиться, обрадоваться, удивиться)* со значением эмоциональных состояний, а семантически производными — морфологически более простые каузативы *(расстроить, обрадовать, удивить)*. В семантической структуре глаголов со значением эмоционального состояния есть указание на причину, которое не может сниматься. Поэтому интерпретация с широкой СД отрицания для эмоциональных глаголов-состояний на *-ся* невозможна. Как справедливо отмечается в работе [Падучева 2005], «в классе психологических глаголов… прибавление *-ся* дает не декаузативацию, а классический диатетический сдвиг».

Однако и сами эмоциональные каузативы неоднородны с точки зрения взаимодействия с отрицанием; как показывает данное исследование, внутри класса эмоциональных каузативов с этой точки зрения выделяются разные подклассы глаголов. У некоторых эмоциональных каузативов отрицание каузативного компонента возможно, у некоторых оно затруднено. Это указывает на разные статусы каузативного компонента в их семантической структуре.

Помимо семантической структуры каузативов, на их интерпретацию с отрицанием влияют и другие семантические и грамматические факторы, которые также рассматриваются в работе.

2. Эмоциональные каузативы и их взаимоотношение с отрицанием

Общую семантическую структуру эмоциональных каузативов можно описать следующим образом:

(4) ‘каузация эмоции’
‘эмоциональная реакция’

Поскольку в принципе для глаголов со значением эмоции, в том числе для каузативов, указание на причину очень важно, оно обычно сохраняется при отрицании, т. е. отрицание с эмоциональными каузативами чаще имеет узкую сферу действия. Соответственно, уместно говорить о том, что каузативный компонент у них чаще имеет статус пресуппозиции:

(5) *Он не удивил собравшихся своей странной выходкой — все уже давно привыкли к его эксцентричности*
‘он совершил выходку’ — пресуппозиция
‘она не удивила собравшихся’ — ассерция

Однако существуют факторы, которые способствуют попаданию каузативного компонента в СД отрицания.

2.1. Факторы, влияющие на интерпретацию отрицания с эмоциональными каузативами

В число факторов, влияющих на интерпретацию отрицания в сочетании с эмоциональными каузативами, входят следующие: тип каузатива; грамматические признаки глагола (вид, наклонение); тип конструкции; тип эмоции.

2.2. Тип каузатива

Основным, хотя и тривиальным, семантическим признаком, определяющим СД отрицания с эмоциональными каузативами, является тип каузатива. Широкая СД отрицания возможна только для агентивных каузативов, то есть для таких, где в роли каузатора выступает человек, а не свойства, поступки или события; ср. широкую СД отрицания в (6а) и ее невозможность в (6б):

(6) а. *Нет, её не отталкивали, не обижали недоверием* (М. Семенова).
‘К ней не проявляли недоверие => Она не обижалась’
Каузатор — человек
б. *Недоверие ее не [обижало]*
‘К ней проявляли недоверие, но она не обижалась’
Каузатор — отношение

Интересно, что это свойство агентивных эмоциональных каузативов никоим образом не фиксируется лексикографически: словари не разграничивают те употребления, где каузатором является агенс (человек), и те употребления, где каузатор не-агентивен (события, свойства, поступки). При этом агентивное и неагентивное употребления различаются своими свойствами. Помимо различия в интерпретации отрицания, агентивные и неагентивные эмоциональные каузативы различаются синтаксической сочетаемостью. В частности, в отличие от агентивных, неагентивные эмоциональные каузативы не употребляются в длительной конструкции или с обстоятельствами, указывающими на длительность:

(7) а. *Он два часа <долго> злил ее своими выходками*
б. **Его выходки два часа <долго> злили ее*

Для неагентивных эмоциональных каузативов также невозможно употребление в конативных конструкциях, например, в конструкциях с удвоением глагола, описанных в работе [Плунгян, Рахилина 2010]:

(8) а. *Он ее злил, злил и разозлил*
б. **Его выходки ее злили, злили и разозлили*

Таким образом, имеются основания для выделения у эмоциональных каузативов двух разных типов значений: агентивного и неагентивного, с разными семантическими структурами:

(9) а. ‘Человек А1 совершает действия А3, которые вызывают эмоцию X у человека А2’ — агентивные каузативы
б. ‘Действия, события или свойства А1 вызывают эмоцию X у человека А2’ — неагентивные каузативы

Однако в реальности картина еще сложнее, т. к. агентивные каузативы делятся на конативные (злить своими выходками) и не конативные (злить своей прямотой), и это также влияет на их взаимодействие с отрицанием (подробнее см. в разделе **2.4**).

2.3. Вид и наклонение каузативного глагола

Следующий фактор, влияющий на интерпретацию отрицания с эмоциональными каузативами — это (как видно из предыдущих примеров) вид каузативного глагола. Для форм НЕСОВ интерпретация с широкой СД более характерна, чем для соответствующих форм СОВ[2]. Это связано с видо-временной семантикой, а именно с тем, что НЕСОВ в подобных сочетаниях имеет узуальное значение, а СОВ — результативное. По этой причине с формами НЕСОВ именная группа ТВОР со значением причины тяготеет к родовому референтному статусу, а с формами СОВ — к конкретно-референтному.

(10) а. *Он не радует ее приходами*
б. *Он не обрадовал ее приходом*

Родовой референтный статус облегчает попадание ИГ в сферу действия отрицания, в то время как конкретно-референтный статус ИГ это затрудняет. Поэтому фразу (10а) естественнее интерпретировать как ‘Он вообще не приходит => Она не радуется’, а фразу (10б) как ‘Он пришел; она не обрадовалась’. Однако это видовое распределение не носит абсолютного характера; при введении дополнительных элементов в контекст, статус именной группы и, соответственно, сфера действия отрицания меняется:

(11) а. *Он не радует ее своими частыми приходами*
б. *За весь этот год он ни разу не обрадовал ее приходом*

Фраза (11а) с узкой СД отрицания указывает на отсутствие ожидаемой эмоциональной реакции на событие, в то время как (11б) — на отсутствие самого события, которое могло бы вызвать эмоцию.

2.4. Тип конструкции

Интерпретация с широкой СД отрицания возможна для эмоциональных каузативов только в одной конструкции — с ИГ со значением причины в форме ТВОР. Конструкции, где валентность причины выражена деепричастием или предложно-именной группой из-за РОД, допускают только узкую СД отрицания:

(12) а. *Мама не огорчилась из-за Петиной двойки*
‘Петя получил двойку, но мама не огорчилась’
*‘Петя не получил двойку’

---

[2] Мы считаем пары типа *радовать-обрадовать* видовыми парами с четвертым типом видового противопоставления по М. Я. Гловинской [Гловинская 1982, 2001].

б. *Петя не огорчил маму, получив двойку*
‘Петя получил двойку, но мама не огорчилась’
*‘Петя не получил двойку’

Этим эмоциональные каузативы отличаются от интерпретативов, которые в аналогичной конструкции (с деепричастием) допускают две интерпретации отрицания:

(13) *Иван не согрешил, изменив жене*
‘Иван изменил жене, но это не грех’ — узкая СД отрицания
‘Иван не изменил жене’ — широкая СД отрицания

2.5. Тип эмоции и семантическая структура эмоционального каузатива

Взаимодействие отрицания, а также других семантически активных элементов с различными эмоциональными каузативами обнаруживает различия в их семантической структуре. Эти различия определяются типом эмоции. Различия касаются статуса каузативного компонента — является ли он подлинной пресуппозицией, т. е. такой, которая не может подвергаться воздействию отрицания и других семантически активных элементов (например, кванторных и экзистенциальных слов), или же «плавающей» пресуппозицией, т. е. такой, которая способна переходить в ассерцию и вступать в семантическое взаимодействие с другими элементами высказывания. На материале эмоциональных каузативов представлена вся шкала возможностей — от подлинно пресуппозитивного статуса каузативного компонента до ассертивного.

Поскольку для эмоциональных каузативов СОВ, вне зависимости от типа эмоции, при прочих равных предпочтительной является интерпретация с узкой СД (т. е., каузативный компонент у них по статусу ближе к подлинной пресуппозиции), рассмотрим глаголы НЕСОВ со значением каузации разных типов эмоций, у которых представлен весь возможный спектр статусов. Рассмотрим пару фраз:

(14) а. *Он перестал злить учительницу своими выходками*
б. *Он перестал злить учительницу своей прямотой*

Фразу (14а) естественно интерпретировать с широкой СД содержащего отрицание глагола *перестать*; она указывает на отсутствие причины для эмоции и, как следствие, отсутствие эмоции. Фразу (14б) естественно интерпретировать с узкой СД отрицательного глагола; она указывает на отсутствие эмоции при наличии причины для нее. При этом (14а) и (14б) различаются характером каузатива — в первой фразе каузатив конативен, указывает на усилия агенса, направленные на достижение результата, а во второй — нет. Можно сделать вывод о том, что у конативных каузативов каузативный компонент «весомее» и в силу этого «ближе к поверхности», к ассерции, менее глубоко

запрятан — и поэтому легче взаимодействует с другими семантическими компонентами, а у не-конативных — менее «весом», ближе к пресуппозиции, глубже запрятан и поэтому с большим трудом вступает во взаимодействие с другими семантическими компонентами.

Предположение о связи конативного/неконативного характера каузатива со статусом каузативного компонента в его семантической структуре подтверждается на более широком материале эмоциональных глаголов.

2.5.1. Тесты на конативность каузатива

Как было показано в разделе **2.1**, на материале каузативов индикатором конативности является способность к употреблению в конструкции вида *X-НЕСОВ, X-НЕСОВ, да и <и наконец> X-СОВ*, где целенаправленное усилие по совершению действия X создает видовую пару третьего типа по М. Я. Гловинской[3] [Гловинская 1982, 2001]:

(15) *Он ее злил, злил и наконец разозлил*

Еще один индикатор конативности — способность употребляться в длительной конструкции вида *два часа* или *весь урок*, указывающими на достаточно короткие периоды времени (длительные обстоятельства типа *десять лет* или *всю жизнь* могут свободно сочетаться с обозначениями постоянных свойств, поэтому не являются индикативными).

(16) *Он весь урок злил учительницу*

Как видно из примеров (16) и (17), каузатив *злить* может обозначать длительное и требующее усилий действие, производимое с целью вызвать эмоцию. Наконец, конативные каузативы, в отличие от не-конативных, характеризуются способностью употребляться в актуально-длительном значении НЕСОВ.

(17) а. — *Что ты сейчас делаешь? — Учительницу злю <друзей веселю>.*
б. — *Что ты сейчас делаешь? — Сестру пугаю.*

(18) — *Что ты сейчас делаешь? — *Бабушку радую.*

Контраст между примерами (17) и (18) демонстрирует различие в семантике между глаголами типа *злить, пугать, веселить* и глаголами типа *радовать*: первые в большей мере являются обычными действиями, а вторые — интерпретационными действиями, по фундаментальной классификации предикатов Ю. Д. Апресяна (Апресян 2006).

---

[3] Вообще же пары типа *радовать-обрадовать*, *восхищать-восхитить*, *обижать-обидеть* можно отнести, как было сказано выше, скорее к четвертому типу видового противопоставления по М. Я. Гловинской.

В совершенном виде различие между конативными и неконативными глаголами в значительном степени стирается, поскольку у глаголов СОВ в фокусе внимания находится результат, а не путь, ведущий к его достижению. Однако и для них возможен тест на конативность — а именно, сочетаемость с наречиями со значением усилия типа *еле*, *едва*, и *с трудом*. Возможность такой сочетаемости указывает на наличие конативного компонента в семантике каузатива. Ср.:

(19) *Он с трудом разозлил <развеселил, испугал> свою учительницу,*

но не

(20) **Он с трудом обрадовал <удивил> свою маму*

Поскольку конативность глаголов СОВ и НЕСОВ коррелирует внутри видовых пар, ниже рассматриваются только глаголы НЕСОВ, в контекстах с которыми конативная семантика проявляется более ярко.

2.5.2. Конативные каузативы в разных типах эмоций

Применим сформулированные тесты к каузативам других типов эмоций. Обнаруживается, что далеко не все типы эмоций позволяют формирование синтетических каузативов; так, они отсутствуют в следующих эмоциональных кластерах в русском языке [о понятии эмоционального кластера см. В. Апресян 2011]: ‘ОТВРАЩЕНИЕ’, ‘ГОРДОСТЬ’, ‘ЖАЛОСТЬ’ (ср., впрочем, *бить на жалость*, *разжалобить*), ‘РЕВНОСТЬ’, ‘ЗАВИСТЬ’, ‘БЛАГОДАРНОСТЬ’, ‘ПРЕЗРЕНИЕ’. Каузация соответствующих эмоций обозначается лексико-функциональным глаголом *вызывать* в сочетании с названием эмоции.

Отсутствие синтетического каузатива в кластере ‘ОТВРАЩЕНИЕ’ является, по-видимому, особенностью русского языка; ср. английский каузатив *to disgust* или немецкий *anekeln* ‘вызывать отвращение’.

То же самое, по-видимому, верно для ‘ЖАЛОСТИ’; ср. немецкое *erbarmen* ‘вызывать жалость, сострадание’ или французское *apitoyer* ‘вызывать жалость’. Каузатив *разжалобить* в данной связи не рассматривается, т. к. он всегда агентивен и конативен, в отличие от обычных эмоциональных каузативов, которые имеют агентивные и неагентивные употребления; ср. Он ее злит vs. *Его злит ее прямота*, но не **Его разжалобили ее раны*.

Что касается эмоций кластеров ‘ГОРДОСТЬ’, ‘РЕВНОСТЬ’, ‘ЗАВИСТЬ’, ‘БЛАГОДАРНОСТЬ’, ‘ПРЕЗРЕНИЕ’, отсутствие синтетических каузативов, по-видимому, носит достаточно универсальный характер (они отсутствуют и в других европейских языках) и связано с относительно произвольным характером их каузации: если для эмоций типа ‘СТРАХА’ существует прямая связь между стимулом и эмоцией, для перечисленных четырех типов эмоций эта связь существенно более опосредована и менее обязательна. В некотором смысле это эмоции-«интерпретации» — их может вызвать любой объект, а также любое состояние, свойство или действие, если экспериенсер воспринимает и расценивает их определенным образом.

Синтетические каузативы присутствуют в следующих эмоциональных кластерах: ‘СТРАХ’ *(пугать, страшить, ужасать)*, ‘ГНЕВ’ *(возмущать, сердить, злить, раздражать, бесить)*, ‘РАДОСТЬ’ *(радовать, веселить, восхищать)*, ‘ГРУСТЬ’ *(расстраивать, огорчать, печалить)*, ‘СТЫД’ (смущать, конфузить; глагол *стыдить* не рассматривается, поскольку указывает на конкретное речевое действие), ‘ОБИДА’ (*уязвлять, задевать*; глагол *обижать* не рассматривается из-за сильной интерференции его основного значения — агрессивного физического действия; ср. *Нельзя обижать маленьких*), ‘УДИВЛЕНИЕ’ *(удивлять, изумлять)*.

Далеко не все из перечисленных каузативов агентивны: так, *страшить* и *ужасать* обычно описывают эмоцию, внушаемую какими-то явлениями или событиями, находящимися вне контроля человека:

(21) а. *Меня страшит неминуемая гибель*
б. *Меня ужасает его вид*

Эмоция не может являться результатом действий агенса:

(22) **Он меня страшит <ужасает> своими мрачными пророчествами*

Неагентивны также каузативы *печалить* и *уязвлять*:

(23) а. *Меня печалит его судьба*
б. **Он печалит меня своими мрачными пророчествами*

(24) а. *Меня уязвляет его равнодушие*
б. **Он уязвляет меня своими презрительными замечаниями*

Прочие каузативы агентивны, однако различаются по степени конативности. Первый тест — на употребление в конструкции *X-НЕСОВ, X-НЕСОВ, да и <и наконец> X-СОВ* — дает следующие результаты.

Кластер ‘СТРАХ’
(25) *Он меня пугал, пугал и наконец испугал*

Кластер ‘ГНЕВ’
(26) а. **Он меня возмущал, возмущал и наконец возмутил*
б. **Он меня раздражал, раздражал и наконец раздражил*
в. *?Он меня сердил, сердил <бесил, бесил> и наконец рассердил <взбесил>*
г. *Он меня злил, злил и наконец разозлил*

Кластер ‘РАДОСТЬ’

(27) а. **Он меня радовал, радовал <восхищал, восхищал> и наконец обрадовал*[4] *<восхитил>*

б. *Он меня веселил, веселил и развеселил*

Кластер ‘ГРУСТЬ’

(28) а. **Он меня огорчал, огорчал и огорчил*

б. **Он меня расстраивал, расстраивал и рассстроил*

Кластер ‘СТЫД’

(29) а. **Он меня конфузил, конфузил и сконфузил*

б. *?Он меня смущал, смущал и смутил*

Кластер ‘ОБИДА’

(30) **Он меня задевал, задевал и задел*

Кластер ‘УДИВЛЕНИЕ’

(31) **Он меня удивлял, удивлял <изумлял, изумлял> и удивил <изумил>*

Второй тест — на употребление в длительной конструкции дает следующие результаты.

Кластер ‘СТРАХ’

(32) *Он два часа пугал нас рассказами о привидениях*

Кластер ‘ГНЕВ’

(33) а. **Он два часа возмущал <раздражал> нас наглым поведением*

б. *?Он два часа сердил <бесил> нас наглым поведением*

в. *Он два часа злил нас глупыми выходками*

Кластер ‘РАДОСТЬ’

(34) а. **Он два часа восхищал нас виртуозной игрой на фортепиано*

б. *Он два часа радовал нас виртуозной игрой на фортепиано*

в. *Он два часа веселил нас шутками*

Кластер ‘ГРУСТЬ’

(35) **Он два часа расстраивал <огорчал> меня своей невнимательностью*

Кластер ‘СТЫД’

(36) **Он меня два часа смущал <конфузил> своими нескромными взглядами*

---

[4] Несобственно видовой парой к *радовать* является глагол СОВ *обрадовать*; делимитативы на *по-* не рассматриваются в данном контексте, поскольку не формируют видовых пар. О семантических различиях между *порадовать* и *обрадовать* см. (Зализняк, Шмелев 2012).

Кластер ‘ОБИДА’

(37) **Он меня два часа задевал своими колкостями*

Кластер ‘УДИВЛЕНИЕ’

(38) **Он меня два часа удивлял <изумлял> новыми фокусами*

Обобщая результаты тестов, можно расположить агентивные эмоциональные каузативы на шкале конативности, от наиболее конативных к наименее конативным (размер шрифта маркирует степень конативности):

(39) **ЗЛИТЬ, ПУГАТЬ, ВЕСЕЛИТЬ**, СЕРДИТЬ, БЕСИТЬ, РАДОВАТЬ, РАССТРАИВАТЬ, ОГОРЧАТЬ, ВОЗМУЩАТЬ, РАЗДРАЖАТЬ, ВОСХИЩАТЬ, УДИВЛЯТЬ, ИЗУМЛЯТЬ, ЗАДЕВАТЬ, СМУЩАТЬ, КОНФУЗИТЬ

Как видно из списка в целом, в него вошли представители ***базовых*** эмоций [согласно Ekman 1999] — гнева, радости, страха, удивления, грусти, стыда. Отсутствует отвращение, но, как было сказано выше, это особенность русской языковой картины мира.

При этом наибольшая степень конативности агентивных каузативов свойственна трем подтипам эмоций — *злить*[5], *пугать* и *веселить*. Их непосредственным стимулом часто служит направленное на их каузацию усилие человека.

Средняя степень конативности свойственна пяти подтипам эмоций — *сердить*, *бесить*, *радовать*, *расстраивать*, *огорчать*. Это эмоции, для которых непосредственным стимулом реже выступает собственно усилие человека, направленное на их каузацию, а чаще — особая оценка или восприятие какого-то события или действия как хорошего или плохого, что, в свою очередь, каузирует эмоцию. При этом каузация *сердить*, *бесить*, *радовать*, *расстраивать*, *огорчать* воспринимается как в некоторой мере контролируемая человеком; ср. возможность императива и прогибитива для них:

(40) *Почаще радуй родителей*

(41) *Не серди <не беси> меня*

(42) *Не расстраивай <не огорчай> бабушку*

При этом побуждение или запрет относятся к совершению действия, которое воспринимается говорящим как способное каузировать эмоцию, а не собственно к каузации эмоции. Конативы *злить*, *пугать* и *веселить* также обозначают контролируемые действия:

---

[5] Ср. наречие *назло*, лексикализующее ту же идею — целенаправленной каузации эмоции *злости*. Ни у какой другой эмоции в русской языке целенаправленная каузация не достигает такой степени лексикализации.

(43) *Давай, весели меня*

(44) *Не пугай <не зли> меня*

Однако у них побуждение и запрет относятся собственно к усилиям каузатора-агенса, направленным непосредственно на каузацию эмоции.

Таким образом, конативность и контролируемость являются коррелирующими, но не совпадающими свойствами; по крайней мере, в случае эмоциональных каузативов контролируемость семантически слабее — конативность включает в себя контролируемость, но не наоборот.

Интересно, что некоторые разновидности ‘ГНЕВА’ и ‘РАДОСТИ’, при том, что их каузатором является человек, возникают не как результат его усилий: так, согласно русской языковой картине мира, люди специально не прилагают усилий, чтобы вызвать *раздражение* или *возмущение*, и, по-видимому, не могут в результате каких-то алгоритмически запланированных действий вызвать эмоцию *восхищения*. Кроме того, ограничена способность *радости* возникать в результате целенаправленных усилий агенса. На самом деле, в результате последовательно прилагаемых усилий возникают только веселье или смех (ср. *веселить, смешить*) — т. е. эмоционально более простые и физиологичные реакции, чем существенно более сложная, оценочная и рациональная радость.

Длительные и последовательные усилия для достижения результата невозможны также для связанной с внезапностью эмоции *удивления* и тем более *изумления*; кроме того, согласно русской языковой картине мира, люди обычно не прилагают усилий для того, чтобы *смутить*, *сконфузить* или *задеть* другого человека — это обычно случается либо помимо их воли, либо без специальных усилий.

Положение каузативов *огорчать* и *расстраивать* на этой шкале указывает не только на то, что люди обычно не прилагают специальных усилий, чтобы каузировать ‘ГРУСТЬ’, но и на то, что они не всегда могут ***избежать*** каузации эмоций этого типа.

Итак, эмоции, которые люди обычно стараются и могут вызвать своими намеренными усилиями — это злость, испуг и веселье. Их объединяет то, что, помимо отвращения, это наиболее физиологичные из всех эмоциональных реакций. Они, во-первых, сопровождаются явными физическими проявлениями (повышение температуры тела для злости, понижение температуры тела для испуга, и смех для веселья), во-вторых, они наименее рационализованы (свойственны, в частности, не только людям, но и животным), а также наиболее непосредственно связаны с внешним стимулом. При этом их стимулы в меньшей степени привязаны к конкретным ситуациям, чем у других эмоций — существуют достаточно универсальные способы *злить*, *пугать* и *веселить* (*смешить*). Например, *испугать* можно любого человека, неожиданно выпрыгнув из-за угла или подкравшись со спины и громко крикнув ему в ухо, а вот для того, чтобы *порадовать*, необходимо искать более индивидуальные подходы, поскольку эта эмоция рационализирована, опосредована и возникает, проходя через некоторую систему оценок, которая у всех людей различна. Эту эмоцию можно каузировать намеренно, однако усилие, которого она требует, направлено не непосредственно

на каузацию эмоции, а на выполнение действия, которое расценивается как хорошее, в результате чего ожидается возникновение эмоции:

(45) *Порадуй стариков, зайди к ним в гости*

Не случайно порадовать, как отмечается в работе (Зализняк, Шмелев 2012) может самостоятельно употребляться как номинация самого действия:

(46) *Исполнители нас сегодня порадовали* [= 'хорошо исполнили']

2.5.3. Конативность каузатива и его семантическая структура

Рассмотрим теперь взаимодействие агентивных эмоциональных каузативов с отрицанием. Рассмотрим три группы примеров — с конативными агентивными каузативами, с не-конативными агентивными каузативами и с группой промежуточных с точки зрения конативности каузативов.

Конативные каузативы:

(47) а. *Он больше не злит учительницу (глупыми выходками)*
б. *Он больше не пугает одноклассников (страшными рожами)*
в. *Он больше не веселит домашних (смешными шутками)*

Для этой группы примеров предпочтительной интерпретацией является интерпретация с широкой сферой действия отрицания: прекращается усилие, направленное на каузацию эмоции, и, как следствие, прекращается сама эмоция. Если бы эта интерпретация была единственной, можно было бы утверждать, что для этой группы статус каузативного компонента — ассерция, статус эмоциональной реакции — импликатив. Однако поскольку возможны фразы типа *Он пугает, а мне не страшно* (Л. Н. Толстой об Андрееве) или *Он меня не злит, а я злюсь,* можно утверждать, что каузативный и эмоциональный компоненты достаточно независимы. Таким образом, наиболее вероятное предположение состоит в том, что статусы каузативного и эмоционального компонентов у конативных эмоциональных каузативов — это две независимые ассерции.

Предлагаемая семантическая структура:

(48) *A1 X-ит A2 своим A3* 'Человек A1 совершает усилия A3, направленные на то, чтобы вызвать эмоцию X у человека A2; человек A2 испытывает эмоцию X'

Не-конативные каузативы:

(49) а. ??*Он больше не возмущает учительницу (бездельем на уроках)*
б. *Он больше не раздражает одноклассников (неуместными шутками)*
в. ?*Он больше не восхищает слушателей (своей игрой)*
г. ?*Он больше не удивляет <не изумляет> окружающих (точными предсказаниями)*
д. *Он больше не смущает меня (неприличными анекдотами)*

е. **Он больше не конфузит меня (вопросами личного характера)*
ж. *Он больше не задевает меня (колкостями)*

Как видно из примеров, для некоторых каузативов этой группы сочетаемость с отрицанием вообще затруднена. Там, где она возможна, предпочтительной интерпретацией является интерпретация с узкой сферой действия отрицания: действие, каузирующее эмоцию, происходит, однако оно перестает вызывать эмоциональную реакцию.

Следовательно, для этой группы статус каузативного компонента — пресуппозиция, статус эмоциональной реакции — ассерция.

Предлагаемая семантическая структура:

(50) *A1 X-ит A2 своим A3* ‘Совершая действия A3, человек A1 вызывает эмоцию X у человека A2’

Промежуточные по конативности контролируемые каузативы:

(51) а. ??*Он больше не сердит родителей (прогулами)*
б. ??*Он больше не бесит сестру (беспорядком в комнате)*
в. *Он больше не радует слушателей (своей игрой)*
г. *Он больше не расстраивает <не огорчает> окружающих пессимистичными прогнозами*

Как видно из примеров, для некоторых каузативов этой группы сочетаемость с отрицанием также затруднена. Там, где она возможна, одинаково возможны обе интерпретации: с широкой СД отрицания, где отрицается само наличие действия, и с узкой СД отрицания, где отрицается только наличие ожидаемой эмоциональной реакции. При этом невозможность фраз типа **Он меня радует, а я не радуюсь*, **Он меня не радует, а я радуюсь* свидетельствует как о не вполне «дотягивающем» до ассерции статусе каузативного компонента, так и о зависимом характере эмоционального компонента.

Следовательно, для этой группы статус каузативного компонента и статус эмоциональной реакции не фиксированы: каузативный компонент «плавает» между пресуппозицией и ассерцией, а статус эмоциональной реакции — между ассерцией и импликативом.

Предлагаемая семантическая структура:

(52) *A1 X-ит A2 своим A3* ‘Человек A1 совершает действия A3, которые вызывают эмоцию X у человека A2’

3. Заключение

Итак, удалось установить зависимость между следующими факторами: тип эмоции; тип каузатива; статус каузативного компонента. Статус каузативного

компонента (пресуппозитивный vs. ассертивный) коррелирует со степенью конативности каузатива: чем больше выражена семантика усилия в каузативе, тем ближе каузативный компонент к семантической поверхности и, следовательно, к ассерции. Наибольшей степенью конативности в русском языке характеризуются каузативы базовых эмоций — страха, гнева, радости, в особенности следующих подтипов этих эмоций — *злить*, *пугать*, *веселить*. В их семантических структурах статус каузативного элемента наиболее близок к ассертивному.

Литература

1. *Апресян В. Ю.* Семантическая структура слова и его взаимодействие с отрицанием // Материалы международной конференции «Диалог 2010». Компьютерная лингвистика и интеллектуальные технологии. Выпуск 9 (16). с. 13–19.
2. *Апресян В. Ю.* Опыт кластерного анализа: русские и английские эмоциональные концепты. Часть 1 // ВЯ, №1, 2011. с. 19–51.
3. *Апресян Ю. Д.* Основания системной лексикографии // Языковая картина мира и системная лексикография. Отв. ред. Ю. Д. Апресян. М.: «Языки славянских культур», 2006. с. 145–160.
4. *Апресян Ю. Д.* Часть первая. Основания системной лексикографии. Глава 2. Фундаментальная классификация предикатов. // Языковая картина мира и системная лексикография. / Апресян Ю. Д. (отв. ред.). — М.: Языки славянских культур, 2006.
5. *Апресян Ю. Д.* Грамматика глагола в Активном словаре (АС) русского языка // Смыслы, тексты и другие захватывающие сюжеты: сборник статей в честь 80-летия Игоря Александровича Мельчука). Под ред. Ю. Д. Апресяна и др. М.: «Языки славянских культур», 2012.
6. *Гловинская М. Я.* Семантические типы видовых противопоставлений русского глагола. М., 1982.
7. *Гловинская М. Я.* Многозначность и синонимия в видо-временной системе русского глагола. М., 2001.
8. *Зализняк А. А., Шмелев А. Д.* Лексика радости // Константы и переменные русской языковой картины мира / А. А. Зализняк, И. Б. Левонтина, А. Д. Шмелев. М.: Языки славянских культур, 2012. с. 462–470
9. *Карловска А. К.* Русские каузативы движения и перемещения (смысловой анализ). Канд. дисс. М., 1990.
10. *Кустова Г. И.* О коммуникативной структуре предложений с событийным каузатором // Московский лингвистический журнал. Т. 2. М.: РГГУ, 1996. с. 240–261.
11. *Падучева Е. В.* Эффекты снятой утвердительности: глобальное отрицание // Русский язык в научном освещении. 10 (2). М.: «Языки славянских культур», 2005. с. 17–42.
12. *Плунгян В. А., Рахилина Е. В.* Тушат-тушат — не потушат: грамматика одной глагольной конструкции // Construction Linguistics. Под ред. Е. В. Рахилиной. М., «Азбуковник», 2010.
13. *Ekman P.* (1999), Basic emotions in Handbook of cognition and emotion, Sussex.

References

1. *Apresyan V. Yu.* Semanticheskaia struktura slova i ego vzaimodeistvie s otricaniem [Semantic structure of words and their interaction with negation]. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, 2010. Pp. 13–19.
2. *Apresyan V. Yu*. (2011), Opyt klasternogo analiza: russkie i angliiskie emotsional'nye kontsepty. Chast' 1 [Cluster analysis: Russian and English emotion concepts], Voprosy iazykoznaniia [Issues in Linguistics], vol. 1, pp. 19–51.
3. *Apresyan Yu. D.* (2006), Osnovania sistemnoi leksikografii [Foundations of Systematic Lexicography], in Iazykovaiia kartina mira i sisemnaia leksikografiia [Linguistic picture of the World and Systematic Lexicography], Iazyki slavianskih kul'tur, Moscow, pp. 145–160.
4. *Apresyan Yu. D.* (2012), Grammatika glagola v Aktivnom Slovare (AS) russkogo iazyka [Grammar of Verbs in the Active Dictionary (AD) of Russian], in Smysly, teksty, i drugie zahvatyvaiushchie siuzhety [Meanings, Texts, and other Fascinating Plots: A Festshrift to Commemorate the 80-th Anniversary of Professor Igor Alexandrovič Mel'čuk], Iazyki slavianskoi kul'tury, Moscow, pp. 42–59.
5. *Ekman P.* (1999), Basic emotions in Handbook of cognition and emotion, Sussex.
6. *Glovinskaia M. Ia.* (1982), Semanticheskie tipy vidovyh protivopostavlenii russkogo glagola [Semantic Types of Aspect Opposition in Russian Verbs], Moscow.
7. *Glovinskaia M. Ia.* (2001), Mnogoznachnost' i sinonimiia v vido-vremennoi sisteme russkogo glagola [Polysemy and Synonymy in Tense-Aspect System of Russian Verbs], Moscow.
8. *Karlovska A. K.* (1990), Russkie kauzativy dvizheniia i peremeshcheniia (smyslovoi analiz) [Russian Causatives of Motion and Movement (Semantic Analysis)]. Ph.D. Thesis, Moscow.
9. *Kustova G. I.* (1996), O kommunikativnoj structure predlozhenii s sobytiinym kauzatorom [On the Communicative Structure of Sentences with Subject Causator], Moskovskii lingvisticheskii zhurnal [Moscow Linguistic Journal], Moscow, RGGU, pp. 240–261.
10. *Paducheva E. V.* (2005), Effekty sniatoi utverditel'nosti: global'noe otritsaniie [Effects of Removed Affirmativeness: Global Negation], Russkii iazyk v nauchnom osveshchenii [Russian Language under Scholarly Analysis], vol. 10 (2), pp. 17–42.
11. *Plungian V. A., Rahilina E. V.* (2010), Tushat-tushat- ne potushat: grammatika odnoi glagol'noi konstruktsii [They are putting and putting out the fire: the Grammar of one Verb Construction], in Construction Linguistics, Moscow, pp. 83–94.
12. *Zalizniak A. A., Shmelev A. D.* (2012), Leksika radosti [Lexicon of joy], in Konstanty i peremennye russkoj jazykovoj kartiny mira [Constants and variables of the Russian linguistic worldview], Moscow. pp. 462–470.

CORRECTING COLLOCATION ERRORS IN LEARNERS' WRITING BASED ON PROBABILITY OF SYNTACTIC LINKS

Azimov A. E. (mitradir@gmail.com)
Lomonosov Moscow State University, Moscow, Russia

Bolshakova E. I. (eibolshakova@gmail.com)
National Research University Higher School of Economics;
Lomonosov Moscow State University, Moscow, Russia

The paper describes a novel method for automatic collocation error correction in NL texts written by language learners or translated from another NL with the aid of machine translators. We assume that the main cause of collocation errors is the strategy of word-by-word translation used by authors of the texts or by machine translators, so the errors essentially depend on the source language. While processing a sentence from the text, the method considers as potential correcting variants all its paraphrases that have the same syntactic structure and are built by replacing all words of the sentence by their substitutes. Substitutes are automatically generated using word translation equivalents taken from a translation dictionary. To detect an error in the sentence, we propose a relevance degree function computed from the probability of the word's syntactic links and applied to the sentence and its paraphrases. If the function value for the sentence is lower than for some of its paraphrases, our method signals an error, then it is corrected by an appropriate sentence paraphrase. The method was evaluated by correcting collocation errors in English texts written by Russian speakers. Stanford Parser and an English text collection were used to gather statistics and compute the probability of English word syntactic links. Within certain limitation, the experiments gave promising results: our method detected about 80% of collocation errors (with words of various POS) and 87% of proposed correcting paraphrases contained a proper correction.

Key words: lexical combinability, collocations, collocation error, error correction, sentence probability, ESL writing

Introduction

Based on computer dictionaries and parsing methods, modern computer text editors and spellers detect both spelling and some syntactic errors in NL texts, but they can't reveal and correct so-called collocation errors. Such errors (e.g., Eng. *heavy tea* instead of *strong tea*; Rus. *играть значение* instead of *иметь значение*) violate norms of lexical combinability. The norms differ for various natural languages and can't be defined by formal rules. As a rule, special collocation dictionaries, for example, [7] include, but only partially, typical collocations for a particular NL.

In modern computational linguistics the concept of collocation may be interpreted in different ways. Following the work [1] we consider a collocation as a stable word combination syntactically related and semantically compatible.

Collocation errors may occur in texts written by native speakers, but more often they arise in texts written by authors for whom the language of the text isn't native, in particular, written by language learners. Such errors also may be the result of automatic translation from one language to another. Below we present several examples of inadequate translation from Russian into English (machine translator Lingvo popular in Russia was used):

- Russian collocation *великий художник* is translated to *great painter*, but in English the collocation *great artist* is more idiomatic;
- Word combination *публичная персона* is translated to *public person* instead of commonly used *public figure*;
- Russian collocation *много денег* is translated into *many money* instead of *much money*.

Such collocation errors often appear in texts as the result of the strategy of word-by-word translation used by either a machine translator or a language learner. The strategy doesn't take into account syntactic and semantic relations between words of the text. The paper [3] argues that the model of native language and the incomplete model of target language (formed in learner's mind or built into the automatic translator) interfere, thus resulting a plenty of mistakes and collocation errors among them.

In the recent decade a number of papers [1, 2, 4, 5, 9, 10] have appeared in the field of computational linguistics, which proposed certain ways to automatically correct collocation errors in NL texts, mainly in English (besides English only Russian and French texts were considered). To detect erroneous collocations the methods proposed in these works use both automatic syntactic analysis of sentences and statistics of word occurrences and cooccurrences. Most of the works employ row statistics, i.e. frequencies of words and word combinations. As a rule, only particular types of word combination are considered while detecting errors: preposition — noun [2], noun — verb [10], collocations with articles [4].

The separate problem is selecting potential word combinations for correcting an erroneous collocation yet detected. Some works, in particular [4], propose to use bases of all possible correcting phrases manually precompiled by human experts, but in practice this way is evidently unreal.

As for accuracy of collocation errors correction, for the works mention above, it varies from 40% [1] up to 69% [10].

This paper describes a method for automatic collocation error correction in NL texts based on probability of word syntactic links computed using statistics of word syntactic links. Unlike previous works, our method handles several types of collocations, including word combinations with content (nouns, adjectives, verbs, etc.) and auxiliary words (prepositions).

While developing the method we suppose the following:

- Texts under correction are written by language learners or are translated from another NL with the aid of certain machine translator.
- The main reason of collocation errors is the applied strategy of word-by-word translation, so the errors essentially depend on the source language.

- Collocation errors don't change the syntactic structure of the sentences; they only substitute erroneous words for correct ones.

Our method follows the idea formulated in [2]: the problem of error correction within a sentence S may be considered as the task to find most probable correcting sentence V^*, among possible sentences V, given sentence S:

$$V^* = arg \max_V P(V|S) \tag{1}$$

Taking into account Bayes' theorem, we get:

$$V^* = arg \max_V \frac{P(S|V)P(V)}{P(S)}$$

and then, using conditional independence $P(S)$ of possible sentence V:

$$V^* = arg \max_V P(S|V)P(V) \tag{2}$$

Therefore, to find the most probable substitute sentence V^*, it is necessary to determine probability of sentence V as correcting variant for S and also to determine probability of sentence V.

The text of the paper is organized as follows. First we describe how to build correcting sentences V, we call them *paraphrases*, and how to determine their conditional probabilities $P(S|V)$. Any paraphrase V is constructed by replacing words from S by their *substitute words* that are automatically generated on the bases of word translation equivalents taken from a particular translation dictionary, for example, Russian-English dictionary.

Next, we explain how to determine the probability $P(V)$ of any sentence V in the text (including paraphrases of the source sentence) given its syntactic structure. Further we describe the way how to detect collocation errors in the sentence. For this purpose we propose *a relevance degree function* computed from the probability of word syntactic links and applied to the sentence and their paraphrases. If the function value for the sentence under correction is less than for some its paraphrase, our method signals an error, and the sentence is corrected by appropriate paraphrase.

Then experimental validation of our method is discussed. The method was evaluated by correcting collocation errors in English texts written by Russian speakers. Stanford Parser [6], English text corpora, and Russian-English dictionary were used to compute necessary probabilities. Within certain limitation, the experiments gave promising results.

Finally, we draw conclusions and outline directions for future work.

Paraphrases and their Probabilities

To determine the probability of paraphrases we need to determine the probability of their components, i.e. substitute words.

Recalling our assumption about word-by-word strategy of translation, we define the map *Translate* as set of ordered pairs ⟨*x*,*y*⟩ where word *x* belongs to the source language *X* and has a set of translation equivalents {*y*} from the target language *Y* — cf. Figure 1. We also define the inverse map *Translate*$^{-1}$ which determines for each word *z* from language *Y* a set of preimages {*x*} from language *X*.

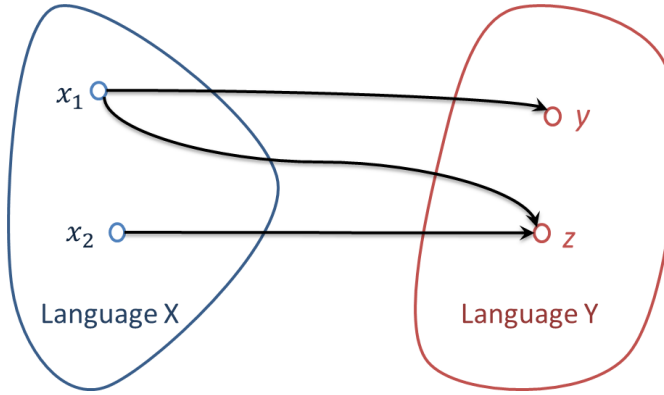


Fig. 1. The map *Translate*

Let $p_+(y|x)$ be conditional probability that word *x* is preimage of word *y* on the map *Translate*. Similarly, let $p_-(y|x)$ be conditional probability on the map *Translate*$^{-1}$.

In accordance with the assumption that collocation errors arise as a result of word-by-word translation strategy from language *X* to language *Y*, a wrong word , as well as a correct one, both are images of certain word *x* from *X*. For every word *y* we consider *DoubleTranslation* set:

$$DoubleTranslation(y) = \{z \in Y \mid \exists x \in X : z \in Translate(x) \& y \in Translate(x)\} \quad (3)$$

The word *y* is also member of the set.

It is reasonable to consider members of this set as substitute words for every word *y* from language *Y*. However, some members of *DoubleTranslation*(*y*) may have the meaning very different from the meaning of the word *y*. For example, for word *future* (according to Lingvo Russian-English dictionary) the set includes words *next, to be, coming, the beyond, after death, beyond the grave, aftertime, approaching, by-and-by, hereafter, weird*. To overcome this problem we should take into account only synonyms of word *y,* thus obtaining the set:

$$Substitutes(y) = \{z \in Y \mid z \in DoubleTranslation(y) \,\&\, z \in Synonyms(y)\} \quad (4)$$

For example, the set of substitutes for word *beautiful* are *attractive, fine, gorgeous, handsome, pretty.*

Let us consider x as a preimage of words z and y from the language Y, i.e. $y \in Translate\ (x)$ & $z \in Translate\ (x)$. The conditional probability of substitute word z, given words y and x, and x is a preimage of y, is defined as follows:

$$p(z|y,x) = p_{-}(x|y)p_{+}(z|x) \tag{5}$$

To compute conditional probability of substitute z for a given word y we must sum values (5) for all common preimages of words z and y:

$$p_{dt}(z|y) = \sum_{\{x|y \in Translate\ (x) \& z \in Translate\ (x)\}} p_{-}(x|y)p_{+}(x|y) \tag{6}$$

Assuming the independence of collocation errors in the sentence under correction we get the following formula for the conditional probability of paraphrase V, given the sentence S:

$$p(S|V) = \prod_{i} p_{dt}(s_i|v_i), s_i \in Substitutues(v_i) \tag{7}$$

Where s_i is a word from the sentence S and v_i is a word of its paraphrase.

Hence we have defined the first factor in formula (2) and need to determine the probability $P(V)$.

Sentence Probability

We build for each sentence its dependency-based parse tree. This tree is a directed graph $G(V, E)$, where V is a set of vertices, they correspond to words of the sentence, and E is a set of edges. Any pair of vertices $(v_1, v_2) \in E$ if and only if the word-vertex v_2 has dependency relation with word-vertex v_1 (v_2 depends on v_1).

The word v_1 is *ancestor* of v_2, if directed path from vertex v_1 to vertex v_2 exists. Let *ancestors*(v_i) denote the set of all *ancestors* for word v_i .

We illustrate the *ancestors* concept with the next sentence: *The main library in the university is one of the largest in Russia*. Its dependency tree is shown in Figure 2, while the Table 1 presents *ancestors* set for each word of the sentence.

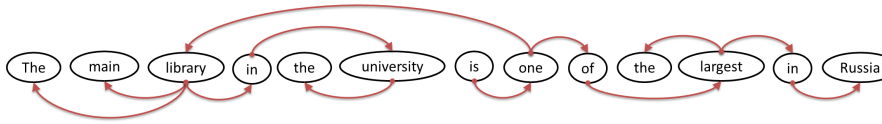


Fig. 2. Example of Dependency Tree

Table 1. Words and corresponding *ancestors* sets

Word	Ancestors
The	library, one, is
main	library, one, is
library	one, is
in	library, one, is
the	university, in, library, one, is
university	in, library, one, is
is	{}
one	Is
of	one, is
largest	of, one, is
the	largest, of, one, is
in	largest, of, one, is
Russia	in, largest, of, one, is

Since we consider collocations as syntactically related word combinations, we assume conditional independence of each word v_i from all other words except its *ancestors*. Thereby the joint probability of the words from the sentence, given the particular sentence parse tree, may be computed as a product of the conditional probabilities:

$$p(v_1, \dots, v_n) = \prod_{i=1}^{n} p\big(v_i | ancestors(v_i)\big) \tag{8}$$

Our method computes probabilities on the bases of word syntactic link statistics gathered on some text collection, so each factor of the above formula is:

$$p\big(v_i | ancestors(v_i)\big) = \frac{N\big(v_i, ancestors(v_i)\big)}{N\big(ancestors(v_i)\big)} \tag{9}$$

where $N(v_i, ancestors(v_i))$ and $N(ancestors(v_i))$ are frequencies of corresponding syntactically related group of words. We should mention that despite of word in the tree root has empty *ancestors* set, the joint sentence probability also depends on root word.

Collocation Errors Correction

Our assumption that collocation errors don't change the syntactic structure of sentences implies that the source sentence *S* and any its paraphrase *V* have isomorphic dependency-base trees — cf. Figure 3. In the source sentence *S* its words s_i were replaced by substitute words v_i (from corresponding *Substitutes* sets), thus resulting paraphrase *V*.

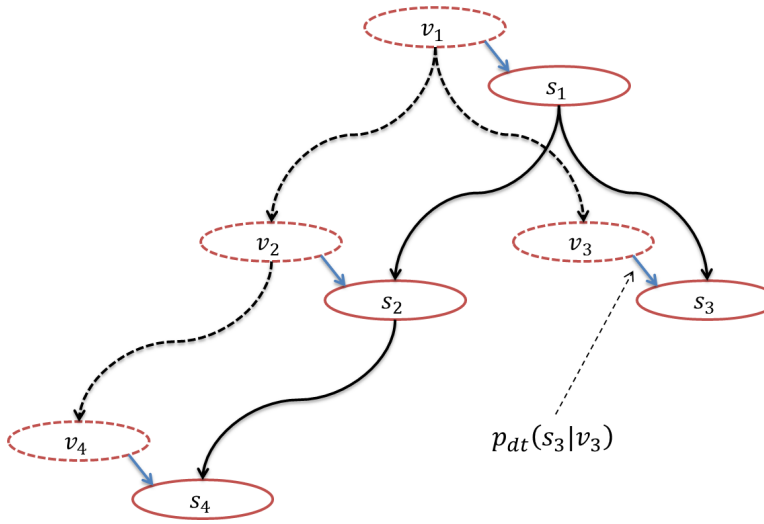


Fig. 3. Isomorphic dependency-base trees

Therefore, the *ancestors* set for word v_i has the following property:

$$ancestors(v_i) = \{v_k \mid v_k \in Substitutes(s_k): s_k \in ancestors(s_i)\}.$$

In such a way, combining formulas (7) and (8), we refine formula (2) and obtain:

$$V^* = arg \max_{v_1,\dots,v_k} \prod_{i=1}^{k} \left(p_{dt}(s_i|v_i)p(v_i|ancestors(v_i))\right) \quad (10)$$
$$\text{where } s_i \in Substitutes(v_i)$$

Next, with the aid of auxiliary coefficient

$$k_{dt}(s_i, v_i) = \frac{p_{dt}(s_i|v_i)}{p_{dt}(s_i|s_i)} \quad (11)$$

For paraphrase evaluation we define a relevance function *Degree*:

$$Degree(v_1, \dots, v_k) = \prod_{i=1}^{k} \left(k_{dt}(s_i|v_i)p(v_i|ancestors(v_i))\right), \quad (12)$$
$$s_i \in Substitutes(v_i)$$

Using this function, formula (10) could be rewritten as follows:

$$V^* = arg \max_{v_1,\dots,v_k} Degree(v_1, \dots, v_k), \quad (13)$$
$$s_i \in Substitutes(v_i)$$

Let us give meaningful explanation for Degree function: it measures the correspondence of a particular set of words to a given parse tree. Our method applies the function to the sentence under correction and their paraphrases. If the function value for the sentence is less than for some its paraphrases, an error is detected. In accordance with (13) the detected error is corrected by paraphrase V^* that give maximum Degree value.

We should add that degree of the source sentence is independent of conditional probabilities of substitute words, so the corresponding formula is

$$Degree(s_1, \dots, s_k) = \prod_{i=1}^{k} p(s_i | ancestors(s_i)) \tag{14}$$

Implementation of the Method and Experiments

The described method was evaluated by correcting collocation errors in English texts within the following limitations: there is no more than one collocation error in any sentence of the text, and for each vertex of sentence parse tree only two *ancestors* are considered, namely parent and grandparent.

In this case, the formula (13) can be rewritten as follows:

$$V^* = arg \max_{v_1,\dots,v_k} k_{dt}(s_i | v_i) \prod_{k=1}^{n} p(v_k | ancestors(v_k)) \tag{15}$$

where $v_k = s_k$ if $i \neq k$ and $s_i \in Substitutes(v_i)$ if $i = k$.

The conditional probabilities $p(v_k | ancestors(v_k))$ are based on statistics of syntactic links between words:

$$p(v_i | v_k) = \frac{N(v_i, v_{k1})}{N(v_{k1})} \tag{16}$$

$$p(v_i | v_{k1} v_{k2}) = \frac{N(v_i, v_{k1}, v_{k2})}{N(v_{k1}, v_{k2})} \tag{17}$$

where $N(v_{k1})$ is the frequency of occurrences of word v_{k1};
$N(v_i, v_{k1})$ and $N(v_{k1}, v_{k2})$ are the frequencies of corresponding syntactically linked pairs of words (v_i, v_{k1}) and (v_{k1}, v_{k2});
and $N(v_i, v_{k1}, v_{k2})$ is the frequency of syntactically linked triple of words (v_i, v_{k1}, v_{k2}).

The probability of the parse tree root is determined and calculated as follows:

$$p(v_i) = p(v_i | root) = \frac{N(v_i, root)}{N(root)} \tag{18}$$

In order to conduct experiments, we built a database containing frequencies of syntactically-linked word pairs, triples, and words computed on a large collection of English texts. Stanford Parser [5] was used for automatic parsing of sentences. 220 billion of words were processed; from this data we extracted 18 billion of syntactically linked word pairs and more than 65 billion of syntactically linked word triples. We also used synonyms from Wordnet [8] to compute formula (4).

Since our method is statistical, for error correction we decide to retain, besides the best paraphrase V^* of the source sentence S, all those paraphrases that have the value of *Degree* function greater than the *Degree* value of *S*. We call the resulted list of paraphrases ordered by *Degree* values *candidate corrections*. They should be suggested for a human editor, in order to make ultimate decision.

The correcting procedure sequentially performs the following steps for each sentence *S*:

Step 1. Parse sentence *S* and obtain its dependency parse tree.
Step 2. For each word from *S* generate its *Substitutes* set and compute conditional probabilities, using the formulas (4) and (6).
Step 3. For each word from *S* generate a paraphrase *V based* on the generated *Substitutes* set of the word, thus forming a set of paraphrases for *S*.
Step 4. Calculate the value of function for sentence *S* by formula (14).
Step 5. Calculate the value of function for all paraphrases *V* by formula (15).
Step 6. If some paraphrases have *Degree* value that exceeds *Degree* value of *S*, signal a collocation error.
Step 7. Suggest the ranked list of paraphrases with high values as *candidate corrections* for human editor

For evaluation of our method, we used 70 sentences with typical collocation errors (one error per each sentence) taken from ESL (English as a Second Language) materials. The sentences include erroneous collocations with words of various POS: prepositions, nouns, and adjectives. Besides erroneous sentences ESL materials contain examples of their proper corrections. Some examples of detected erroneous collocations and their proper corrections are presented in Table 2.

Table 2. Detected collocation errors and their corrections

Erroneous sentence	Proper Correction
I think it is a **spend** of my money.	I think it is a **waste** of my money.
To make **understandable**.	To make **plain**.
I have **done** a mistake.	I have **made** a mistake.
The jar was full **with** oil.	The jar was full **of** oil
This is great **painter**.	This is great **artists**.
The **ghost** of the opera.	The **phantom** of the opera.

The experiments showed that 80% of collocation errors were detected. For detected errors, 87% of candidate corrections lists included the proper correction.

In order to analyze the rank of proper corrections within the candidate corrections lists, we choose mean reciprocal rank (*MRR*), which is the arithmetic mean of the inverse ranks of the proper correction in the list:

$$MRR = \frac{1}{L}\sum\frac{1}{r} \tag{19}$$

where L is the number of sentences we used in our experiments, and r is the rank of proper correction in a candidate corrections list. If the list doesn't include proper correction, we assumed that r equals infinity. The Table 3 shows how the number K of corrected sentences depends on the size of candidate correcting list; the last column presents resulting MRR.

Table 3. Results of automatic evaluation MRR

Rank of proper correction	$r = 1$	$r \leq 2$	$r \leq 3$	$r \leq 100$	**MRR**
K	35	45	48	49	0.5

According to presented data the size of candidate correcting list may be equal to 2–3. We should also note that some candidate correcting lists included correct alternative paraphrases that are different from the proper correction.

Conclusions and Future Work

We proposed a novel method for collocation errors correction in learners' writing, based on assumption that the strategy of word-by-word translation is used by authors of the texts. The method automatically generates possible correcting paraphrases, for this purpose it uses translation dictionary from the native language of learners to the language of the text under correction. A relevance degree function was proposed to estimate generated paraphrases and to detect an error; the function is evaluated from the statistics of word syntactic links.

We implemented and evaluated our method supposing only one collocation error in the sentence. The experiments gave promising results: our method detected about 80% of collocation errors and 87% of proposed correcting paraphrases included proper correction.

Directions of our future research are:

- to use Bayesian networks to make the detecting procedure more efficient and test our method on sentences with several collocation errors;
- to expand *Substitutes* sets with word forms and homophones in order to detect additional type of collocation errors.

References

1. *Bolshakova E. I., Bolshakov I. A.* (2007) Automatic detection and computer-aided correction of Russian malapropisms [Avtomaticheskoe obnaruzhenie i avtomatizirovannnoe ispravlenie russkikh malapropizmov], Nauchnaya i Tekhnicheskaya Informatsiya. Ser. 2, No. 5, 2007, p. 8–13.
2. *Brockett C., Dolan W., Gamon M.* (2006) Correcting ESL Errors Using Phrasal SMT Techniques, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics ACL.
3. *Brown P., Pietra S., Mercer R., Pietra V.* (1993) The Mathematics of Statistical Machine Translation, Computational Linguistics, Vol. 19(2).
4. *Gamon M. (2010) Using Mostly Native Data to Correct Errors in Learners' Writing: A Meta-Classifier Approach, Proceeding HLT '10 Human Language Technologies.*
5. *Hermet M., Désilets A., Szpakowicz S.* (2008) Using the Web as a Linguistic Resource to Automatically Correct Lexico-Syntactic Errors, Proceeding The 6th Edition of the Language Resources and Evaluation Conference (LREC 06).
6. *Manning C., Jurafsky D.* Stanford Parser 2012 [html] (http://nlp.stanford.edu/software/lex-parser.shtml).
7. *Oxford Collocations Dictionary* for Students of English. Oxford University Press, 2003.
8. *Wordnet* 2013 [html] (http://wordnet.princeton.edu/)
9. *Wu J., Yu-Chia Chang Y., Teruko Mitamura T., Chang J.* (2010) Automatic Collocation Suggestion in Academic Writing, Proceedings of the ACL 2010 Conference Short Papers.
10. *Yi X., Gao J., Dolan W.* (2008) A Web-based English Proofing System for English as a Second Language Users, Proceedings of IJCNLP.

СЕМАНТИКА УГРОЗЫ В ЛИНГВИСТИЧЕСКОЙ ЭКСПЕРТИЗЕ ТЕКСТА

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

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

Ключевые слова: речевой акт, угроза, лингвистическая экспертиза

SEMANTICS OF THREAT IN FORENSIC LINGUISTICS

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

V. V. Vinogradov Russian Language Institute RAS, Moscow, Russia

The paper considers the semantics and pragmatics of threat as a speech act. In lexical semantics, the concept of a threat is often explained as a unified (single) notion. It is shown that speech acts of threat in Russian are divided into two types: threat-penalty and threat-warning. The latter type of threat — threat-warning — has a specific variety — threat-compulsion. Threat-penalty is a kind of a threat situation in which something bad occurred and speaker informs the hearer (who is responsible for this) that he will be punished. Threat-warning presupposes that no bad thing has occurred yet and the speaker shows the hearer that he should not do this bad thing. The realization of threat-compulsion assumes that the speaker tries to force the hearer to do something under threat of penalty.

Distinguishing the three kinds of threat is important for forensic linguistics. In cases of extremism, murder, bribe, exaction and others articles of law detection of body of the crime presupposes an analysis of criminal intention, which is reflected apart from everything else in kinds of threat.

Implicit ways of threatening are the most complicated to analyze in forensic linguistics. The analysis of implicit threat presupposes that all parts of semantic representation of this speech act (variables with terms and constants) should be identified in the text. The paper focuses on the case of implicit threat. The specific feature of the case analyzed consists in the implicit expression of penalty.

Key words: speech act of threat, forensic linguistics, lexical semantics

1. Семантика речевого акта угрозы и его типы

Выявление семантики угрозы в диалоге и письменном тексте необходимо при проведении экспертных исследований по ряду составов преступлений, предусмотренных российским законодательством. В частности, это преступления, связанные с подстрекательством, вымогательством, угрозой убийства, с принуждениями разного рода, а также по делам об экстремистской деятельности, взятках, рейдерстве и пр.

Угрозы как речевые действия (речевые акты) разделяются на два основных типа: угрозы-наказания и угрозы-предупреждения. **Угрозы-наказания** реализуются в ситуациях, когда адресат сделал что-то нежелательное для говорящего и он предупреждает адресата о наказании, которое за это последует[1]:

(1) *В октябре 1918 года Литке был командиром полка особого назначения, и В. Трифонов часто отдавал ему разного рода письменные распоряжения и приказания <…>. В одной записи, например, за какое-то нарушение дисциплины он* ***грозил*** *предать весь командный состав полка суду полевого трибунала.* [Ю. Трифонов. Старик]*;*

(2) *Начальник Канцелярии принял слова Данилова на свой счет, бился в ужасном гневе <…>,* ***грозил*** *упечь Данилова в расплавленные недра Земли.* [В. Орлов. Альтист Данилов].

Угрозу-наказание в примерах указанного типа можно истолковать следующим образом:

X угрожает$_1$/грозит$_1$ Y-у, что сделает P = ‘X говорит Y-у, что сделает P — нечто плохое для Y-а, из-за того, что Y сделал что-то плохое X-у (= Q), чтобы Y боялся P и осознал, что нехорошо делать плохое для X-а (в том числе и Q) и не делал это в будущем’.

Основной смысловой акцент в семантической экспликации угрозы-наказания в том, чтобы заставить адресата бояться наказания. Как следует из толкования, при возможности повторения нежелательного действия Q угроза-наказание может выполнять и воспитательную функцию (как и самое наказание как таковое). Угроза-наказание не относится ни к прямым, ни к косвенным побуждениям, хотя из компонента ‘чтобы Y боялся P и <…> и не делал это в будущем’ выводится следствие ‘X не хочет, чтобы Y делал плохое X-у’, которое может влечь побуждение Y-у не делать плохого для X-а. Впрочем, это побуждение довольно слабое. Тем самым выявление в диалоге или письменном тексте угрозы-наказания не указывает, например, на ситуацию вымогательства. Действительно, нечто плохое Q уже сделано. Однако для экстремистского дискурса угроза-наказание вполне органична.

---

[1] Поскольку далее рассматривается речевой акт угрозы, то в иллюстрациях используются и глагол *грозить* и глагол *угрожать*. Различия между ними относительно невелики.

В ряде случаев угроза-наказание эксплуатируется и при вымогательстве вымогательства. В этом случае квалификация Q, сделанного Y-ом, как плохого для X-а носит декларативный, притворный характер и используется как фактор давления на Y-а для получения от него желаемого — денег, уступки прав на собственность и пр. В этом случае семантическая структура вымогательства усложнена.

Иная ситуация в случае угроз-предупреждений. **Угрозы-предупреждения** используются в тех случаях, когда говорящий понимает, что адресат может совершить нечто нежелательное для него (или связанных с ним лиц) и пытается предотвратить это. Ср. примеры следующего типа:

(3) — *Все доводил.* ***Грозил****, если не скажу кто любовник, то убьет.* — *И ты сказала? — Да. Он про тебя все знает.* [Е. Кукаркин. Рассказы и повести]*;*

(4) *Великий Урфин приказал особенно зорко сторожить этих заключенных и за небрежность* ***грозил*** *страшными карами.* [А. Волков. Огненный бог Марранов].

Для угрозы-предупреждения можно предложить следующую семантическую экспликацию:

X угрожает$_2$ / грозит$_2$ Y-у, что сделает P, [чтобы Y не делал Q] = ‘X говорит Y-у, что сделает P — нечто плохое для Y-а, если Y сделает Q — что-то плохое для X-а, чтобы Y боялся P и из-за этого не делал Q’.

Здесь побуждение семантически вполне прозрачно: оно следует из компонента ‘чтобы Y боялся P и из-за этого не делал Q’. Впрочем, и в этом случае оно остается косвенным, хотя и вполне очевидным.

Угроза-понуждение — это вариант угрозы-предупреждения, реализующийся в ситуации, когда говорящий хочет от адресата не бездействия (отказа от совершения поступка), а действия: *Совсем не помню, что я отвечала этому полковнику. Кажется, я больше молчала, только изредка повторяя: «Не подпишу!». Он то* ***грозил****, то уговаривал, обещал свидание с мужем, с детьми.* [Е. Гинзбург. Крутой маршрут]. Угроза-понуждение в целом описывается приведенным толкованием угрозы-предупреждения с тем отличием, что угроза понуждение часто преследует интересы адресата (разумеется, в понимании говорящего), ср. *Если не сделаешь уроки, не пойдешь гулять!* Впрочем, это не мешает говорящему бездействие адресата рассматривать как событие, нежелательное для самого говорящего.

Угроза-понуждение не идентична шантажу, поскольку последний представляет собой более узкий глас угроз: это такие угрозы-понуждения, которые связаны с обнародованием компрометирующей информации (см. по этому поводу также [Гловинская 2004]).

Следует отметить, что семантика глагола *угрожать* и его синонимов весьма подробно и обстоятельно рассмотрена в [Гловинская 2004]. Основная

идея описания значения данного глагола (и соответствующего речевого акта) сводится в этой работе к тому, что у глагола *угрожать* не выделяется разных значений, а в качестве толкования берется их общая часть: *X угрожает <грозит> Y-у, что сделает P* = ‘X говорит Y-у, что сделает плохое для Y-а P, чтобы Y боялся P’. В качестве обоснования такого подхода привлекаются контексты типа *В пьяном виде он часто угрожал <грозил, грозился > поджечь квартиру; Сколько раз, когда становилось особенно невмоготу, он грозился бросить все и уехать,* в которых нежелательное для адресата действие говорящего, предположительно, вызвано не поведением адресата, а просто раздражением, усталостью, агрессивностью и другими особенностями психического состояния говорящего. Анализ показывает, однако, что приведенные контексты, конечно, редуцированы. В реальных ситуациях такого рода нежелательное положение дел, действие, состояние и т. д. имеется. Другое дело, что адекватность интерпретации говорящим (или субъектом действия) действительности может вызывать сомнения, но негативный фактор Q, за который, по мнению говорящего (или субъекта действия), несет ответственность адресат или какая-то обобщенная инстанция (судьба, фатум, рок и пр.), есть. Ср. расширенные примеры, лексически приближенные к приведенным выше редуцированным контекстам:

(5) *Савушкин **грозился** бросить все к чертовой бабушке. Он не может позволить себе роскошь мучиться два-три года, чтобы получить отрицательные результаты. Ему нужно защитить диссертацию. У него семья, дети.* [Д. Гранин. Иду на грозу]*;*

(6) *Муга рассказывал, как после второй стражи в дверь постучали именем короля, и Уно кричал, чтобы не открывали, но открыть все-таки пришлось, потому что серые **грозились** поджечь дом.* [А. и Б. Стругацкие. Трудно быть богом].

В примере (6) Q — это требование открыть дверь. В примере (5) ситуация несколько более сложная, однако вполне понятная, если знать содержание повести: Савушкин стремится защитить диссертацию, однако его научный руководитель Данкевич, талантливый ученый, находится в творческом поиске, он пробует различные варианты, не гарантирующие успеха. Савушкин, угрожая, конечно, имеет в виду Данкевича, но не может ему это сказать сам, предпочитая выпускать гнев и недовольство в присутствии Крылова, хотя угрозы направлены Данкевичу.

Представляется, что расширение примеров и привлечение достаточного контекста при употреблении глаголов *угрожать* и *грозить* всегда позволяет восстановить нежелательное для говорящего действие, событие и т. д., с которыми так или иначе связан адресат. Так, в ситуации, описываемой фразой *В пьяном виде он часто угрожал <грозил, грозился > поджечь квартиру*, отрицательную реакцию пьяницы может вызвать неправильно понятое поведение членов семьи, окружающих, друзей, коллег и т. д.

Отметим, что в [Гловинская 2004] выделяется целый ряд релевантных семантический оппозиций, по которым различаются синонимичные глаголы *угрожать, пригрозить, грозить* и *грозиться*. Однако неразличение угроз-наказаний, угроз-предупреждений и угроз-понуждений осложняет использование семантической экспликации угрозы, предложенной М. Я. Гловинской, в лингвистической экспертизе текста. Более того, эта экспликация не вполне отвечает и языковой интуиции. Разграничение указанных типов угроз как речевых актов существенно и с чисто лингвистической точки зрения, поскольку прослеживается в изменении условий успешности. Что касается собственно толкований глаголов *угрожать* и *грозить*, это требует особого обсуждения.

Рассмотрим пример выявления семантики угрозы в практике лингвистической экспертизы.

2. Case study: скрытая угроза

Выявление угрозы предполагает установление компонентов акта угрозы, то есть актантов, заполняющих соответствующие валентности, и фиксированных частей семантической экспликации этого речевого акта (констант). Наибольшую сложность представляют случаи, когда угроза не передается прямо, а по понятным причинам скрывается. Рассмотрим пример выявления скрытой угрозы.

Фабула дела такова. Лицо М2 (рейдер) пытается побудить лиц М1 и М3 к продаже недостроенного здания. Свое коммуникативное намерение М2 эксплицитно выражает в форме речевого акта предложения, используя для этого эксплицитную перформативную формулу: *М2 <…> Значит, в связи с этим, да, мы* ***предлагаем*** *договариваться, да. <…>; — Мы* ***предлагаем*** *урегулировать эту ситуацию <…>; В этой связи, да, мы* ***предлагаем****, значит, ситуацию эту выкупить.* Однако эксплицитное указание на коммуникативное намерение может быть неискренним в том случае, когда говорящий вынужден скрывать криминальный характер своих действий. Неискренность в некоторых случаях может быть установлена в результате анализа семантики и прагматики языковых форм, используемых говорящим.

Жанр исследуемой беседы — это переговоры («торг») относительно купли-продажи здания. Ситуация «купли-продажи», как известно из лексической семантики, задается четырьмя участниками: покупателем, товаром, продавцом и ценой за товар: «Глаголы <…> — *купить* и *продать* — обозначают ситуацию купли-продажи <…>: в результате этих операций тот, кто платит, приобретает нужную ему вещь в свою собственность» [НОСС: 357]. Для дальнейшего изложения важны некоторые семантические свойства ситуации купли-продажи, следующие из приведенного определения:

1) продавец в обычном случае является владельцем товара или, по крайней мере, имеет разрешение от владельца на продажу, что признается участниками ситуации купли-продажи **(Свойство 1)**;

2) при купле-продаже имеется консенсус между участниками относительно того, что является товаром, то есть стороны (покупатель и продавец) должны быть уверены, что под товаром имеется в виду одно и то же **(Свойство 2)**;

3) участники ситуации купли-продажи добровольно вступают в роли покупателя и, соответственно, продавца **(Свойство 3)**.

Анализ показывает, что свойства 1, 2 и 3, обязательные для ситуации купли-продажи, в исследуемой беседе между лицом М2, с одной стороны, и лицами М1 и М3 — с другой, не выполняются.

Свойство 1 не выполняется из-за того, что участник М2 не признает, что участники М1 и М3 являются собственниками недостроенного строения:

М1 — Для нас цель понятна. Мы владельцы инвестиционного проекта. <…> Девелопментская компания с именем. М2 — ***Вы владельцы инвестиционного проекта, а-а, с непонятным происхождением и с очень хилыми вариантами какого-то дальше развития*** <…> .

М3 — <…> Есть дом, оформленный в собственность на нашу компанию. М2 — ***Но есть ряд определенных решений, которые аннулируют массу, значит, правовых оснований владения этим домом*** <…> ***Есть собственность, которая находится в данный момент …***[2] ***в спорах*** <…>.

М3 — То есть Вы предлагаете купить у нас нашу собственность, правильно или нет? М2 — Мы предлагаем урегулировать эту ситуацию, потому что ***это не собственность, не собственность.*** <…> ***Вы позиционируете себя владельцем здания*** <…> М2 — Хорошо, на вашем языке, может быть, да, это … продать здание, да. Я***, все-таки, не считаю, пока еще это как бы окончательным товаром.*** <…> Именно так. ***Есть конкретная ситуация, которая не является товаром, которую нельзя продать.***

Таким образом, в приведенных репликах участник М2 передает следующее содержание:

- высказывает сомнения в юридической «чистоте» актива *(с непонятным происхождением и с очень хилыми вариантами какого-то дальше развития)*;
- указывает на то, что здание является объектом судебных споров *(собственность, которая находится в данный момент … в спорах)* и, тем самым, не может быть продано *(Есть конкретная ситуация, которая не является товаром, которую нельзя продать)*;
- выражает уверенность в том, что право собственности будет аннулировано судебными решениями *(это не собственность, не собственность)*;
- отмечает, что участники М1 и М3 только «позиционируют» себя владельцами здания *(вы позиционируете себя владельцем здания)*, не являясь таковыми на деле,
- и, наконец, резюмирует, что спорное здание не является товаром *(Я, все-таки, не считаю, пока еще это как бы окончательным товаром)*.

---

[2] Отточия без скобок указывают на паузы.

В нормальном случае отказ покупателя признавать, что объект продажи принадлежит продавцу, сразу прекращает процесс торга. Между тем, участник М2, отрицая наличие собственности на здание у участников М1 и М3, все равно продолжает диалог, стремясь склонить участников М1 и М3 к продаже того, что он не считает их собственностью.

Свойство 2. Вторая важная характеристика ситуации купли-продажи — это общий взгляд на то, что предлагается к продаже — на товар. Из текста разговоров видно, что если участники М1 и М3 считают в качестве возможного товара недостроенное здание, то участник М2 — некую «ситуацию»:

М2 — Вот. Э-э. Значит, наши предложения здесь фактически исходят из того, что ***мы хотим купить …ситуацию***, да, как-то ее урегулировать со всех сторон, зачистить и, соответственно, заработать всем нам на этой ситуации;

М2 — Мы готовы, значит, как-то ***эту ситуацию а-а купить***, да, с вами разойтись.

В следующем фрагменте участник М3 выражает удивление по поводу интерпретации участником М2 объекта продажи как ситуации:

М3 — (Дело в том, что) ***купить ситуацию — это что-то там на вашем сленге.*** Мы такого как бы не очень приемлем, да? ***Есть дом, оформленный в собственность на нашу компанию.***

Несмотря на возражения участников М1 и М3, участник М2 вновь и вновь возвращается к выбранному им способу категоризации (понимания) объекта продажи:

М2 — <…> В этой ***связи***, да, мы предлагаем, значит, ***ситуацию эту выкупить***.

М2 — Есть ***конкретная ситуация***, которая не является (товаром), которую (нельзя) продать. Мы хотим иметь товар, да?

Коммуникативная цель номинации объекта продажи как *«ситуации»* заключается в том, чтобы указать на судебные дела, которые связаны с недостроенным зданием:

М3 — <…> Есть дом, оформленный в собственность на нашу компанию. М2 — Но ***есть ряд определенных решений, которые аннулируют*** … М3 — Что аннулируют? М2 — ***Которые аннулируют массу, значит, правовых оснований владения этим домом*** <…>.

М2 — ***Есть правовые основания признания недействительными этих прав,*** правильно?

М1 — Есть собственность, официальным, должным законным образом оформленная. М2 — ***Есть собственность, которая находится в данный момент … в спорах***.

Различное понимание товара — того, что покупается в ситуации купли-продажи — обычно ведет к прекращению обсуждения сделки, однако в рассматриваемом случае этого не происходит: участник М2 продолжает коммуникацию, повторяя свои аргументы с минимальным перефразированием.

Еще одно свойство ситуации купли-продажи — Свойство 3 — готовность участников выступать в качестве покупателя и продавца. В данном случае участник М2 выступает в роли покупателя, однако участники М1 и М3 неоднократно заявляют о том, что они не собираются продавать принадлежащее им здание:

М3 — Да нет. ***Дом не продается просто, как бы нет смысла.*** Вы можете из-за *забора посмотреть, оценить район и так далее*. М2 — <…> Просто у нас разное позиции, да. М3 — ***Мы ничего не продаем, мы строили для себя.***

Опять-таки обычно заявление одной из сторон, что она не хочет участвовать в сделке, прекращает переговоры (торг). Однако в данном случае участник М2 маниакально настаивает на продолжении обсуждения:

М2 — Ну хорошо. Да, давайте … (согласуем) позиции. <…> Сейчас нам просто, да, нужно понимать, как нам, значит, можно заинтересовать друг друга, какими цифрами. Да? Просто принципиально ***у вас подход такой: вы ничего не продаете, да, но вы говорите двадцать миллионов из расчета того, что это гипотетически может быть продано, да, … цены вот такой.*** <…> ***— Давайте, давайте прямей будем, да? То, что может вас устроить в принципе.*** М3 — Да нет. Нас вряд ли может устроить эта цифра, когда мы должны имущество, которое мы не предназначали, так сказать, к продаже, ну, Вам, тем более, продавать.

М3 — Значит, ***Вы сами себя ввели в заблуждение. В основном Вы просто хотите денег заработать, поэтому а-а обсуждать здесь нечего абсолютно нечего.*** М2 — ***А у меня другое мнение, я считаю, что нам есть что обсуждать.***

Различие позиций сторон с точки зрения участия в ситуации купли-продажи особенно отчетливо видно в последней паре реплик. Участник М3 говорит, что *«обсуждать здесь нечего, абсолютно нечего»*, то есть в явном виде отказывается от сделки, а участник М2, напротив, настойчиво повторяет: *А у меня другое мнение, я считаю, что нам есть что обсуждать*. Такая настойчивость участника М2 сама по себе могла бы просто указывать на большое желание участника осуществить сделку, однако в сочетании с отрицанием участником М2 права собственности участников М1 и М3 на объект купли-продажи и различным пониманием участником М2, с одной стороны, и участниками М1 и М3 — с другой, того, что является предметом торга («ситуация» и недостроенное здание), переговоры о купле-продаже здания предстают как абсурд в духе Ионеско.

Объяснить выявленную семантическую аномальность исследуемого разговора можно в том случае, если предположить, что коммуникативная цель участника М2 заключается в скрытой передаче какого-то иного коммуникативного намерения, модифицирующего ситуацию торга при обсуждении купли-продажи товара или дополняющую ее.

Рассмотрим в этой связи речевой акт угрозы-побуждения, проанализированный выше. При наложении сформулированного выше толкования на ситуацию, представленную в исследуемой беседе, можно получить следующую семантическую экспликацию:

- X (угрожающий) — М2;
- Y (тот или те, кому угрожают, адресат угрозы) — М1, М3;
- не Q (действие, желательное для угрожающего) — согласие продать здание, принадлежащее М1, М3, участнику М2.

Содержание, или сущность, угрозы (Р) передается участником М2 в следующих фразах:

М2 — ***Мы включаем свои механизмы. Мы теперь используем свои ресурсы для того, чтобы добиться в этом проекте максимального эффекта.***

М2 Как бы суть понятна, как бы наша позиция здесь ясна, да, правда за нами. <…> Есть сложившаяся на настоящий момент ситуация, которая дает нам основания полагать, опять же, что мы здесь выиграем, да. И ваши свидетельства, ваши методы, <…> скажем так, (вилами на воде), да. При успешном завершении всех разбирательств судебных, да, мы понимаем, что ***и можем все это организовать как самовольную постройку*** и ***согласование префектов, с адресом, все эти махинации, да.***

М1 — Вот этот документ, согласование, которое префект <…> М2 — ***Но этот документ может быть признан нулевым.*** <…> ***Значит, нормально довести это мы вам не дадим.***

Содержание угрозы (Р) сначала довольно неопределенно: это максимальное использование имеющихся у М2 (компании, которую он представляет) ресурсов *(Мы включаем свои механизмы. Мы теперь используем свои ресурсы для того, чтобы добиться в этом проекте максимального эффекта)* и создание препятствий завершению строительства *(Значит, нормально довести это мы вам не дадим)*. В двух других репликах — это более конкретные действия М2: признание (видимо, по суду) правоустанавливающих документов как не имеющих силы *(Но этот документ может быть признан нулевым)*, а также создание ситуации, когда здание будет признано самостроем *(и можем все это организовать как самовольную постройку)*.

В обычном случае угроза-понуждение предполагает, что содержание угрозы (Р) выражается в более явном виде. В данном случае это не так: угроза завуалирована, скрыта — ее суть излагается в разных репликах и сопровождается лексическими маркерами модальности возможности *(**может** быть признан; **можем** все это организовать как самовольную постройку)*. Чтобы

обеспечить ее действенность («серьезность»), участник М2, как было показано выше, модифицирует свойства ситуации купли-продажи, то есть, во-первых, он отказывается признавать недостроенное здание товаром и, во-вторых, считает объектом купли-продажи (товаром) не здание, а «ситуацию», то есть здание, обремененное судебными тяжбами.

Проведенный анализ, выявивший наличие скрытой угрозы в репликах участника М2, позволяет объяснить, почему этот участник не прекращает попытки продолжения переговоров, когда участники М1 и М3 прямо заявляют, что не собираются продавать здание (см. выше анализ Свойства 3). Использование угрозы-понуждения — хотя и в скрытой форме — позволяет участнику М2 надеяться на продолжение переговоров и принятие решения участниками М1 и М3 в свою пользу.

Дополнительным доказательством наличия элементов семантики угрозы в рассматриваемых разговорах может служить то, что на прямые утверждения участников М1 и М3 о том, что участник М2 представляет рейдеров, последний не отрицает такой квалификации:

М3 — Понятно. Вы Виктора Ломакина знаете давно? М2 — Но это, как бы, не имеет значения, имеет значение лишь договоренности, которые есть на данный момент. М3 — Нет. Дело в том, что это имеет значение. Вот ***Виктор Ломакин — это рейдер***, да? М2 — ***Ну да.*** М3 — ***Вы рейдер на рынке более известный, чем он.*** То есть один рейдер боролся три года, ничего у него не получилось. Он пришел к более сильным рейдерам, так сказать, продолжать борьбу. М2 — ***Нам, да, нам интересно все, что может принести прибыль.*** А какое время мы знакомы с Ломакиным — не по существу. Вопрос прибыли и вопрос экономики. Вот и все.

Участник М3 говорит, что Ломакин и участник М2 (его компания) являются рейдерами, а М2 с этим соглашается. Использование угроз — пусть и в скрытой форме — естественно вписывается в методы действия рейдеров.

Отметим, что выявленная угроза является угрозой-побуждением, а не угрозой-наказанием или угрозой-предупреждением. Данный факт существенен для квалификации речевых действий участника М2 как рейдера. Тем самым, разграничение типов угроз необходимо для полного и точного анализа семантики текста.

Литература

1. *Гловинская М. Я.* (2004), Словарная статья «Угрожать, пригрозить, грозить, грозиться», в Новый объяснительный словарь синонимов русского языка, Языки славянской культуры, М., с. 1190–1194.
2. *НОСС* — Апресян Ю. Д., Апресян Ю. Д., Бабаева Е. Э., Богуславская О. Ю. и др. (2004), Новый объяснительный словарь синонимов русского языка, Языки славянской культуры, М.

References

1. *Glovinskaja M. Ja.* (2004) Lexical entry “Threaten, menace, promise” [Slovarnaja statja “Ugrozhat’, prigrozit’, grozit’, grozit’sya”], in New explanatory dictionary of synonyms of Russian [Novyj ob”asnitelnyj slovar’ sinonimov russkogo jazyka], Jazyki slavjanskoj kultury, M.
2. *NOS* — Apresjan Ju. D., Apresjan Ju. D., Babajeva E. È., Boguskavskaja J. Ju. (2004), New explanatory dictionary of synonyms of Russian [Novyj ob”asnitelnyj slovar’ sinonimov russkogo jazyka], Jazyki slavjanskoj kultury, M.

КОРПУС КАК ЯЗЫК: ОТ МАСШТАБИРУЕМОСТИ К ДИФФЕРЕНЦИАЛЬНОЙ ПОЛНОТЕ

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

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

Пиперски А. Ч. (apiperski@gmail.com)
РГГУ, Москва, Россия

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

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

Основным вопросом всякого корпусного исследования, будь то эксперименты с Интернетом, работа с НКРЯ или иным корпусом, должен быть вопрос об объекте наблюдения: изучается конкретный корпус, поисковая машина или собственно язык? К сожалению, почти всегда исследователь принимает в качестве не требующего доказательства предположения «масштабируемость» результатов частного корпусного исследования на весь язык. В статье рассматриваются критерии, с помощью которых разработчики корпусов обосновывают возможность такого масштабирования частных корпусных данных и предполагается новый подход к оценке границ действия обнаруженных исследователем фактов, принятый в рамках продолжающегося проекта создания Генерального интернет-корпуса русского языка (ГИКРЯ). Одним из базовых положений этого проекта является идея, что само масштабирование результатов является операцией весьма ограниченного применения. Для большинства лингвистических и лексикографических задач корпусной анализ должен проводиться с точностью до четко определенных жанровых и социолингвистических границ.

CORPUS AS LANGUAGE: FROM SCALABILITY TO REGISTER VARIATION

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

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

Piperski A. (apiperski@gmail.com)
RSUH, Moscow, Russia

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

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

The main research question of any corpus investigation, either while experimenting with the Internet or working with the RNC or any other corpus, should be the question of the object of investigation: do we study a particular corpus, search engine or the language "overall"? Unfortunately, researchers usually accept as self-evident the assumption of "scalability" of the results obtained with a specific corpus study to the whole body of language. The article examines the criteria to justify the possibility to scale specific data and proposes an approach to assessing the limits of discovered facts, as adopted in the framework of an ongoing project to create the General Internet Corpus of Russian (GICR). One of the basic ideas of this project is that scaling the results is a very limited operation. For the majority of linguistic and lexicographical problems, corpus analysis should be carried out within a well-defined genre and sociolinguistic parameters.

О проекте ГИКРЯ[1]

Данная статья является продолжением работы [Беликов, Селегей, Шаров 2012], в которой обосновывалась необходимость запуска еще одного корпусного проекта для русского языка и определялись основные технологические принципы создания сверхбольшого корпуса на основании полностью автоматических методов сбора и лингвистической, и метатекстовой разметки.

---

[1] Проект ведется при финансовой поддержке Министерства образования и науки Российской Федерации (гос. контракт № 07.514.11.4142) и программы стратегического развития РГГУ.

В данной работе мы не касаемся вопросов технологии корпусного строительства. Рассматриваются только вопросы оценки достоверности корпусных исследований. ГИКРЯ в данный момент находится в фазе сбора первой тестовой версии, позволяющей тем не менее проводить отдельные лингвистические исследования, некоторые результаты которых будут представлены ниже.

Объем этой версии в данный момент составляет около 4 млрд слов, по составу текстовая версия ориентирована на исследования блогосферы (более 20 млн записей и комментарии к ним) и актуальной художественной литературы и публицистики (около 45 тыс. текстов объемом 250 млн слов).

Масштабируемость и сбалансированность

Скрупулезный анализ адекватности полученных результатов не стал еще, к сожалению, частью культуры корпусных исследований. Этот удивительный факт находится в разительном противоречии с популярностью и значимостью этих исследований в современной лингвистике.

Большое количество работ основаны на абсолютном доверии к полученному количественному результату, часто вся выводная часть строится на сопоставлении частот, полученных запросом по всему доступному корпусному материалу, будь то НКРЯ или Интернет. Типичное резюме выглядит следующим образом:

Материалом для исследования стали данные корпуса N. Количество вхождений для каждой из сравниваемых конструкций составило (приводятся несколько цифр, часто одного порядка). Далее следует вывод *о приоритете конструкции с большей частотой в исследуемом языке*. Вопросы, которые обычно остаются без внимания:

- сравнение данных по числу вхождений, документов и авторов;
- анализ временной динамики;
- анализ распределения результатов по типам источников (параметрам метатекстовой разметки);
- наличие дублетов или иных систематических факторов, «накручивающих» счетчик.

Обобщение частных корпусных результатов на весь язык (включая и отрицательные результаты!) является господствующей тенденцией, которую в значительной степени поддерживают сами создатели корпусов.

Если исследователь может найти в корпусе примеры на интересующее его явление, то его выводы во многом зависят от того, как позиционируется этот корпус. В названии некоторых корпусов содержится информация об их составе (например Michigan Corpus of Academic Spoken English), и пользователи таких ресурсов едва ли рискнут масштабировать полученные результаты на весь язык. Однако корпуса, позиционирующие себя как национальные, намного сильнее навязывают своим пользователям представление, что по ним можно делать выводы про язык в целом.

Обычно корпуса пытаются так или иначе обосновать эту претензию. Возможность масштабирования выводов на язык связывается с понятиями сбалансированности и представительности/репрезентативности, которые часто

упоминаются в описаниях корпусов. Иногда эти понятия рассматриваются как эквивалентные. Например, создатели Национального Корпуса Русского Языка (НКРЯ) указывают [НКРЯ, 2012]:

> *Национальный корпус … характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус* ***по возможности пропорционально их доле в языке соответствующего периода*** [выделение наше].

Едва ли не текстуально совпадающие автохарактеристики можно найти на сайтах других национальных корпусов, например, болгарского и британского[2].

Как создатели корпусов доказывают наличие указанных свойств? Здесь успехи достаточно скромны. Даже НКРЯ, увы, «не защищен» солидными формальными доказательствами сбалансированности и репрезентативности. Имеется лишь одна работа, посвященная статистическому анализу НКРЯ в целом [Шаров, Ляшевская, 2009]. При создании частотного словаря, основанного на текстах этого корпуса, применялись определенные принципы подсчета частот, позволяющие доказать лексическую сбалансированность этого ресурса (сегментирование корпуса и подсчет частот с учетом равномерности распределения по сегментам). Но оценка сбалансированности в целом не была дана, было показано только, что относительные частоты в ядре языка устроены в НКРЯ близко к тому, что показывает Интернет.

Чтобы доказать свою представительность и сбалансированность, некоторые корпуса эксплицитно указывают принципы отбора текстов: например, Корпус современного американского английского языка — COCA за каждый год начиная с 1990-го включает примерно по 4 млн словоформ в устный, художественный, журнальный, газетный и научный подкорпуса. Другие корпуса просто сообщают свой состав как данность. Так поступает, например, Венгерский национальный корпус, который приводит информацию об объеме текстов в зависимости от их жанровой и географической принадлежности[3]. Наконец, есть корпуса, которые ограничиваются лишь общими рассуждениями о типологическом разнообразии и не указывают конкретных цифр (в их число входит НКРЯ).

В целом, апробированных научным сообществом способов обеспечения представительности и сбалансированности корпусов не предложено ([Hunston 2008]; [McEnery, Hardie 2011]). Эта тема активно обсуждается в компьютерной лингвистике (ср. [Biber 1993], [Leech 2007]), но пока что получается, что корпус

---

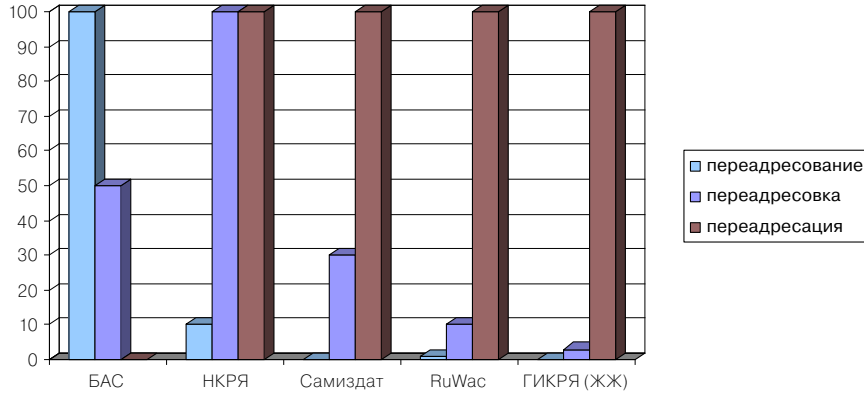
[2] http://www.ibl.bas.bg/BGNC_classific_bg.htm; http://www.natcorp.ox.ac.uk/corpus/index.xml.

[3] Эта информация суммирована в таблице 5×5 (**жанры**: пресса, литература, наука, официальные тексты, частная сфера; **регионы**: Венгрия, Словакия, Прикарпатье, Трансильвания, Воеводина).

считается представительным и сбалансированным тогда, когда на этот счет имеется негласный договор между его создателями и пользователями.

Масштабируемость и состав корпуса

Препятствием для масштабирования исследовательского результата на весь язык (что бы ни понималось под таковым — см. далее) является не только типологическая несбалансированность (несоответствие пресловутой доле в языке), но и случайность подбора текстов каждого типа, неизбежная при ограниченном объеме корпуса, а также лексическая неактуальность корпусного контента. На графике ниже показано примерное соотношение приведенных частот употребления 3-х вариантов именования понятия: *переадресование* (основной вариант по данным словарей[4]) — *переадресовка* — *переадресация*.



Эти данные, к анализу которых мы вернемся несколько позже, показывают, что два корпуса, претендующих на представительность, НКРЯ (ручная сборка) и RuWac (полностью автоматическая сборка), показывают весьма различающиеся данные, что плохо соотносится с идеей масштабирования[5]. Различие картин мира БАС («национального» толкового словаря) и блогов ГИКРЯ не требует комментария.

---

[4] В БАС (т. 15, 2011) и в словаре Ефремовой (2006) *переадресовка* отсылает к *переадресование*; в словаре Шведовой *переадресовка* помечена *разг*. *Переадресация* в толковых словарях отсутствует.

[5] Вообще говоря, количество вхождений слов в НКРЯ (до 20) делает результат в некотором смысле случайным, чувствительным к процедуре подсчета, но что самое важное, никак не демонстрирует процесс ухода варианта *переадресовка*.

Недокументированные особенности интернет-поиска

Максимально возможным приближением к полному *множеству* текстов является Интернет. Однако, как было показано в [Беликов, Селегей, Шаров 2012], Интернет не является *корпусом*, поскольку не предоставляет адекватных инструментов доступа и анализа содержащихся в нем текстов.

Это не означает, что интернет-поисковиками не нужно или тем более — нельзя пользоваться. Необходима аккуратная постановка эксперимента, позволяющая надежно верифицировать полученные результаты. Это возможно только при ясном понимании того, как этот результат получен.

Некритичное использование Интернета для лингвистических целей стало беспокоить и самих разработчиков этого ресурса, по крайней мере тех, кто имеет дело с лингвистами. Так, в лекции о принципах поиска Яндекса в Политехническом музее [Плахов 2012], было в очередной раз, но теперь — уже из самых первых уст сказано, что любые цифры, которые нельзя перепроверить вручную (до 1000 поисковых результатов) являются результатами не прямых подсчетов, а различных аппроксимаций. Не зная алгоритмов аппроксимации и их параметров, используемых на момент поиска, добросовестный лингвист не может полагаться на результаты поиска, которые часто можно рассматривать как случайные.

Мы не станем перегружать текст доказательствами этого тезиса, приведем только один новый пример, полученный авторами в ходе сравнения данных ГИКРЯ и Яндекса. В таблице ниже видны не только обычные нарушения аксиом арифметики и алгебры, но и совершенно неожиданные и не имеющую ничего общего с изучаемым явлением зависимости результата поиска от предполагаемого или явно указанного местонахождения исследователя (все скриншоты сохранены авторами):

	«украину»	«на украину»	«в украину»
Поиск от 12.08.2011 в Угловке Новгородской обл.			
без ограничения региона	136 млн	310 млн	321 млн
Поиск от 14.03.2013 в Петербурге			
без ограничения региона	3 млн	138 тыс.	196 тыс.
✓ в Санкт-Петербурге	2 млн	951 тыс.	2 млн
Поиск от 15.03.2013 в Москве			
без ограничения региона	5 млн	4 млн	14 млн
✓ в Москве	69 млн	3 млн	6 млн

Работая с поисковиками, исследователь использует в качестве рабочего инструмента то, что программисты назвали бы недокументированными возможностями (не описанными в официальных руководствах и свободными от обязательств поддержки). Это относится как к общим алгоритмам поиска и подсчета частот отдельных словоформ и коллокаций, так и к специальным инструментам «лингвистического» анализа, которые периодически появляются.

Например, в течение некоторого времени Яндекс поддерживал специальный инструмент «Пульс блогосферы». В феврале 2013 г. он был закрыт как

непопулярный. Между тем среди лингвистов он был довольно популярен. В готовящемся в изд. НЛО сборнике «Русский язык как глобальный ресурс и новые технологии» он независимо упоминается двумя авторами. Б. В. Орехов называет его «удобным электронным инструментом» для исследования микроистории лексики. М. А. Кронгауз, основываясь на данных «Пульса блогосферы», пишет о динамике популярности интернет-мемов. Между тем, популярность мема в соответствующей среде может расти, а его частотность в постоянно перестраивающейся в социальном отношении блогосфере — падать. При этом принципы подсчета частоты совершенно скрыты от пользователя[6].

Масштабируемость как цель

Таким образом, в основе представлений о возможности масштабирования результатов поиска в корпусе лежат три идеи о его контенте:

1. о его сбалансированности (остается, как мы показали, совершенно неформальным свойством корпуса);
2. о его актуальности;
3. о его общем количестве, достаточном для генерализации выводов (также совершенно неформальный критерий).

Но что именно получает ведомый идеей масштабируемости исследователь, работая пусть даже с идеальным корпусом, удовлетворяющим этим условиям?

Он полагает, что можно обращаться к корпусу как к самому языку с вопросами, ответы на которые будут относиться также ко всему языку (получая нечто вроде усредненной частоты явления в языке).

Но о каком потенциальном множестве текстов идет речь, когда создатели корпуса говорят о «доле в языке» некоторого типа, представленного в корпусе? Является ли идеалом сбалансированности гипотетический полный корпус текстов, содержащий все написанное на данном языке в некотором временном промежутке (с точностью до фактов устной речи, требующей фиксации иного типа)?

Хорошо известно мнение некоторых лингвистов о том, что «Интернет — это большая помойка, в нем полно неграмотных текстов». В этом случае Интернету противопоставляется некоторый свод текстов на Правильном Русском Языке (ПРЯ) со своей идеальной картиной типового распределения. Получение корпуса этого ПРЯ является задачей, достойной с какой угодно точки зрения, но не с точки зрения лингвистики и лексикографии.

Заметим, что методы недифференцированного анализа корпуса часто применяются в компьютерной лингвистике, где условная сбалансированность позволяет создавать модели, эффективные в среднем (обученные на предполагаемой в деятельности пользователей жанровой и тематической взвеси).

---

[6] Объем статьи не позволяет входить в детали, для интересующихся более полные данные опубликованы в расширенном варианте статьи на сайте конференции Диалог.

На наш взгляд вопрос о «процентном составе языка» является лингвистически совершенно бессмысленным, и любые полученные цифры отражают процессы, имеющие малое отношение к лингвистике и даже социолингвистике. Так, если в практике создания текстов носителями языка в данном историческом периоде на один протокол осмотра происшествия приходится 0,01 некролога, 10 медицинских заключений и 1000 «объявлений о знакомстве», вряд ли разумно требовать от корпуса, чтобы и в нем соблюдалось то же соотношение.

Более «справедливая» модель отбора, известная со времен Ноева ковчега, также порождает при запросах «в среднем» совершенно неадекватную картину.

Кроме того, относительно самого набора «корпусных тварей» пока не наблюдается ни малейшего единодушия [Sharoff, 2010]. Количество категорий классификации текстов варьирует от 15 (Брауновский корпус), 70 (БНК), 181 (НКРЯ), 349, отобранных в исследовании предпочтений пользователей [Crowston, Kwasnik, Rubleske, 2010] до более 4 тысяч [Adamzik, 1995].

О неоднородности корпусных данных

Принятие гипотезы о масштабируемости корпуса порождает исследовательские ошибки не только из-за несбалансированности, недостаточности объема и прочих бед, о которых было сказано. Самой большей проблемой является то, что даже идеально сбалансированный корпус будет приводить исследователя к неверным выводам, если не учитывается принципиальная неоднородность языковых данных. Некоторая типология явлений, требующих дифференциального подхода была дана в работе [Беликов, Селегей, Шаров 2012].

К сожалению, сегодня дифференциальные исследования остаются на периферии интересов лингвистов и лексикографов. Одно из немногих исключений — масштабное исследование региональной вариативности в русском языке в проекте «Языки русских городов» (http://community.lingvo.ru/goroda/, [Беликов 2010]).

Удивляет то, что недифференциальный подход свойственен даже многим работам по собственно вариативности. Исследователи рассматривают дивергенцию в языке как некоторое о б щ е я з ы к о в о е свойство, не пытаясь анализировать параметры этой вариативности. Например, на сайте НКРЯ объявлен научный проект по изучению вариативности «Проблемы русской стилистики по данным НКРЯ» (http://studiorum.ruscorpora.ru/stylistics/intro.html). Приведены несколько дающих проектные установки работ, остановимся только на одной: М. В. Шкапа «Склонение топонимов на *-ово (-ево), -ыно (-ино)*». Автор подсчитывал количество вхождений в текстах за 2000–2010 гг., работая с корпусом в 50,0 млн слов. Уже то, что автор работает с вхождениями, а не документами, резко снижает ценность исследования. Так, один только появившийся после исследования текст Льва Дурнова «Жизнь врача» на 40% увеличил число склоняемых вхождений топонима *Перово*. Кроме того, газетный корпус НКРЯ составлен

из публикаций семи изданий, в разном объеме и за разные годы; редакционная политика изданий заметно влияет на результаты. Все это указывает на то, что решаются не проблемы *русской стилистики*, а проблемы стилистики *конкретного собрания текстов*.

Но важнее всего то, что многие такие частные проблемы решать «в среднем по корпусу» просто не следует. Так, даже беглый анализ вариативности в склонении топонимов по данным блогосферы показывает, что на нее влияют самые разные факторы. Прежде всего — региональные (см. версию статьи на сайте). Имеют место различия в выборе варианта для своего и чужого топонима, и т.д. Для усмотрения каких-то тенденций и параметров, которые на них влияют, нужен существенно больший материал, чем имеется в НКРЯ.

Мы провели небольшое исследование по сравнительному анализу двух условно сопоставимых по жанровой смеси и объему корпусов: художественных текстов НКРЯ и журнального подкорпуса ГИКРЯ, полученного на основании текстов Журнального зала (magazines.russ.ru).

Помимо упомянутого выше анализа употребления тройки «*передресация* и т. п.», исследовалась структура значения слова *окучивать* по данным этих корпусов. Выяснилось, что на собранном вручную (НКРЯ) и автоматически (ГИКРЯ) корпусах обнаруживаются заметные отличия, связанные, например, с обилием остросюжетных текстов в современной части НКРЯ, представленных в основном романами (при поиске *окучивать* их оказалось более половины) и известной «однобокостью» поэтического корпуса НКРЯ[7]. В результате в ГИКРЯ среди 142 вхождений глагола *окучивать* 5,5% оказывается в поэзии, на традиционное агротехническое значение падает 65% всех вхождений, а в НКРЯ они составляют лишь 45% от 42 вхождений.

Причин этих отличий несколько: от серьезного влияния невыявленных дублетов (к каковым в случае НКРЯ относятся, например, многократные повторы об условиях подписки на издание), до влияния неслучайности отбора (тут могут быть важны и соображения доступности, авторских прав и т. п.).

Выводы просты: создатели корпусов должны добиваться независимости любого подкорпуса, выделенного по какой-либо группе параметров, например, жанровой, от процедуры отбора корпусного материала. То есть, корпус должен обнаруживать устойчивость в результатах анализа по тем свойствам, которые являются объектом исследования. В корпусах небольшого объема добиться такой устойчивости можно только резким сужением круга изучаемых явлений.

От сбалансированности к дифференциальной полноте

Как показано выше, масштабируемость является не универсальным свойством корпуса, а частным следствием из доказанной однородности корпусных данных относительно данного языкового факта.

---

[7] 203 наболее поздних текста, датированные периодом (1981–1995), принадлежат трем авторам 1903–1907 г. р.: И. В. Чиннову (126), Г. А. Глинке (74) и А. А. Штейнбергу (3).

Для того, чтобы вывод о масштабируемости был надежным, нужен корпус с особыми свойствами. Вместо понятия сбалансированности мы предлагаем использовать понятие *дифференциальной полноты* корпуса.

В отличие от понятия сбалансированности-репрезентативности, дифференциальная полнота означает не просто предположительную типологическую полноту корпуса, но и

- явное указание типа за счет метатекстовой разметки и использование типологических данных при обработке запросов;
- доказуемые предположения относительно необходимого и достаточного объема текстов каждого типа в корпусе.

В дифференциально полном корпусе результат обработки любого запроса может быть разложен по типологическим координатам и оценен с точки зрения однородности. Разумеется, при желании исследователь может дать веса типам для того, чтобы получить какие-то обобщенные данные, соответствующие его представлению о составе языка.

Вопрос о формальных критериях дифференциальной полноты корпуса не является решенным. Собственно говоря, это является одной из основных «математических» целей проекта ГИКРЯ.

Мы можем говорить о достаточной дифференциальной полноте корпуса относительно некоторого языкового явления в следующих смыслах:

- при наличии метатекстовой разметки (аналог обучения с учителем, supervised learning), которая сама может быть порождена автоматическими методами с разной степенью надежности;
- при отсутствии такой разметки, как оценка максимального варьирования параметров в рамках автоматически определенных кластеров (аналог обучения без учителя, unsupervised learning).

В первом случае, например, дифференциальная полнота по регионам определяется наличием в корпусе статистически значимого количества текстов, авторы которых происходят из представительного списка регионов; полнота по жанрам предполагает наличие значимого количества текстов по всем жанрам, которые должны быть представлены в корпусе.

С другой стороны, этот вид полноты предполагает, что все категории классификации известны нам заранее. Это явно не имеет места для жанровой классификации, и даже для классификации региональной мы не можем быть априорно уверены в значимости параметра дробности для отдельных категорий. Мера полноты может быть оценена выбором признаков (вся лексика, избранная лексика ключевых слов, частеречные коды и т.п.), кластеризацией корпуса и оценкой различий между различными кластерами [Sharoff, 2007].

При этом существенным для определения значимых различий между сегментами корпуса (либо определенными по метатекстовой разметке, либо кластеризацией) является:

- выбор языкового явления, которое измеряется в числовом выражении и может быть надежно извлечено автоматическими методами на большом корпусе (обычно это частота леммы, словоформы или грамматической конструкции);
- количество примеров этого явления в рамках отдельных текстов каждого из сегментов (частота по документам по сегментам);
- числовое выражение явления в рамках документов в сегменте.

В любом случае, полнота представлена не как абсолют, а как относительное понятие: один корпус полнее другого (или дает более достоверную картину) в рамках того или иного языкового явления, либо два корпуса более похожи по степени своей полноты в сравнении с третьим.

Рассмотрим пример для региональных частот. С точки зрения статистики мы исходим из предположения («основной гипотезы» в терминах статистики), что два региональных сегмента в сравнимых жанрах не отличаются по использованию лексемы X. Мы можем оценить:

1. статистическое распределение частот употребления этой лексемы в текстах каждого региона и
2. определить значимость «альтернативной гипотезы» о различии в этих частотах.

Пункт 1 считается как вероятность по документам (частота в документе, деленная на его размер), в результате мы получаем распределение частот по текстам каждого региона. Если f_d — частота термина в документе d, а N_d — размер этого документа, то:

$$p_d = \frac{f_d}{N_d}$$

Пункт 2 считается стандартными методами аппроксимирования альтернативной гипотезы, например, используя t test распределения вероятностей, либо коэффициент логарифмического правдоподобия отклонений абсолютных частот (a и b) в двух сегментах (размерами соответственно c и d), который учитывает как абсолютные значения, так и относительные размеры каждого сегмента:

$$E1 = c\frac{(a+b)}{(c+d)}$$

$$E2 = d\frac{(a+b)}{(c+d)}$$

$$G^2 = a\log\frac{a}{E1} + b\log\frac{b}{E2}$$

В результате сравнение разных результатов для разных регионов или феноменов может быть оценено с точки зрения вероятности того, что «альтернативная гипотеза» верна не случайным образом. Например, значение G^2 большее 3,84 дает 95 % вероятность того, что разница не случайна [Rayson, Garside, 2000].

Заключение

Основным содержательным результатом проекта ГИКРЯ через год после начала систематической работы авторов над этим проектом можно считать появление первой тестовой версии корпуса, позволяющей верифицировать результаты корпусных исследований (проводимых как в рамках существующих корпусов РЯ, так и в Интернете). ГИКРЯ пока едва достиг 10 % планируемого объема, но уже сейчас дает исследователю результаты, каждый шаг получения которых является совершенно прозрачным. Другим важным результатом является формирование понятия дифференциальной полноты. Что касается ее вычислимых критериев, тут предстоит еще большая работа, требующая нескольких итераций сбора корпуса и коррекции используемых классификаторов. Важным результатом станет возможность тестирования лингвистической устойчивости корпуса при расширении его состава и возможность объективного сравнения свойств разных корпусов.

Литература

1. *Adamzik, K.* 1995. Textsorten — Texttypologie. Eine kommentierte Bibliographie, Nodus, Münster. http://www.unige.ch/lettres/alman/adamzik/akt/naments.pdf
2. *Biber D.* (1993), Representativeness in Corpus Design, Literary and Linguistic Computing, Vol. 8, No. 4, pp. 243–257
3. *Crowston, K., Kwasnik, B., Rubleske, J.* 2010. Problems in the use-centered development of a taxonomy of web genres. // Mehler, A., Sharoff, S., Santini, M., editors, Genres on the Web: Computational Models and Empirical Studies. Springer, Berlin/New York.
4. *Hunston S.* (2008), Collection strategies and design decisions, in A. Lüdeling and M. Kyto (eds.), Corpus linguistics: an international handbook, Walter de Gruyter, Berlin/New York, pp. 154–168.
5. *Leech G.* (2007), New resources, or just better old ones? The Holy Grail of representativeness, in M. Hundt, N. Nesselhauf and C. Biewer (eds.), Corpus Linguistics and the Web, Amsterdam, Rodopi, pp. 133–49.
6. *McEnery T., Hardy A.* (2011), Corpus linguistics, Cambridge University Press, Cambridge, 2011.
7. *Rayson, P. and Garside, R.* 2000. Comparing corpora using frequency profiling. // Proceedings of the workshop on Comparing Corpora at ACL 2000, Hong Kong.

8. *Sharoff, S.* 2007. Classifying Web corpora into domain and genre using automatic feature identification. // Proc. of Web as Corpus Workshop, Louvain-la-Neuve.
9. *Sharoff, S.* 2007. Creating General-Purpose Corpora Using Automated Search Engine Queries.
10. *Sharoff, S.* 2010. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Mehler, A., Sharoff, S., Santini, M., editors, Genres on the Web: Computational Models and Empirical Studies, Springer, Berlin/New York.
11. *Беликов* 2010. Методические новости в социальной лексикографии XXI века // Slavica Helsingiensia 40. Instrumentarium of Linguistics. Sociolinguistic Approaches to Non-Standard Russian. Helsinki, 2010 / Ed. by A. Mustajoki, E. Protassova, N. Vakhtin.
12. *Беликов В. И., Селегей В. П., Шаров С. А.* 2012. Пролегомены к проекту Генерального интернет-корпуса русского языка. // Труды конференции Диалог 2012.
13. *НКРЯ. Что такое Корпус?* Сайт Национального корпуса русского языка. http://www.ruscorpora.ru/corpora-intro.html. 2012 г.
14. *Ляшевская О. Н., Шаров С. А.* Частотный словарь современного русского языка Издательство Азбуковник, 2009. Вступительная статья http://www.dialog-21.ru/digests/dialog2008/materials/html/53.htm.
15. *Плахов А.* Системы поиска в Интернете: как обрабатывается запрос пользователя. Лекция в политехническом музее 16.10.12. http://pmlectures.ru/video/Cistemy_poiska_v_Internete_kak_obrabatyvaetsya_zapros_polzovatelya-38

КОМПЬЮТЕРНЫЙ АНАЛИЗ КОНСТРУКЦИЙ С ГЛАГОЛОМ ПОДДЕРЖКИ В РУССКОМ И ИТАЛЬЯНСКОМ ЯЗЫКАХ

Ключевые слова: глаголы поддержки, оценка компьютерных систем для анализа лингвистических корпусов, семантическая разметка

COMPUTATIONAL TREATMENT OF SUPPORT VERB CONSTRUCTIONS IN ITALIAN AND IN RUSSIAN[1]

Benigni V. (valentina.benigni@uniroma3.it)
Università "Roma Tre", Roma, Italia

Cotta Ramusino P. (paola.cottaramusino@unimi.it)
Università degli Studi, Milano, Italia

We aim at comparing some corpora-based computational resources that enable us to analyse the collocational profiles of the SVCs in both languages. The resources include SketchEngine, which works for both languages, Lexit for Italian and NKRJA for Russian.

The case study focuses on the Italian verb *mettere* followed by a prepositional phrase with the prepositions *in* and *a*, and the corresponding Russian verb *stavit'/postavit'* followed by a prepositional phrase with the prepositions *v* and *na*.

We discuss the options offered by the tools at the syntax-semantic interface. A closer comparison of the three tools shows that they provide different data. We have explored some aspects of the semantic tagging of Lexit and NKRJA and propose an integration of the two tools. It seems that further development of semantic tagging could be helpful in creating Italian-Russian lexicographic resources.

Keywords: Support Verb Constructions, evaluation of corpus-based computer resources for linguistic research, semantic tagging

[1] The article is the result of close collaboration between the two authors, whose names are listed in alphabetical order. For academic purposes only, Valentina Benigni is responsible for sections 1, 2.2.1, 2.2.3, 4.1, 4.2 and Paola Cotta Ramusino for sections 1.1, 2.1, 2.2.2, 3.1, 3.2.

1. Introduction

In this paper, devoted to Support Verb Constructions (henceforth SVCs) in Italian and Russian (cf. also Benigni & Cotta Ramusino 2011), we aim at comparing some corpora-based computational resources, that enable us to analyse the collocational profiles of SVCs in both languages. These resources are: SketchEngine, which works for both languages, Lexit for Italian only and NKRJA for Russian only.

The subject of the present case study is the Italian verb *mettere* "to put" followed by a prepositional phrase with the prepositions *in* "in, into" and *a* "to", and the corresponding Russian verb *stavit'/postavit'* followed by a prepositional phrase with the prepositions *v* and *na*.

A closer comparison of the three tools shows that they provide different data. We suggest that the integration of this data could be of great help in creating Italian-Russian lexicographic resources.

1.1. Support Verb Constructions: a case study

In Benigni & Cotta Ramusino 2011 we identified a number of morpho-syntactic criteria (tests) on the basis of which we could distinguish SVCs from free constructions. We then attempted at categorizing SVCs with the Italian verb *fare* "to make" by the following steps:

- first of all, we divided all SVCs into semantic classes within which the SV has the same meaning and combines with a N object that has similar semantic characteristics;
- then we grouped these classes into larger actional classes, according to Vendler's classification.

The data, which needed further processing, was obtained using a CQL query in SketchEngine. One of the most significant results of that classification was to propose relevant parameters for identifying SVCs, i.e., first of all the semantic class of the direct object and, secondly, when dealing with particularly opaque constructions, the semantic and actional class of the whole construction, which could be better treated as a single lexical item (*fare mente locale* "to try to remember", lit. "to make local mind").

2. Computational tools for the Italian language

2.1. SketchEngine

SketchEngine is "a corpus query system incorporating **word sketches**, one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour" (http://www.sketchengine.co.uk). The Italian corpus available on SketchEngine is the *itTenTen Corpus* (3,076,908,415 tokens).

The query has been carried out by using the "WordSketch" function (henceforth WS). The WS function presents a list of the grammatical relations the word participates in and provides a list of collocates for each grammatical relation (subject, object, prepositional objects, modifying adverbs…).

As for indirect objects, always introduced by prepositions in Italian, WS allocates simple and compound prepositions to different patterns. We obtained the following profiles for each preposition: **pp_in-x, pp_nel-x, pp_nella-x, pp_nell'-x, pp_a-x, pp_al-x, pp_alla-x, pp_all'-x.**

This function does not account, however, for singular and plural, so that SVCs like *mettere nei guai* "to get into trouble", in which the filler is always plural, are assigned to the singular profile **pp_nel-x**.

Table 1. Prepositional complements of the Italian verb *mettere* according to SketchEngine

mettere in	**MI**	**mettere nel**	**MI**	**mettere nella**	**MI**	**mettere nell'**	**MI**
evidenza	10.68	panno	10.76	condizione	7.19	angolino	7.33
discussione	10.31	guaio	9.67	pentola	6.60	armadio	5.29
luce	9.74	mirino	8.81	ciotola	6.35	impossibilità	5.09
atto	9.53	cassetto	8.12	valigia	6.33	angolo	5.05
scena	9.49	calderone	8.07	mano	6.32	agenda	4.54
moto	9.26	carrello	7.93	padella	6.21	ottica	4.38
risalto	9.20	dimenticatoio	7.78	teglia	6.10	impasto	4.21
campo	9.16	culo	7.55	tasca	6.08	urna	3.76
guardia	9.12	sacco	7.18	bara	5.80	orecchio	2.74
pratica	9.08	frullatore	7.14	culla	5.66	animo	2.40
mettere a	**MI**	**mettere al**	**MI**	**mettere alla**	**MI**	**mettere all'**	**MI**
disposizione	11.69	riparo	10.53	berlina	10.24	asta	9.21
nudo	9.52	bando	9.36	gogna	9.81	indice	7.00
punto	9.46	corrente	8.87	stretta	9.63	incanto	6.98
fuoco	9.45	volante	8.85	prova	9.55	inseguimento	6.53
segno	9.02	rogo	7.94	corda	8.20	angolo	6.20
rischio	8.97	centro	7.77	frusta	7.46	odg	5.99
confronto	8.89	tappeto	7.36	guida	7.25	occhiello	5.12
frutto	8.50	collo	7.15	spalla	7.15	opera	4.72
prova	7.97	posto	7.12	porta	6.90	ordine	4.30
agio	7.44	fornello	7.08	calcagno	6.90	incasso	4.30

So, the WS function applied to the Italian corpus allows us:

- to single out verb collocates; in particular WS extracts the prepositional collocates to the right of the verb and the noun collocates both left and right of the verb. The collected data contains a substantial amount of noise: for example, one of the most frequent left collocates is the word *santa*, which is part of the multiword noun *santa messa* ("Holy Mass"), where *messa* is a noun and not the past participle of the verb *mettere*;
- to view the contexts in which the token occurs;
- finally, to infer some information about noun gender starting from the compound preposition gender, except for hyphened prepositions like *nell'* and *all'*, which can be either masculine or feminine.

On the other hand, the WS function does not allow us:

- to extract all possible syntactic frames within which the target lemma occurs (for instance, when extracting a prepositional phrase, it does not provide the user with information about presence or absence of a direct object);
- to separate reflexive and non-reflexive verb forms.

2.2. Lexit

Lexit is a corpus-derived lexical resource for the analysis of Italian verbs, nouns and adjectives that extracts distributional profiles at the syntax-semantic interface.

The current version of Lexit contains information gathered from two different corpora: the *La Repubblica* corpus (Baroni *et al.* 2004, about 331 millions tokens) and the Italian section of *Wikipedia* (ca. 152 millions tokens) (Lenci *et al.* 2012: 3713)

The resource has been developed by Lenci at the University of Pisa (Computational lab, Department of Linguistics) and is available at the address http://sesia.humnet.unipi.it/lexit.

The resource allows us to:

- Extract the syntactic frames[2] of a target lemma, going beyond the traditional distinction between argument and adjunct;
- Extract all the fillers of a target syntactic slot;
- Get the semantic classes of the fillers.

2.2.1. Syntactic frames

The syntactic frames of our target lemma *mettere* are extracted and ordered by decreasing values of LMI (Local Mutual Information, used to measure the association between verb and subcategorization frames, frame slots and their lexical fillers) (Lenci *et al.* 2012: 3713).

---

[2] Lenci refers to them as Subcategorization frames (SCF), i.e. "a pattern of syntactic dependencies headed by the target lemma" (Lenci 2012: 3713).

This is the list of frames individuated by Lexit:

subj#obj#comp-*in*
subj#obj#comp-*a*
subj#comp-*in*
subj#si#inf-*a*
subj#obj#comp-*su*
subj#obj#comp-*sotto*
subj#si#comp-*in*
subj#si#obj#comp-*in*
subj#comp-*sotto*
subj#si#comp-*a*

The extracted data needs further analysis, given that Lexit treats reflexive and non-reflexive forms as the same lemma (*mettersi/mettere*), while allocating them to different frames. Lexit also shows that the same prepositional phrase can occur within different frames, i.e. the **compl-in** is used both with and without a direct object.

By clicking on a specific frame we obtain general information about the lexical fillers in the different syntactic slots of the frame. It is not possible, however, to combine information about a specific argument filler with the data concerning the fillers of other complements.

2.2.2. Lexical fillers

The "slot function" allows extraction of all the lexical fillers of the prepositional phrases compl-*in* and compl-*a*, regardless of the syntactic frame in which they occur.

Table 2 shows the first 20 fillers of **compl-*in*** and **compl-*a*** by decreasing values of LMI.

Table 2. Prepositional complements of the Italian verb *mettere* according to Lexit

mettere in	**LMI**	**mettere a**	**LMI**
discussione	31,804	punto	49,883
moto	27,716	disposizione	45,121
scena	18,489	segno	21,385
evidenza	17,300	repentaglio	10,141
guardia	15,668	prova	9,488
luce	14,715	fuoco	8,927
piede	11,978	riparo	8,202
difficoltà	10,969	confronto	7,083
dubbio	10,125	posto	6,774
pericolo	9,307	bando	6,724
crisi	9,285	nudo	5,305

mettere in	**LMI**	**mettere a**	**LMI**
atto	8,712	asta	4,220
vendita	7,794	lavoro	4,112
ginocchio	7,550	rischio	4,084
campo	7,479	frutto	3,701
condizione	7,034	soqquadro	3,153
pratica	6,213	verbale	3,009
risalto	5,578	servizio	2,873
cantiere	5,372	portafoglio	2,411
conto	5,217	mondo	2,159

The "slot function" shows some shortcomings:

- It is possible to view the lexical set of the most prototypical fillers, but it is not yet possible to see the examples in context (the resource is being implemented);
- Lexit currently does not supply information about the preposition (whether simple or compound): *mettere in scena* vs *mettere nei guai* ("to put on stage" vs "to get into trouble").
- Lexit does not supply information about the number value of the noun, which can be singular or plural: *mettere in un guaio* vs *mettere nei guai* (lit. "to put into a trouble" vs "to put into troubles").

(These last two shortcomings limit the use of the tool in self-learning, because the data needs further checking to get rid of redundant results.)

- At the moment, Lexit does not supply information on the position of the prepositional object or the presence of other lexical items between verb and prepositional phrase (e.g. *mettere in un guaio* vs *mettere in un grosso guaio*; lit. "to put into a trouble" vs "to put into a big trouble"), as the frames are formed by unordered sets of slots representing the syntactic constituents.

2.2.3. Semantic classes

As previously specified, Lexit not only provides the lexical set for the most prototypical fillers for each syntactic slot; it also supplies semantic information on the semantic classes (ordered by LMI) to which the nouns belong.

Nouns are classified into 24 groups[3] corresponding to the 24 "top-nodes" dominating the semantic noun ontology in the Italian section of MultiWordNet (Pianta *et al*. 2002), a multilingual lexical database linked to WordNet and structured in hierarchically organized semantic classes.

Unfortunately, this semantic resource is not fully developed as yet, so that we cannot see which fillers correspond to each semantic class, and the link between filler and semantic class has to be reconstructed by manually checking on MultiWordNet the top-node corresponding to each filler.

[3] Animal, Artifact, Act, Attribute, Food, Communication, Knowledge, Body Part, Event, Natural Phenomenon, Shape, Group, Location, Motivation, Natural Object, Person, Plant, Possession, Process, Quantity, Feeling, Substance, State, Time.

This hitherto partially developed semantic classification could be of great help both in teaching and in NLP, as it provides an array of lexical fillers for a given slot.

Therefore, in terms of semantic classification of the fillers, which appears to be Lexit's potentially strong point, we would like to point out that:

- The link between semantic classes and fillers has been carried out automatically without disambiguation;
- The automatic processing does not account for regular polysemy, whereby the same word can be linked to different top-nodes, for instance the filler *posto* "place" is present in the same prepositional **compl-a** with three different syntactic profiles: *mettere* $X_{[+anim]}$ *a posto* "to put sb into place", *mettere* $X_{[-anim]}$ *a posto* "to put sth in order", *mettere* X *al posto* di Y "to put X in the place of Y". In the first and in the second case *posto* "place" is a State, because the entire construction *a posto* "in place" has this meaning, in the third case *posto* is a Location, albeit a metaphorical one.
- In the same way, the words *mano* "hand", *bocca* "mouth", *testa* "head", *piede* "foot", *ginocchio* "knee", are linked to the top-node BodyPart, nevertheless in SVCs such as *mettere in mano* "to put in the hand", *mettere in bocca* "to put in / into mouth", *mettere in testa* "to put in / into head", they acquire the meaning Location, whereas in SVCs as *mettere in piedi* "to set sthg up", lit. "to put on feet", *mettere in ginocchio* "bring to one's knees" they mean Position.
- Although there are different criteria for semantic classification, it seems more appropriate to classify fillers based on productive categories: for example there are many fillers of **compl-in** which could be classified as Location (*prigione* "prison", *galera* "jail", *pista* "track") or as "position" (*fila* "line", *linea* "row", *cerchio* "circle"), but there are no other SVCs, apart from those listed above, with a BodyPart acting as filler of **compl-in** (with the exception of idiomatic expression *mettere la pulce nell'orecchio* "plant a seed of doubt", lit. "to put a flea into the ear").
- Even if the primary meaning of words like *mano, bocca* is BodyPart, their coerced meaning within the SVC is Location or Position, so it would probably be more appropriate to link them to these semantic classes, or, at least, to add a semantic tag which could account for the semantic shift, something like BodyPart — Location, BodyPart — Position.
- Lastly, the semantic classification does not supply information about the degree of idiomaticity of a SVC, which is often due to the desemantization of the filler, because of metonymic or metaphorical processes (for instance *mettere in palio* "to raffle" lit. "to put as a flag". It would seem therefore more useful to tag these constructions as a single lexical item.

3. Computational resources for the Russian language

3.1. SketchEngine for the Russian language

The Russian corpus uploaded on SketchEngine is the *ruTenTen* Corpus (20,162,118,568 tokens).

We used the WS function to carry out a query on the lemma *stavit'* "to put". The tool extracts:

- MI-ordered collocational profiles of the lemma, in particular internal, external arguments and prepositional phrases. The first problem is that for Russian SketchEngine easily mixes up internal and external arguments as it selects by default the noun on the left of the verb as Subject and the one on the right as Object. On the other hand, it distinguishes different internal arguments as type 1 (acc), type 2 (gen) and type 3 (gen part);
- Allomorphs of the same preposition, specifically *v/vo*[4] "in";
- Word contexts;
- The overall frequency and the MI, i.e. information on the collocational nature of the contructions.

Table 3. Prepositional complements of the Russian verb *ставить* according to SketchEngine

ставить в	**MI**	**ставить во**	**MI**	**ставить на**	**MI**
тупик	9.91	МХАТе	6.04	кон	8.34
известность	9.21	глава	5.57	огонь	6.79
духовка	8.85	всеуслышание	4.69	подоконник	6.39
упрек	7.46	фрунт	3.33	стол	5.91
холодильник	7.36	двор	3.03	повестка	5.84
вина	7.28	главу	2.88	полка	5.76
укор	6.88	Фронтеры	2.87	пауза	5.72
микроволновку	6.85	ГЛАВУ	2.84	колено	5.68
морозилка	6.32	флоп	2.81	ручник	5.6
кавычка	6.06	гла	2.62	подставка	5.54

3.2. NKRJA — Russian National Corpus

The NKRJA is a corpus of modern Russian incorporating 300 million words. Although it is not a computational resource for the extraction of statistical information on words and constructions, its rich morphological and semantic tagging makes it a useful tool for linguistic research, including identification of collocational profiles and systematic semantic patterns.

In particular, in this section we will discuss the functions supplied by the corpus' semantic tagging.

Semantic tagging is based on the classification system developed for the Lexicograph database from 1992 onwards under the leadership of Paducheva and Rakhilina (http://www.ruscorpora.ru/en/corpora-sem.html).

---

[4] It should be noted that among the results of *vo* complements quite a substantial amount of noise can be found, and both the overall frequency and MI are very low, so that the data is not fully reliable.

The set of semantic and lexical parameters is different for different parts of speech. Nouns, which are the POS we are dealing with here, are divided into three subclasses: concrete nouns, abstract nouns, and proper names, each one with its own hierarchy of tags.

Lexical and semantic tags are grouped as follows:

1. Taxonomy (a lexeme's thematic class) — for all nouns;
2. Mereology ("part — whole" and "element — aggregate" relationships) — for concrete and abstract nouns;
3. Topology — for concrete nouns;
4. Evaluation — for abstract and concrete nouns.

In the first place, in order to compare NKRJA with other tools, we carried out a search of the verb *stavit'* using morphological tagging. The query was as follows:

first slot
Word: ставить
Distance: from 1 to 1

second slot
Word: в /на
Distance: from 1 to 2
third slot
Gramm. features: noun, accusative

The result of this first query is the following: 1825 tokens with *stavit' na* "to put on" and 1271 with *stavit' v* "to put in/into"; unfortunately, this search does not supply any information regarding the MI of each filler.

We thus proceeded to the second step of the query by adding some semantic features, selecting them among those that showed the highest MI for the corresponding Italian SVC in Lexit. Results are as follows:

natural phenomena = 3
mental sphere = 224
space and places = 247
human body parts and organs = 10

A brief examination of the results reveals a substantial amount of noise: among abstract natural phenomena we find *sneg* "snow" (which is a natural phenomenon, but a concrete one) and *vedro* "bucket", which is neither a natural phenomenon nor an abstract noun, so it becomes clear that the results need further processing.

For this reason we decided to select more generic semantic features. In the second step we selected only abstract nouns and got 1067 results with *stavit' na* and 982 results with *stavit' v*. The choice of abstract nouns is associated with the peculiarities of SVCs, in which abstract nouns show a higher frequency.

The results have been processed manually and ordered by overall frequency.

Table 4 reports the first 20 results for each pattern:

Table 4. Prepositional complements of the Russian verb *ставить* according to NKRJA

ставить на + $N_{ACC[+ABSTRACT]}$	**FQ**		**ставить в+ $N_{ACC[+ABSTRACT]}$**	**FQ**
X[5] место / места	119		вину	84
(X) карту	85		пример	84
место	76		тупик	81
вид	52		положение / положения X	79
стол	44		известность	62
колени	31		X зависимость	46
голосование	29		связь с чем-то	40
(X) огонь / огонек =1	28		один ряд	38
X план	28		упрёк	30
пол	26		заслугу	19
кон	24		место / места	19
счет	15		соответствие	14
очередь	14		условия X	11
полку / полки	11		основу	9
сцену	11		ряд / ряды	8
повестку дня	10		затруднение	7
X почву	9		недоумение	7
обсуждение	9		необходимость	7
работу	9		строку	7
учёт	9		счет	7

We expected a clear-cut and fully reliable result, but on the contrary, we immediately spotted several concrete nouns (*ogon'* "fire", *stol* "table", *pol* "floor", *voda* "water", *škola* "school", *kotël* "pot").

Moreover, polysemy represents a serious problem for semantic tagging: among the results, there are nouns which acquire an abstract sense if used metaphorically or metonymically (*stavit' na kartu* "to stake", lit. "to put on the card", *stavit' na koleni* "bring/force sb to his knees", *stavit' v tupik* "to lead into a dead end", *stavit' v rjady* "to set in the ranks of"). In particular, semantic tagging does not account for nouns like *mesto* "place", which appears in contexts like *stavit' na X*[6] *mesto* "to put in X place" */ stavit' na mesto kogo-to/***čego-to** "to put in sb's/sth's place", where the noun means *place* and the verb maintains its primary meaning of motion verb, but also in constructions like *stavit' kogo-to na mesto* "to put sb into place" and *stavit'* **čto**-*to na mesto* "to put sth in order", where the noun exhibits the more abstract meaning of Position and the verb undergoes desemantization and functions as support verb.

---

[5] X indicates the presence of a modifier in that syntactic position.

[6] Cf. footnote 5.

NKRJA pros: the query system allows the user:
a) to query the different syntactic frames in which the predicate appears, but does not provide a list of them;
b) to find, by means of the *distance* function, more lexical items (adjectives, adverbs, and so on) within the frame, so that it is possible to obtain collocates that occur far away from their prepositional head (*stavit' v polnuju/polnejšuju/prjamuju zavisimost'* lit. "to put in full/fullest/direct dependence) and multiword expressions (*stavit' na pervyj plan* "to put in the foreground").
c) to refine the search by inserting more semantic features, although this function is not totally reliable, as shown by the examples above.

NKRJA cons: the query system does not provide the user with:
a) a list of all the possible frames in which the target lemma may appear;
b) the overall frequency and MI data of the filler, so that we don't get any information about the association of the analyzed words;
c) a semantic classification of the filler (what Lexit is trying to do): in other words we can verify the absence/presence of fillers with certain semantic features, but the tool does not tell us which semantic classes occupy a given syntactic slot. We can obtain this information by manually processing the data.

4. Three different computational resources: final remarks

4.1. The collocational profile

In this paper we discussed how three different computational resources can provide information for linguistic research. We chose SVCs for testing purposes, since for this kind of structures both morpho-syntactic and semantic features are relevant (and should be identified by the tools). As far as morpho-syntactic features are concerned, both Lexit and SketchEngine, although to different degrees, extract the syntactic and collocational patterns of the target verb. The data, when subject to further manual processing, results in a selection of SVCs. In particular, Lexit extracts a more exhaustive list of the syntactic pattern of a target lemma. On the other hand, NKRJA allows a query of the syntactic frames within which the predicate occurs, but does not identify or list them all.

4.2. The semantic profile and semantic tagging

With respect to semantic classification, which could introduce significant advantages both for the creation of lexicographic resources and for self-learning and teaching activities, we should make some observations on hitherto unsolved problems.

In particular, with respect to Lexit we were confronted with:

- the absence of a link between semantic class and actual filler;
- a choice of semantic classes which is not always functional in creating a classification with a predicting character (a case in point being the BodyPart tag);
- the association of the filler with top-nodes that were too high in the hierarchy.

On the contrary, NKRJA provides the user with a very refined semantic tagging, although:

- it is not fully satisfactory yet, especially when the manual disambiguation of homonimy has not been carried out;
- it requires an "expert" use of the query system.

Moreover, at present, both tools fail to reflect the polysemy produced at the *parole* level, where lexemes may acquire new senses, by metonymical or metaphorical processes (a phenomenon usually called polysemy (Apresjan 1974), coercion (Pustejovsky 1993), deferred reference (Nunberg 1995)).

NKRJA seems to suggest a partial solution to this problem, as it allows, once you obtain the query results, to click on a given token and check all the assigned tags. Semantic tagging is organized in a two-tier system, which includes *main* and *secondary* semantic features, which accounts for both the connotation level and the semantic shift, as shown by the semantic tags for the lemma *tupik*:

Semantics main	r:concr, t:space
Semantic shifts	ev:neg, r:abstr, r:concr, t:space

This kind of two-tier semantic tagging suggests that it could be possible to reflect this hierarchy in the query system.

Finally, we observed that information on preposition semantics could be included in the semantic tagging. In Italian, for instance, *in* can have different meanings:

- it describes a movement towards sth (≃into) in SVCs like *mettere in testa* "to put in/into head",
- it can refer to the way sth is done (≃how) in SVCs like in *mettere in ginocchio* "bring to one's knees".

Both patterns are regular and productive: *in* (≃into) occurs also in other SVCs, like *mettere in tasca* "to put in pocket", *mettere in galera* "to put in jail", whereas *in* (≃how) occurs in SVCs like *mettere in difficoltà* "to hinder sb", lit. "to put in difficulty", *mettere in pericolo* "to put in danger".

At the same time, we observe the same regularity in Russian; *v* means:

- "towards" or "into" sth, in SVCs like *stavit' v tupik* "to lead sb into a dead end", *stavit' v kavyčki* lit. "to put into inverted commas";
- "in which way", "how" sth is done, in SVCs like *stavit' v rjad* lit. "to put in line", *stavit' v parallel'* "to put in parallel", *stavit' v zatrudnenie* "to hinder sb", lit. "to put in difficulty";

- "as, like" in SVCs like *stavit' v primer* "to cite as an example", *stavit' v osnovu* "to assume sth as a basis".

We thus suggest that further research and implementation of these computational resources should focus first of all on semantic tagging: the link between fillers and semantic classes, as in Lexit, and the semantic tagging by means of a two-tier system adopted by NKRJa, seem to be useful devices in clarifying polysemy. Moreover, as previously observed, prepositions show a high degree of semantic regularity and distributional similarity, even cross-linguistically, and we maintain that this kind of tagging which takes into account large lexical contexts should be implemented.

References

1. *Apresjan Ju. D.* (1971), Regular polysemy [O reguljarnoj mnogoznachnosti], Bulletin of the Academy of Sciences of the USSR: Section of. Language and Literature [Izvestija Akademii Nauk SSSR: Otdelenie literatury i jazyka], T. XXX, Nauka, Moskva, pp. 509–523.
2. *Baroni M., Bernardini S., Comastri F., Piccioni L., Volpi A., Aston G., Mazzoleni M.* (2004), Introducing the "La Repubblica" corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian, Proceedings of LREC 2004, Lisboa, pp. 1771–1774.
3. *Benigni V., Cotta Ramusino P.* (2011), Italian constructions with support verb *fare* in comparison with russian [Italjanskie konstruktsii s glagolom podderzhki *fare* v sopostavlenii s russkin jazykom],
4. *Computational* Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011"], Bekasovo, pp. 68–83.
5. *Lenci A., Lapesa G., Bonansinga G.* (2012), LexIt: A Computational Resource on Italian Argument Structure, available at: www.lrec-conf.org/proceedings/lrec2012/pdf/622_Paper.pdf.
6. *Nunberg G.* (1995), Transfers of meaning, Journal of Semantics, Vol. 12 (2), pp. 109–132.
7. *Pianta E., Bentivogli L., Girardi C.* (2002), Multiwordnet: developing an aligned multilingual database, Proceedings of the First International WordNet Conference, Mysore, pp. 293–302.
8. *Pustejovsky J.* (1993), Type Coercion and Lexical Selection, in Pustejovsky J. (ed.), Semantics and the Lexicon, Kluwer Academic Publishers, Dordrecht, pp. 73–94.

МОРФОЛОГИЧЕСКАЯ РАЗМЕТКА КОРПУСА СИЛАМИ ВОЛОНТЁРОВ

Бочаров В. В. (bocharov@opencorpora.org),
Алексеева С. В. (alexeeva@opencorpora.org),
Грановский Д. В. (granovsky@opencorpora.org),
Протопопова Е. В. (protoev@gmail.com),
Степанова М. Е. (mariarusia@gmail.com),
Суриков А. В. (ksurent@opencorpora.org)

Проект «Открытый Корпус», OpenCorpora.org

Ключевые слова: корпус, разметка, морфологическая омонимия, краудсорсинг

CROWDSOURCING MORPHOLOGICAL ANNOTATION

Bocharov V. V. (bocharov@opencorpora.org),
Alexeeva S. V. (alexeeva@opencorpora.org),
Granovsky D. V. (granovsky@opencorpora.org),
Protopopova E. V. (protoev@gmail.com),
Stepanova M. E. (mariarusia@gmail.com),
Surikov A. V. (ksurent@opencorpora.org)

OpenCorpora.org

Manually annotated corpora are very important and very expensive resources: the annotation process requires a lot of time and skills. In OpenCorpora project we are trying to involve into annotation works native speakers with no special linguistic knowledge. In this paper we describe the way we organize our processes in order to maintain high quality of annotation and report on our preliminary results.

Key words: corpora, annotation, part of speech tagging, ambiguity, crowd-sourcing

1. Introduction

Corpora with manual annotation are required for testing and development of text analysis tools. In the OpenCorpora project we have already created a 1 million of words corpus[1] of Russian texts with human-verified words, sentences and paragraphs boundaries [2]. Morphology is the next level of annotation we are working on. We do this work in two steps: first of all each word gets all possible morphological hypothesis according to dictionary and later all wrong hypothesis are removed by human annotator. Handwork of linguists experts is expensive and we are trying to use native speakers with no linguistic knowledge as much as possible while maintaining high quality of annotation. It has been demonstrated that crowd-sourcing is a suitable method for obtaining linguistic data and "the quality is comparable to controlled laboratory experiments, and in some cases superior" [4]. We have involved several thousands of volunteers into annotation works by providing them with simple annotation questions. In each question we are asking about one grammatical category of one word within a sentence context. We have collected more than 1.1 million of answers[2]. In order to annotate 1 million of words about 4 millions of questions are to be asked.

2. Morphological annotation process

As we have stated before we use morphological dictionary (taken from AOT project [5] with some modifications in the tag set and complex cases' interpretations [1] to find all possible hypothesis for each word. No postprocessing or heuristics is applied to the set of hypothesis so we accept even very rare interpretations such as *ИЗА* (noun, feminine, plural, genitive case, personal name) for the word *ИЗ*. An example of our dictionary-based annotation is shown in Figure 1 (this way to display annotation is described in [1] and [3]).

Fig. 1. Dictionary-based annotation

[1] Statistics on corpus size is always up to date on page http://opencorpora.org/?page=genre_stats

[2] Statistics on contribution to corpus annotation is located on page http://opencorpora.org/?page=stats

The final goal is to have only one interpretation for each word for sentences with no syntactic or semantic ambiguity (as show in Figure 2). For ambiguous sentences several interpretations are allowed.

Мама	мыла	раму
v мама x	v мыть x	v рама x
СУЩ, од, жр, ед, им	ГЛ, несов, перех, жр, ед, прош, изъяв	СУЩ, неод, жр, ед, вн

Fig. 2. Unambiguous annotation

In OpenCorpora project the choice of the right morphological interpretation is done by hand by volunteers. In order to simplify this work we have splitted the annotation of each word into a set of simple annotation questions. Each question is asked about one grammatical category of one single word within a sentence context. In our example “Мама мыла раму” according to the set of hypothesis for the word МЫЛА following questions can be asked:

- is МЫЛА a verb or a noun?
- is МЫЛА singular or plural form of noun?
- is МЫЛА in nominative or accusative case?

This questions form a decision tree (Figure 3) where the next question is asked only in case it is meaningful after the previous answer. For the word МЫЛА in our example the correct answer for the first question is VERB and no other questions will be asked.

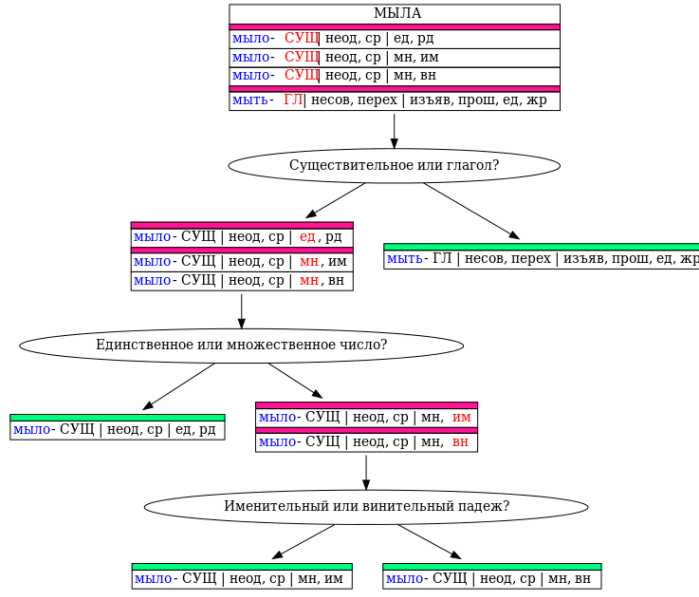


Fig. 3. Annotation decision tree for word МЫЛА

Annotation questions are grouped by type (i.e. “VERB or NOUN”, “singular or plural”, …) and volunteers can choose the specific type of questions they want. Each question type has its own instruction where general guidelines and the explanation of tricky cases are provided. The purpose of guidelines is to refresh background knowledge of linguistic categories and to specify issues which need interpretation different from one given in the secondary or high school. We don’t expect volunteers to know grammar perfectly. Instead we always ask them (both in guidelines and on web pages) to skip questions they don’t understand instead of making doubtful contribution.

3. Annotation quality estimation

Each question is answered by several (three or four) people and then it goes to moderation for approval. Moderators have a good linguistic background and they are able to make a correct decision. It will be very time-consuming to review and approve all answers. At first we have decided to do manual approval only for answers where there is some disagreement between volunteers or comments added. This decision was based on following calculations: let’s assume that all volunteers make random mistakes in 10% of answers (this is high error rate to simple questions like “singular or plural?”). Thus the probability of the event “all three annotators are wrong” is $0.1^3 = 0.001$ i.e. one annotation mistake per 1,000 words (0.1%) will be automatically approved if moderators will review only examples with disagreement.

In practice it turned out differently: an error rate for questions “is noun singular or plural?” is between 0.5% and 10% for most of volunteers and we have found 2% cases where all annotators were wrong. This means that our initial assumption of random error distribution isn’t true and the probability of an annotation error depends on the annotated word itself and on its context.

In order to find features that cause annotation errors we have splitted contexts into a set of simple context features. A context feature consists of position (0 is a position of word being annotated, −1 is one word to the left, +1 — one word to the right) and a word at that position. For each feature we have calculated the number of annotation disagreement events in examples with this feature. Following table includes top features ordered by percentage of disagreement events for questions of type “is noun singular or plural?”. In the rightmost column we show the expected error probability assuming that in case of disagreement between three annotators two of them are wrong (i.e. the worst case).

Table 1. Disagreement statistics for singular vs. plural disambiguation

Context feature	Position	Total samples	Samples with disagreement	Samples without disagreement	Disagreement rate	Expected error probability
word = четыре	−1	64	47	17	73.44%	48.96%
word = две	−1	136	89	47	65.44%	43.63%

Context feature	Position	Total samples	Samples with disagreement	Samples without disagreement	Disagreement rate	Expected error probability
word=три	−1	115	75	40	65.22%	43.48%
word = два	−1	93	60	33	64.52%	43.01%
word = две	−2	58	36	22	62.07%	41.38%
word = одна	4	13	8	5	61.54%	41.03%
word = две	0	226	135	91	59.73%	39.82%
word = копейки	0	17	10	7	58.82%	39.22%
word = четыре	–	95	55	40	57.89%	38.60%

This statistics reflect the norm of Russian grammar stating that the noun after the numeral ending in 1, 2, 3 or 4 must be in the singular. This is counterintuitive and most of people without linguistic knowledge make mistakes.

With these results we have decided to include into manual approval list for moderators all examples with context features provoking errors. The final list of such features will influence the overall precision of the annotation. In order to illustrate this we have plotted all context features occurring in questions of "is noun singular or plural?" type in the 2d space (Figure 4). The estimated error probability is on X-axis and the total number of examples is on Y-axis (logarithmic scale).

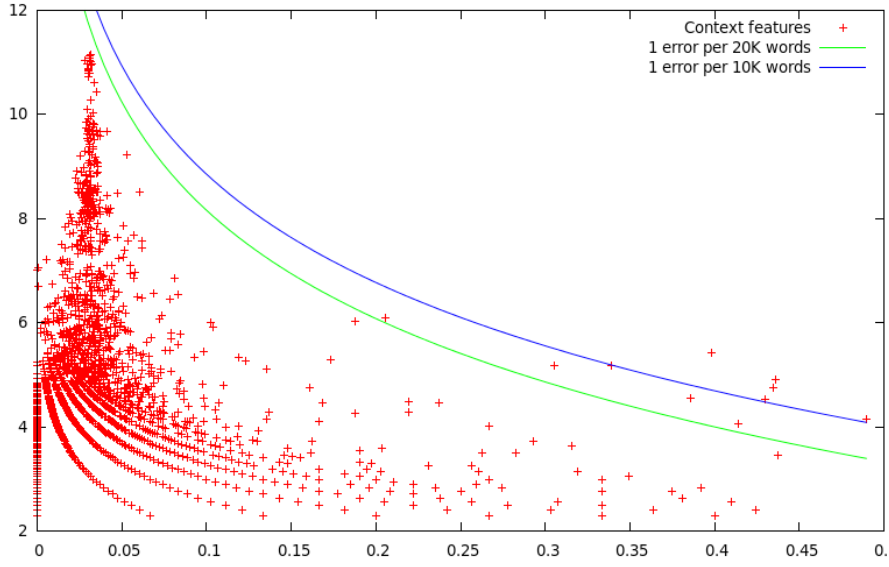


Fig. 4. Context features for "is noun singular or plural?" question type

Each dot on the plot corresponds to one context feature. Lines represent quality goals: the blue one — one error per 10K questions of this type (i.e. words), the green line — one error per 20K of question. Examples with context features above the line are to be included into manual approval list in order to meet quality goal chosen.

The feature with highest frequency is the pseudo-feature that is available in 100% of examples. The quality goal line that intersects with this feature denotes the highest possible annotation precision achievable with partial manual approval process. Better annotation requires all examples to be reviewed by people with expert knowledge in linguistics.

4. Conclusion

In this paper we have described our experience of crowd-sourcing morphological annotation in OpenCorpora project: the way annotation process is organized, our preliminary results and quality estimations technique based on disagreement rate between several annotators.

During the annotation process we collect not only annotation results but also the information about participants' interaction with user interface including timestamps of clicks on buttons. These data allow deeper analysis of both annotation and text understanding process. All the data we have collected are provided in the Download section on http://opencorpora.org and are licensed under the terms of Creative Commons CC-BY.

References

1. *Bocharov V., Bichineva S., Granovsky D., Ostapuk N., Stepanova M.* Quality assurance tools in the OpenCorpora project. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 101–109
2. *Bocharov V., Bichineva S., Granovsky D., Ostapuk N., Stepanova M. Surikov A.* Text segmentation in the OpenCorpora project. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"]
3. *Bocharov V., Granovsky D.* Software tools for collaborative morphological markup [Programmnoje obespechenije dlja kollektivnoj taboty nad morfologicheskoj razmetkoj korpusa]. Trudy mezhdunarodnoj konferentsii "Korpusnaja lingvistika — 2011" [Corpus Linguistics — 2011: Proceedings of the International Conference]. Saint-Petersburg, 2011
4. *Munro R., Bethard S., Kuperman V.* et al. Crowdsourcing and language studies: the new generation of linguistic data // Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. — 2010.
5. *AOT*, available at http://www.aot.ru.

ОПИСАНИЕ РУССКИХ КОНСТРУКЦИЙ С ВНЕШНИМ ПОСЕССОРОМ В СИСТЕМЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Богданов А. В. (bidon@inbox.ru),
Леонтьев А. П. (taonick@yandex.ru)

ABBYY, Москва, Россия

Ключевые слова: внешний посессор, автоматическая обработка, машинный перевод, семантико-синтаксический интерфейс

DESCRIPTION OF THE RUSSIAN EXTERNAL POSSESSOR CONSTRUCTION IN A NATURAL LANGUAGE PROCESSING SYSTEM

Bogdanov A. V. (bidon@inbox.ru),
Leontyev A. P. (taonick@yandex.ru)

ABBYY, Moscow, Russia

The paper shows how Russian external possessor constructions are treated in the ABBYY Compreno ® system. The specific tasks of our system require that sentences with external possessor constructions be considered as synonymous with those with internal possessors. Accordingly, the semantic structure is generated in such a way that the possessor, whether external or not, and the possessum form a single constituent. This is not the case with the syntactic structure because there is much evidence that the external possessor is not syntactically dependent on its possessum. The semantic and syntactic structures of external possessor constructions are not isomorphic so we have to apply a syntax-semantic interface to derive one from the other. We show that two different kinds of interface must be used. For constructions with strong lexical restrictions we use a special normalization module while leaving the syntactic description relatively simple. In contrast, constructions with fewer lexical restrictions require a more sophisticated syntactic description where movements are postulated.

Key words: external possessor, natural language processing, machine translation, syntax-semantics interface

In memory of Aleksandr Yevgenyevich Kibrik

Our research is a part of work on the natural language processing system ABBYY Compreno©. The main task of this system is to convert the input text into a semantic structure that is a tree where nodes are concepts and arcs are relations between these concepts. Further information on the project see in [Anisimovich et al. 2012].

Our system is not designed for processing only one specific language. We claim that the approaches we use in our system are applicable to any natural language. Hence the semantic trees we get must not contain any language specific phenomena and therefore the semantic structure for a certain sentence in Russian must be exactly the same as that of its full analogue in English. That can be illustrated by the example (1a) and its full analogue (1b)

(1a)	Мальчик	съел	хлеб.			*Russian*
	[[$_{\text{Subject}}$ Мальчик]	есть	[$_{\text{Object\_Direct}}$ хлеб]]			
(1b)	The	boy	ate	the	bread.[1]	*English*
	[[$_{\text{Subject}}$ [$_{\text{Article}}$The]	boy]	eat	[$_{\text{Object\_Direct}}$ [$_{\text{Article}}$ the]	bread]]	

The syntactic structures of these sentences which are given below them are although slightly but different. English has articles while Russian does not. That is why in the semantic structure (1c), which corresponds to the both of the sentences, articles are not represented as separate nodes

(1c) [[$_{\text{Agent}}$ BOY[2]] TO_EAT [$_{\text{Object}}$ BREAD]]

Such a guideline may be good for making semantic structures truly universal. Nevertheless it can lead to some problems. Namely when the sentences we are inclined to consider as synonymous have considerably different syntactic structures. This is often the case when some syntactic phenomenon is language specific and cannot be syntactically reduced to a more universal one. Here we have got no other way out but to make the syntactic and semantic structures not isomorphic and, since one cannot be derived from the other by regular procedures, a special syntax-semantics interface will be needed. That is exactly the approach we apply regarding Russian external possessor constructions

---

[1] We use $_{\text{lower index}}$ to mark up semantic or syntactic relations. Square brackets [] mark constituent borders.

[2] Words written in CAPITALS mark semantic concepts. The concepts in our system are mostly named in English. But that is so only for convenience and does not mean that the concepts can not correspond to words in other languages

Problems in description of external possessor constructions

External possessor constructions[3] are generally postulated in sentences (2) to (4)

(2) **У меня** болит **шея**.
on me aches neck
My neck aches.

(3) Мальчик наступил **девочке** **на ногу**.
boy stepped girl.Dat on foot
The boy stepped on the girl's foot.

(4) Мальчик поцеловал **девочку** **в губы**.
boy kissed girl.Acc in lips
The boy kissed the girl's lips.

These sentences share one thing in common, namely the semantic equivalence[4] with sentences like (5) to (7) where the nouns in bold type are replaced with a complex NP headed by one of these nouns (further possessum) and the other (further possessor) is its genitive modifier or possessive adjective.

(5) **Моя** **шея** болит.
my neck aches

(6) Мальчик наступил **на ногу** **девочки**.
boy stepped on foot girl.Gen

(7) Мальчик поцеловал **губы** **девочки**.
boy kissed lips girl.Gen

Thus the semantic structures for (2) to (4) have to be exactly the same as for (5) to (7). But this is not the case for their syntactic structures. The genitive NPs always form a single constituent, but in the external possessor constructions there is some evidence that the possessor is mostly a modifier of a verb and not that of the possessum. In (8a) it is shown that in external possessor constructions possessor and possessum can be on the different sides of the verb. While for the genitive NP such an order is impossible.

---

[3] More on external possessor see: [Kibrik 2003], [Kibrik et al.. 2006], [Payne, Barshi 1999].

[4] There is a lot of evidence that the sentences with the external possessor are not truly semantically equivalent to the sentences with genitive NPs. For example see [Shibatani 1994], [Podlesskaya Rakhilina 1999], [Brykina 2005]. Still in our project we have to neglect it for the reason given in next paragraphs.

(8a) **У Васи** вчера сломалась **машина**.
on Vasya yesterday broke car
Vasya's car broke down yeasteday.
(8b) ?**Васина/Васи** вчера сломалась **машина**.
Vasya.Poss/Vasya.Gen yesterday broke car

In the external possessor constructions the possessum can be a pronoun whereas the corresponding genitives NP is impossible.

(9a) Голова, спрашиваешь? **Она** **у него** круглая.
it on him round
(You asked about the head). He has it round.
(9b) ***Его** **она** круглая.
his it round

Thus we have to admit that in the external possessor constructions the semantic structure is not isomorphic to the syntactic and cannot be derived from it automatically. Further we shall describe the approaches we used to solve this problem.

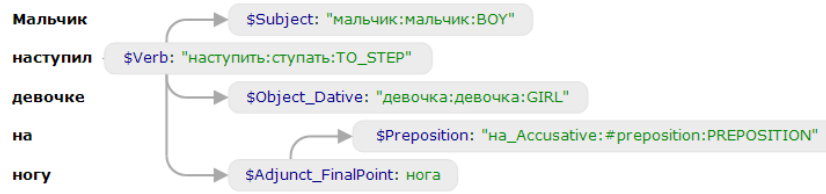
The first approach: simpler syntax, more complicated semantics

The first approach used in our system may be described like this: we no longer assume that the sentences with external possessor are semantically equivalent to the sentences with genitive NPs and hence their semantic structures are not obliged to be the same. In that case nothing prevents us from making the semantics and the syntax isomorphic that is making the external possessor a modifier of verb in the sentences like (3).

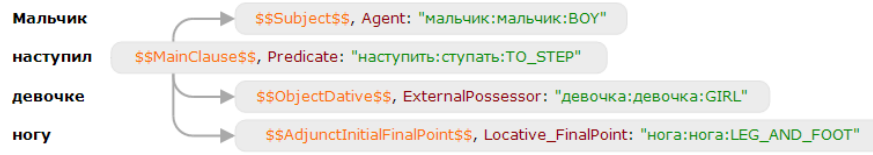
There is only one problem that remains: what semantic role must be assigned to the external possessor. This is a well known issue in theoretical linguistics discussed for example in [Podlesskaya, Rakhilina 1999] and [Shibatani 1994]. It is often proposed that a special extra-thematic role (like Affectee) must be postulated. This is the approach we used for the sentences like (10) where the external possessor is an NP in dative.

(10) Мальчик наступил **девочке** **на ногу**.
boy stepped girl.Dat on foot
The boy stepped on the girl's foot.

The semantic (Picture 1) and the syntactic (Picture 2) structures we proposed are as follows



Picture 1



Picture 2

The NP in dative (*девочке*) is a modifier of verb in the syntactic structure. In the semantic structure it remains the same: the concept correspondent to this NP takes the role ExternalPossessor and is a modifier of the predicate. No complicated interface is needed and the syntax remains quite simple. Which is the most prominent argument for this approach.

Nevertheless we have to face another kind of problem. The extra thematic role we use must have a strange peculiarity: in some languages (e.g. English) this role is never overtly expressed. In that case we will get different semantic structures for the Russian sentence (10) and its English translation. That is not very good for machine translation, which is one of the tasks of our system.

However this problem can be easily solved by a transfer module that transforms the semantic structure in Picture 2 into the suitable structure for the English sentence, and vice versa. A part of this module is presented below (Picture 3).

```
"TO_WALK"
[
 ExternalPossessor: y,
 Locative_FinalPoint: loc "PART_OF_ORGANISM"
]
=>
[
 loc
    [Whole: y]
];
```

Picture 3

Capitalized text in quotation marks («TO_WALK») marks a semantic concept. Sequence of symbols ended with a column (Whole:) marks a semantic role, *loc* and *y* are variables used to mark nodes, and the arrow => divides the input structure from what we get as a result. The sense of the rule is as follows: if a concept of the semantic

class TO_WALK has a dependent with the role ExternalPossessor and another dependent of the semantic class PART_OF_ORGANISM with the role Locative_FinalPoint, the first dependent must be reattached to the second with the role Whole.

This transformation can easily be described due to the special characteristics of the external possessor in dative. This construction is possible only for certain semantic classes of verbs and the semantic roles that are available for the possessor and the possessum have very precise restrictions. That helps us to define the semantic role of the possessor in the output structure. We know that the range of possible concepts in that context is very restricted and so is the range of the possible relations between them.

The range of possible syntactic positions in the input structure is also restricted. Thus we may be sure that our transformation will apply to all the sentences where it must and on the same time will never apply where it must not.

The second approach: simpler semantics, more complicated syntax

The previous approach does not, however, fit with the most frequent construction with an external possessor, namely the construction with the preposition *y* (see (2)). Unlike a dative one, this construction is restricted with neither a class of a verb nor a syntactic position of possessum, nor a possible semantic relation. It makes impossible the application of semantic normalization rules like that in Picture 3.

For description of this construction we had to use a special semantic-syntactic interface. There are several candidates for this role in our system, but the movement was declared as most suitable. It's a mechanism which is typically used in our system for description of complex cases of communicative dislocation and other phenomena associated with violation of projectivity of syntactic structure.

Let's consider the architecture of the system and the mechanism of movement in detail.

In general case the analysis module works as follows: syntactic structure is being built; it means that every word form of the input text takes some syntactic position of some parent. Further the transition from syntactic structure to the semantic one follows as a result of which each arc between a parent and a child is interpreted — each child gets semantic role related to its parent. The process of a switch from syntactic positions to semantic roles is possible because each lexeme has a diathesis description — a list of correspondences between syntactic positions that can connect to it and their semantic roles. Let's consider an example.

(12a) Мальчик дал девочке яблоко.

(12b) [[$_{\text{Subject}}$ мальчик] дать [$_{\text{Object\_Dative}}$ девочка] [$_{\text{Object\_Direct}}$ яблоко]]

(12c) [[$_{\text{Agent}}$ BOY] TO_GIVE [$_{\text{Possessor}}$ GIRL] [$_{\text{Object}}$ APPLE]]

(12d) *The boy gave the girl an apple.*

The input Russian text is represented in (12a). In (12b) one can see the syntactic structure of the input text. All the arcs in syntactic structure are marked with syntactic positions. Semantic structure is represented in (12c) where the arcs are marked with semantic roles and the lexemes are replaced with their semantic concepts. For example, the node GIRL receives its semantic role of Possessor because the lexeme *дать* has among others the diathesis <Object_Dative — Possessor>.

The analysis module is organized such a way that syntactic structure must be projective. Therefore in all cases of real not-projectivity in an input text we have to use a special mechanism — movement. It works as follows: in syntactic structure the word form can connect to its parent into special moved position which does not have any diathesis (does not correspond to any semantic role), but has a rule of movement assigned to it. The rule of movement consists of a path via the syntactic tree from a parent of the moved position to the initial position of movement (pro). The paths can be arbitrarily deep but the moved position (the target of movement) always C-commands the pro. In case of successful search of position for a pro, corresponding to one of the paths, the target of movement takes the position of the pro in semantic structure. At the same time the position of a target erases.

Let's consider the example.

(13a) *Девочке* *мальчик* *хочет*
дать *яблоко.*

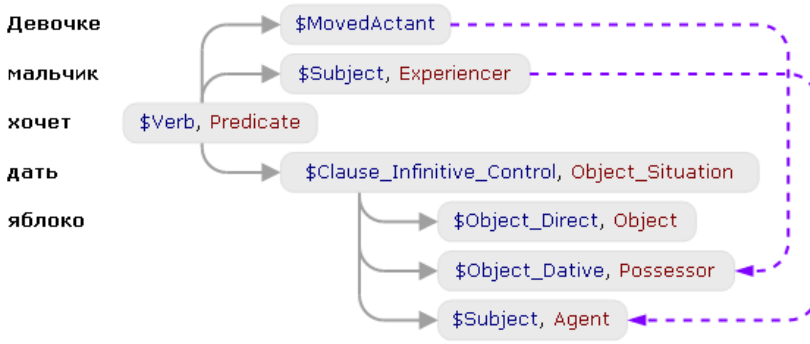
(13b) [[$_{\text{MovedActant}}$девочка] [$_{\text{Subject}}$мальчик] хотеть
[$_{\text{Clause_Infinitive_Control}}$дать [$_{\text{Object_Direct}}$яблоко]]]

(13c) [[$_{\text{Experiemcer}}$ BOY] TO_WANT [$_{\text{Object\_Situation}}$ TO_GIVE
[$_{\text{Object}}$ APPLE] [$_{\text{Possessor}}$ GIRL]]]

(13d) *The boy wants to give the girl an apple.*

An input text with non-projective structure is represented in (13a). Syntactic structure where *девочка* is connected to *хотеть* into moved syntactic position MovedActant is given in (13b). Semantic structure where GIRL is connected to its real parent TO_GIVE and has a semantic role Possessor is shown in (13c).

This structure can be seen at Picture 4.



Picture 4

In Picture 4 one can see schematically shown structure of (13). The syntactic positions are shown with the symbol "$" in blue color. Semantic roles are shown in red color. So-called non-tree links are shown with the dotted arrows, one of which shows movement of *девочке* from the MovedActant position to Object_Dative position under *дать*. The second arrow shows a link of control between a subject of a matrix verb and a subject of infinitive.

For the shown movement to be possible in the movement rule there is a path like this:

Clause_Infinitive_Control.Object_Dative;

In accordance with this path the analysis module restores a pro in Object_Dative position under a constituent in the position Clause_Infinitive_Control which has the same parent as the moved position.

In case of external possessor the position with *y* is the target position of movement, which connects to a verb without diathesis (without semantic role). In movement rule there is a path like this:

(Subject | Actants).Object_Indirect_У;

In accordance with this path the pro restores in the position Object_Indirect_У under subject or some actants like in example below.

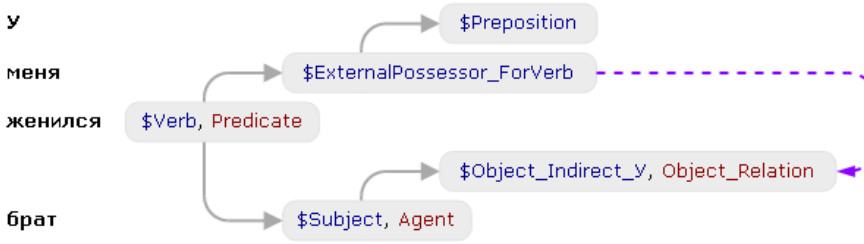
(14a) *У меня* *женился* *брат.*

(14b) [[$_{\text{ExternalPossessor\_ForVerb}}$ я] жениться [$_{\text{Subject}}$ брат]]

(14c) [TO_MARRY [$_{\text{Agent}}$ BROTHER [$_{\text{Object\_Relation}}$ I]]]

(14d) *My brother has married.*

In syntactic structure (14b) external posessor *у меня* is connected to the verb *жениться*, whereas in semantic structure (14c) it is connected to its posesssum BROTHER having semantic role Object_Relation. It can be seen at Picture 5.



Picture 5

Picture 5 shows that *у меня* is connected to the verb in the moved position without semantic role and the movement to the position Object_Indirect_У under the subject can also be seen. Note that the preposition is connected to the noun and does not have a semantic role because it takes the so-called grammatical syntactic position.

Thus in the semantic structure we have an original external possessor under its possessum with necessary semantic role. Using of movement allows us to solve a problem of not-isomorfism of a syntax tree and a semantic tree.

Therefore, although the Russian external possessor constructions have some common characteristics, it may be plausible to apply different approaches while describing them in a system of natural language procession.

References

1. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012) Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, Computational linguistics and intellectual technologies vol. 11 (18).Izdatel'stvo RGGU, Moscow
2. *Brykina M. M.* (2008) The linguistic means of coding of possessivity (basing on a corpus research of Russan) [Yazykovye sposoby kodirovaniya posessivnosti (na materiale korpusnogo issledovaniya russkogo yazyka)], PhD thesis, Moscow.
3. *Kibrik A. E., Brykina M., Leontiev A., Leontieva A. L.* (2007) A single phenomenon-oriented corpus: benefits and problems, Second international conference "Grammar and Corpora". Abstracts, Ustav pro jazyk cesky AV CR, Liblice.
4. *Payne, D. L., Barshi, I.* (eds.) (1999) External Possessor. John Benjamins, Amsterdam/Philadelphia.
5. *Podlesskaya, Vera I., Rakhilina, Ekaterina V.* (1999) External Possession, Reflexivization and Body Parts in Russian, in External Possessor. John Benjamins, Amsterdam/Philadelphia:
6. *Shibatani, M.* (1994) An integrational approach to possessor raising, ethical datives and adversative passives, Proceedings of the 12th Annual meeting of the Berkeley Linguistic Sosciety. Berkeley
7. *Apresyan Yu. D.* (2006) Types of correspondence between semantic and syntactic actants [Tipy sootvetstviya semanticheskikh i sintaksicheskikh aktantov], Problems of typology and general linguistics [Problemy tipologii i obshchey lingvistiki]. St. Petersburg., pp. 15–27
8. *Kibrik A. E.* (2000) External possessor as a result of valence splitting [Vneshniy posessor kak resultat rasshchepleniya valentnosti], A word in a text and dictionary. Papers presented to academician Yu. D. Apresyan at his 70th birthday [Slovo v tekste i slovare. Sbornik k semidesyatiletiyu akademika Yu. D. Apresyana]. Languages of Russian culture, Moscow.
9. *Kibrik A. E., Brykina M. M., Khitrov A. N., Leontyev A. P.* (2006) Russian possessive constructions in the light of corpus statistics research [Russkiye posessivnye konstrukcii v svete korpusno-statisticheskogo issledovaniya]. Problems of linguistics [Voprosy yazykoznaniya] 2006, vol. 1, Nauka, Moscow.

КТО ИЩЕТ — ВСЕГДА ЛИ НАЙДЕТ? (о поисковой функции вербальных хезитативов в русской спонтанной речи)[1]

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

Филологический факультет СПбГУ, Санкт-Петербург, Россия

В докладе анализируются вербальные хезитативы (ВХ) русской спонтанной речи, используемые говорящими в функции поиска. Поиск (конкретной лексической единицы или целой предикативной конструкции) является одной и самых распространенных функций ВХ, он во всех случаях сопряжен с хезитацией, иногда — с самокоррекцией и заканчивается не всегда удачно. Материалом для наблюдений и выводов служит Звуковой корпус русского языка (сбалансированная аннотированная текстотека и блок «Один речевой день»).

Ключевые слова: русская спонтанная речь, паузы хезитации, вербальные хезитативы, хезитационная конструкция, поисковая функция, неудачный поиск, Звуковой корпус русского языка

THOSE WHO SEEK, WILL THEY FIND? (search function of verbal hesitations in Russian spontaneous speech)

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

Saint-Petersburg State University, Philological Faculty, St. Petersburg, Russia

The article is dedicated to verbal hesitations used in Russian spontaneous speech when a speaker is trying to find a better way of expressing his idea. The search process always mates hesitation, sometimes self-correction, and sometimes stays incomplete. Our conclusions are based on the reflections on the Russian Speech Corpus (balanced annotated textothec and the One Speech Day block).

Key words: Russian Spontaneous Speech, Hesitations Pauses, Verbal Hesitations, Discourse elements, Hesitation Construction, Incomplete Search, Search Function, Russian Speech Corpus

[1] Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение 8554 «Звуковой корпус русского языка: комплексный анализ звучащей речи».

Основополагающими тенденциями при порождении устного спонтанного текста являются *принцип экономии и принцип избыточности*, неотделимые от самого понятия естественного языка. Закон экономии работает на всех языковых уровнях (*фонетическом, лексическом* и *синтаксическом*), связан с прогрессивным характером языкового развития в целом и имеет двойственный характер: лингвистическая экономия, являясь, по сути, *деструктивной*, оказывается в то же время и *созидательной*, что особенно хорошо видно на «оси речи» (*Frei* 1929: 108; о редуцированных формах русской речи как источнике пополнения лексикона см., например: *Богданова* 2009).

Однако *на текстовом, дискурсивном, уровне* нашей устной речи обнаруживаются явления диаметрально противоположного толка: абсолютно неэкономное расширение, приращение звуковой формы без всякого приращения смысла — здесь работает принцип избыточности, объясняемый, в частности, тем, что говорящий вынужден в условиях временно́го дефицита решать сразу две задачи — обдумывать речь и собственно говорить. В результате он с неизбежностью вербализует весь процесс порождения текста, включая поиск слова, разного рода хезитации (колебания), самокоррекцию и проч., ср.: «Устная речь необратима — такова ее судьба. Однажды сказанное уже не взять назад, *не приращивая к нему нового* (курсив автора. — *Н. Б-Б.*); “поправить” странным образом значит здесь “прибавить”» (*Барт* 1989: 541).

Одновременное действие обоих законов приводит к тому, что акцентируя наиболее значимые объекты и цели своей речи, говорящий уделяет меньше внимания тем, которые представляются ему менее значимыми. При этом менее значимое может как сокращаться до нуля (*языковая экономия*), так и разрастаться до объемной бесформенной массы (*языковая избыточность*).

Приращение устного текста может осуществляться разными способами (см. об этом, например: *Богданова* 2011), среди которых важное место занимает использование в речи разного типа *неречевых* (*э-э, а-а, гм-м* и т. п.) и *условно-речевых единиц*. К последним традиционно относят частицы (*вот, так, именно, только, это*), союзные средства (*поэтому, потому что*), вводные слова и конструкции (*может быть, наверное, пожалуй*), а также так называемые *вербальные хезитативы: это самое, (я) не знаю, (…) скажем (…), (я) (не) думаю (что), боюсь (что), (…) знаешь (…), собственно (говоря), как его (ее, их), короче (говоря)* и мн. др. Подобные элементы звуковой цепи рассматриваются как разновидность *речевых автоматизмов*, индикатором которых можно считать встречаемость в высказываниях многих говорящих (*Верхолетова* 2010: 6768). Актуальной представляется задача максимально полного, в том числе лексикографического, описания таких единиц на материале естественной русской речи, что и является перспективной целью настоящего исследования.

Из всего многообразия терминов для единиц, способных вербально заполнять паузу хезитации в устной речи (*лишние слова, пустые лексемы, слова-паразиты* и т. п. — см. подробнее, например: *Богданова* 2012), наиболее удачным представляется наименование *«вербальные хезитативы»* (ВХ), поскольку именно вербально выраженная хезитация является универсальной функцией этих единиц в речи. Все они, как правило, полифункциональны, т. е., помимо

хезитационной и параллельно с ней, выполняют в речи еще множество других функций: дискурсивную, поисковую, метакоммуникативную, ритмообразующую, функцию маркера самокоррекции и некот. др. Детальное описание всех функций необходимо для заполнения специальной функциональной зоны в словарной статье *Словаря вербальных хезитативов*, который может оказаться полезным и лингвистам, исследователям повседневной русской речи; и создателям грамматики русской речи; и переводчикам спонтанных текстов на другие языки, хотя бы в рамках художественного произведения, при передаче речи персонажей; и преподавателям русского языка иностранцам, которые вынуждены учиться воспринимать и правильно понимать спонтанную речь как устно, так и письменно, при чтении русскоязычных текстов. Думается, что отдельного исследования требует как каждый из вербальных хезитативов — во всем многообразии выполняемых им функций, так и каждая выделенная функция — во всем многообразии способов ее языкового (речевого) воплощения.

Конкретным объектом внимания в настоящей статье стала *поисковая функция* ВХ, наиболее распространенная в устной речи и реализующаяся по преимуществу вместе с хезитационной. Многообразие способов ее воплощения в спонтанной речи, а также различие результатов производимого говорящим поиска вынуждают взглянуть на эту функцию ВХ более внимательно.

Материалом для анализа стали два блока *Звукового корпуса русского языка*: САТ (сбалансированная аннотированная текстотека) и ОРД («Один речевой день»)[2].

Поисковая функция вербальных хезитативов весьма распространена в устной спонтанной речи и легко поддается дальнейшей (весьма детальной) систематизации, поскольку задержка речи (хезитация) может возникнуть у говорящего в случае поиска разных лексических средств для дальнейшего продолжения монолога, а сам поиск может привести к разным результатам. Наиболее распространены в этой функции ВХ *это, это самое, как его (ее, их)*, хотя встречаются и другие варианты. Анализ материала позволил выявить более или менее полный реестр подобных единиц, а также создать некоторую типологию поисковых ситуаций.

1. Поиск глагола-сказуемого,

который может следовать сразу после ВХ или дистантно, а может не последовать вовсе.

(1) *Танечка ! *П будь ласка / *П* ***это самое*** */* *спроси у Анечки* (ОРД)*;*

(2) *ну вот ну он значит был / грязный / рыжий / как они его в общем* ***это самое*** *нашли / подобрали / так вроде бы* (САТ).

В примере (2) результатом поиска явились сразу два глагола-сказуемых (нашли / подобрали), т. е. колебания говорящего связаны не только с поиском

[2] Подробнее о Звуковом корпусе русского языка см., например: *Богданова* и др. 2011.

слова, но и с элементами саморедактирования. В примере (3) говорящий использовал после ВХ целый ряд глаголов, обозначающих последовательность действий героя:

(3) *в общем / *В *П какой-то маразм / такое впечатление создаётся // *П и старушка эта / *П* ***вот так знаете*** *(э…э) *П взяла / разорвала рецепт / и бросила там ей* (ОРД).

Иногда поиск продолжения начинается прямо с середины составного сказуемого:

(4) *вообще люблю ∫ сходить на рыбалку // ну / на рыбалку // в лес сходить / за грибами // ну вот // а в выходные дни я* (…) *<кашель> люблю* ***это самое*** *отдыхать* (САТ).

Видно, что и в этом, и во многих других примерах пауза хезитации, вызванная поиском слова для продолжения монолога, заполняется говорящим не только конкретным вербальным хезитативом, но и другими средствами *(ну вот / и значит; ∫; в общем, э…э)*, т. е. ВХ встраивается в довольно протяженную *хезитационную конструкцию*.

Иногда поиск сопряжен с обрывом слова, вызванным самокоррекцией:

(5) *она гово… она приходила говорила / люди при(:) (…) люди уже(:) по несколько раз при…* ***это самое*** */ приходят / *П не… не могут договор никак / (э-э) никак получить* (ОРД).

Часто в результате заминки, связанной с поиском глагола и попыткой саморедактирования, выстраиваются неправильные грамматические конструкции (6)–(7):

(6) *я думаю бедных павлинов / тоже во время пло… плохих времен / тоже на* ***это самое*** *∫ пожарили наверное скорее всего* (САТ)*;*

(7) *ой / ха-ха я чего-то / я чего-то / я чего-то запомнила только конец // как они кота накормили / это самое // он начал /* ***это самое*** */ э-э* ***ну это*** */* ***как его*** *// э-э* ***ну*** *з… / ж… /* ***ну*** *жареной свининой //* ***значит*** */ окунями // и он начал кататься валять по полу // кататься и валяться по полу* (САТ).

В примере (7) поиск глагола начинается с середины составного сказуемого. Произнеся вспомогательный фазовый глагол (*начал*), информант хезитирует в поисках субъектного инфинитива, а найдя (и не один, а сразу два), сначала возвращается к первому фрагменту монолога и рассказывает, чем именно накормили кота, а затем повторяет всю конструкцию с составным глагольным сказуемым, начиная с подлежащего.

(8) *мы орем с женой / помедленнее-помедленнее / а он* ***это самое*** *∫ <усмешка> а ему ээ […] Александр Иванович его зовут / а он* ***так это*** */ а чего вы / говорит / боитесь / я тут и так все знаю* (САТ)*;*

(9) *ну ты*(:) (э) ***это самое*** *… свежий завари / вон там в другую @ потому что* (…) *то что в чайнике … *П там что-то осталось / да ?* (ОРД)*;*

(10) *иди* ***это самое*** */ теленяню смотри // @ ты не устал / Хома ?* (ОРД).

Видно, что иногда ВХ и найденный в результате поиска глагол разделяет довольно объемный фрагмент текста (8), а порой всего одно-два слова (9)–(10).

Довольно часто глагол в результате поиска оказывается так и не найден. Факт того, что поиск все же производился, подтверждается не только наличием ВХ, но и всей грамматической структурой фрагмента монолога — например, присутствием в нем подлежащего или второстепенных членов, заполняющих валентности отсутствующего глагола (в примерах все эти элементы подчеркнуты):

(11) **Н / это какой-то ужас // мы () это самое уже / *П вчера / в… они вышли из зала заседания // *П я говорит … *П я говорю / Коля / тебя прессуют* (ОРД)*;*

(12) *как бы он не* ***это самое*** (ОРД)*;*

(13) *а это вы что ли усыновили там кого-то / да ? ага / да да да да // смотрите / молодец какой // а мы /* ***как его*** */ в Доме малютки короче вот / угу* (ОРД).

В примере (12) используется отрицательная хезитационная конструкция *не это самое*, свидетельствующая, что говорящий искал глагол именно с отрицанием. Наличие частицы *не* косвенно также подтверждает, что поиск глагола говорящим действительно производился.

2. Поиск наименования

(чаще — имени существительного), которое, так же как глагол-сказуемое, может следовать сразу после ВХ или дистантно, а может не последовать вовсе. При этом — в случае с ВХ *это самое* и *как его (ее, их)* — наблюдается либо гармония форм конструкции и найденного имени — (14)–(20), либо дисгармония — (21)–(23). Рассмотрим подробнее все эти ситуации.

(14) *ну на фотографиях я посмотрел / он там //–// как бы вылетел он из* ***этой самой*** *∫ из машины и где-то вот ∫ метра–ах наверно ∫ в пяти от машины / лежит тело* (САТ)*;*

(15) *у меня тут @ *Н @* ***это самое*** */ письмо(?) тут ещё есть* (ОРД)*;*

(16) *а / всё …а у меня на да… д… дача на* ***этом*** */* ***как его*** */ на Дунае* (ОРД)*;*

(17) *а* ***ну вот этот вот*** */* ***как его*** *?* ***вот*** *травматолог / это вот как раз таки / и он начальник третьего отделения травматологии* (ОРД)*;*

(18) *там одна женщина раб… живёт значит / прилично с… прилично построено у неё всё / @ угу // @ и ей там делали /* ***ну знаете вот*** *кирпичный* (э-э) *в… забор / там всё* (ОРД).

Гармония форм ВХ и найденного имени (словосочетания) видна во всех приведенных примерах, равно как и разнообразие самих ВХ. В текстах (14), (16) видно также, что предлог в рамках предложно-падежной формы повторяется дважды — до ВХ и после, уже рядом с найденным существительным. Создается впечатление, что грамматика планируемой конструкции для говорящего ясна изначально и он вынужден искать только ее лексическое наполнение (ср. выше о поиске глагола с отрицанием или именной части составного сказуемого).

Из следующих примеров видно, что грамматическая гармония форм не разрушается даже в случае довольно большой дистанции между ВХ и искомым наименованием:

(19) *тут ещё есть / и… и как были вот* ***это самое*** */ я… так и говорится / да /* ***вот эти*** */ *П как(?) (…) называется ? *П все документы* (ОРД)*;*

(20) *ну она не бу… господи / Дора понятно что люди платить не будут / я тебе говорю это те люди / которые берут* (…) */* ***этот самый*** */ правильно как ты сказала / на пятерых / один шведский стол* (ОРД).

Дисгармония грамматических форм хезитатива и найденного в результате поиска имени, так же как и гармония в предыдущем блоке примеров, хорошо видна даже неспециалисту:

(21) *а! он еще тоже […] судьба какая! он же в эту ∫ в первую чеченскую кампанию / он еще* ***этим самым*** *∫ <вздох> в солдаты попал // на срочную службу* (САТ)*;*

(22) *а потом нас еще и повезли / как бы на* ***эти самые*** *∫ на день рождения* (САТ)*;*

(23) *ну вот я Данилевского что-то посмотрел уже / когда-то // в прошлом году / наверное / она кстати есть в Интернете если ты тебе интересно эта монархическая* ***эта как ее*** */ государство у нас есть в Интернете* (ОРД).

Сам факт использования в роли заполнителя паузы хезитации не наиболее типичной «классической» формы ВХ *это самое* свидетельствует о том, что поиск говорящим действительно производился, просто результатом его стала не только

конкретная форма имени, но и некоторая общая коррекция произносимого текста. Несколько иначе обстоит дело в тех случаях, когда говорящий использует «классическую» форму данного ВХ — *это самое*. Отсутствие грамматической гармонии после хезитации (такой, как, например, в тексте (15) — ***это самое*** / письмо), может свидетельствовать в пользу другой интерпретации данных примеров: в них налицо не поиск имени, а просто заминка речи общего характера, без явной мотивации:

(24) *ну что говорю воровали / он говорит / несли все подряд / они говорят / даже ∫* ***это самое*** *∫ даже паркет отрывали* (САТ)*;*

(25) *это… это @ не пирожное @ это не пирожное / это…* ***это самое****… @ мираλгин* (ОРД).

Приведенные примеры позволяют сделать еще ряд интересных наблюдений. Так, в тексте (24) видим повтор усилительной частицы *даже*: она употреблена сначала до ВХ и двух физических пауз хезитации, свидетельствующих в целом о длительном колебании говорящего, а затем еще и после ВХ, перед найденным существительным. По-видимому, есть разница между автоматическим *воспроизводством* в спонтанной речи различных формальных единиц текста (предлоги, частицы) и порой трудоемким, сопряженным с поиском, *производством* — значимых.

Иногда имя в результате поиска так и не найдено:

(26) *страшно боже! я это вцепился там всеми силами ээ души ∫ в* ***эти самые*** *∫ в эту бедную […] и она скрипит еще и думаешь / сейчас она отвалится* (САТ)*;*

(27) *Настя / достань* ***это самое*** *и @ *Н @ выкини / давай* (ОРД)*;*

(28) *правда удобные* ***эти самые*** *?* (ОРД)[3].

Видно, что и в этом случае формы ВХ весьма разнообразны, что лишний раз доказывает, что поиск существительного говорящим действительно производился. В ряде примеров есть даже определения к этому так и не найденному имени: *в* ***эти самые*** *∫ в эту бедную*; *удобные* ***эти самые***.

Анализ этой части материала позволяет сделать еще ряд любопытных наблюдений. Так, довольно часто, не найдя подходящего имени, говорящий вставляет вместо него своеобразный описательный оборот (в примерах подчеркнут), помогая слушателю понять невысказанную или плохо высказанную мысль:

(29) *ну там не пятка вернее / а вот* ***это самое*** */ перед пяткой /* ***как её*** *? на стопе* (ОРД)*;*

---

[3] Примеры такого рода, весьма, надо сказать, немногочисленные в материале исследования, могут быть, тем не менее, выделены в особый разряд, так как представляют скорее не монологическую, а диалогическую речь, в которой отсутствие имени может быть восполнено, например, жестом или общей ситуацией разговора.

(30) *а я… не от меня зависит / я бы подсадила на *Н # ну дак подсадила людей на* ***это самое*** */ ты знаешь / да ? это зависимость появилась* (ОРД).

В примере (30) вместо описательного оборота говорящий использует прием обращения к слушающему, предполагая, что тот знает, о чем идет речь.

В ряде контекстов конструкция ВХ *это самое* входит в состав сравнительного оборота — и в этом случае нахождение имени почти и не предполагается:

(31) *и вот // Лукашенко сидел там как* ***этот самый*** *∫ значит ∫ и все вывешивали там флагами этими* (САТ)*;*

(32) *скажите / что такое йока? а это говорит / наше такое блюдо в общем / очень интересное / что за йока ? что за блю… блюдо такое йока ? йока это когда / значит / как бы яйцо / яйцо вместе с блином пожаренное /* ***вот такое вот*** */ типа* ***это самое*** *∫ типа такого чего-то […] потому что он необычный / это блин! но блин с яйцом* (САТ).

Фактически здесь мы имеем дело с неким устойчивым оборотом, свойственным только разговорной речи и выражающим в разных текстах разное, зачастую трудноопределимое, значение. В примере (32) ВХ входит к довольно протяженную «поисковую конструкцию», также имеющую характер сравнительного оборота, но предполагающую все же (не всегда успешное) нахождение нужного слова.

Из приведенных примеров видно также, что конструкции с ВХ без искомого имени существительного носят подчеркнуто разговорный характер.

3. Поиск определения

Такие примеры в материале исследования оказались немногочисленными, но и их удалось некоторым образом систематизировать. Так, в примерах (33)–(36) определение находится довольно быстро и следует сразу после ВХ:

(33) *ну просто здесь случай / *П* ***скажем так*** *не самый простой* (ОРД)*;*

(34) *я был один раз на () маслобойке вот на Украине // да ? в Донецке () куда я ездил // мы возили семечки в обмен на масло* ***такое знаешь*** *() ну ароматное* (ОРД)*;*

(35) *я уже думал / может опять отправить их одних // *П всё-таки с ребёнком / там море* ***такое знаешь*** *(…) опасное* (ОРД)*;*

(36) *и дальше поехали на один из курортов на / берегу / Каспийского / моря // ну / э-э конечно там уникальная / природа /* ***вот этот*** */ э-э совершенно нетронутая / Каспийское / море* (САТ).

В примере (36) снова видим дисгармонию форм ВХ и найденного определения (равно как и определяемого): *природа / **вот этот** / э-э совершенно нетронутая*.

(37) *церковь когда ехали мм в аэроду… аэропорт где-то небольшая **такая** церковь **вот** прямо **такая вот** ∫ с луковками сс звездочками* (САТ).

В отличие от предыдущих примеров, определение в тексте (37) найдено несогласованное, похожее на описательный оборот, характерный для ситуации неудачного поиска, ср.:

(38) *нет / а я думаю знаешь что ? // вот этот вот дом / он … вот это всё / можно тоже **таким** // **это как его** / ну как брёв… брё… брёвна / как () что ? доски // а вот эту сторону … она будет немножко что / потемнее / в тени потому что* (ОРД).

В данном примере (38) говорящий пытается назвать цвет, которым следует покрасить дом, и, не найдя нужного слова, снова дает сравнительный оборот (в тексте подчеркнут).

4. Поиск обстоятельства

(39) *она говорит да ладно *Н и вот знаешь говорю / у меня уже собеседование говорю / **знаешь как** *Н критически так смотрю @ угу* (ОРД).

Подобный контекст в материале исследования оказался пока только один, хотя в нем представлен «поисковый» ВХ *знаешь как*, за которым сразу следует найденное обстоятельство.

5. Поиск предиката (не действия),

который может либо следовать после ВХ (сразу или дистантно), либо не последовать вовсе.

(40) *и соответственно значит вся Олимпиада наша […] без электричества как они будут? естественно это очень это самое ∫ сложно будет* (САТ)*;*

(41) *снимочки мы делали / там вроде косточки нормальные / да ? у вас ? // да / там* (:) *они **это** / **как его** ? ближе к спрессованным* (ОРД)*;*

(42) *у нас сейчас () акция действует / скидка десять процентов на библиотеку // то есть это у вас получается **как бы** () **так скажем** бесплатная сборка* (ОРД).

В примере (40) говорящий осуществляет поиск всего предиката (сложно будет), в то время как в примере (42) поиск начинается только перед именной

частью сказуемого, что снова свидетельствует о разных механизмах *воспроизводства* в живой речи грамматической структуры синтаксических единиц и *производства* — их лексического наполнения.

(43) *и вот мы с ним этим самым конечно с женой мы все в шоке / по этому поводу / потому что жена-то на* ***это самое*** *∫ Татьяна-то на девятом месяце беременности* (САТ)*;*

(44) *Настя / *П* (э-э) *коша // а всё / всё* ***это самое*** */ всё анонимно* (ОРД).

Дистанция между ВХ и предикатом в обоих примерах заполнена повторением (в первом случае — с поправкой-заменой) подлежащего *(Татьяна-то, всё)*. В роли именной части сказуемого выступают краткое прилагательное (анонимно) или развернутая именная конструкция (на девятом месяце беременности). В контексте (43), в отличие от (44), нет формальной гармонии между ВХ и найденным словом.

6. Поиск предикативной единицы (подчеркнута)

Фактически здесь происходит поиск не одной лексической единицы, а целого предикативного фрагмента, связанного, тем не менее, с конкретным словом в предыдущей части монолога:

(45) *от есть / вещи такие / вот / ну / у людей хобби например / да ? *П *В ну(:) /* ***там скажем*** */ *П* ***ну*** */* ***не знаю*** */ паяет что-то* (ОРД)*;*

(46) *он ммможет умереть / вообще невозможно / потому что он так любил жизнь / у него же* ***вот это вот*** *он постоянно всё / пытался везде успеть // я помню когда мы еще только познакомились* (САТ)*;*

(47) *тоже цветные / да ? // цветные / да / да // цветные матовые // цветные матовые / хорошо // надеюсь* ***этот самый*** */* ***как его*** */ шестнадцать штук три на четыре еще надо* (ОРД)*;*

(48) *нет / не было так мне… так сильно / ярко выражено // у меня* (:) (э-э) */* ***как его*** */ угу // когда как там было* (ОРД)*;*

(49) *так а у меня чуть чего-то тоже* (э-э) ***как его*** *? после этого и панкреатит может развиться* (ОРД)*;*

(50) *у него же после менингита / у него*(:) ***это самое*** */ *П он медленно говорит / он* (ээ) */ долго ему надо ду… обдумывать все эти … # длинная реакция видимо* (ОРД)*;*

(51) *т… ты сама … # а сейчас с флэшкой всё таки / ага / сейчас // он ушёл в* ***этот самый*** *@ он ушёл @ в (э-э) … бельё понёс / я ему сказала / он сейчас бельё отнесёт и придёт* (ОРД)*;*

(52) *извини меня / хахаль из двухкомнатной хаты / угу //* ***как его*** *? там евроремонтик у него / пластиковые окна / кухонька новая* (ОРД)*;*

(53) *ой-ой-ой / *П* ***ну*** *(:)* ***знаете*** */* ***как это*** */ вашими устами / да мед пить / *П вот* (ОРД).

Из примеров видно, что найденный в ходе хезитативного поиска предикативный фрагмент соотносится с предыдущей частью контекста как с помощью союзных средств *(когда, после этого)*, так и бессоюзно. В целом эта ситуация отражает саму сущность спонтанного речепроизводства, когда любое продвижение монолога вперед может вызвать хезитационную заминку, в ходе которой говорящий обдумывает свою речь, подбирает ту или иную конкретную языковую единицу, что-то корректирует или меняет микротему. Поисковая функция вербальных хезитативов оказывается практически неотделимой от собственно речевого колебания, хотя попутно решается еще и ряд других задач.

Думается, что вся информация такого рода должна найти свое место в словарной статье на конкретную единицу в Словаре вербальных хезитативов, что может придать ему, как мне уже приходилось отмечать, весьма специфическую форму: это может стать своего рода собранием исследовательских эссе по каждому ВХ. Более того, такая статья должна оставаться открытой для включения в нее новых данных или для коррекции имеющихся, поскольку Звуковой корпус русского языка, ставший основным источником материала для всех приведенных здесь наблюдений, продолжает пополняться, равно как продолжается и работа по его описанию.

Литература

1. *Барт Р.* Гул языка // Р. Барт. Избранные работы. Семиотика. Поэтика / Сост., общ. ред. и вступ. статья Г. К. Косикова. М., 1989. С. 541–544.
2. *Богданова Н. В.* Об одном из путей пополнения русского лексикона (к формированию учебных материалов нового типа) // Речевая коммуникация в современной России: Материалы I международной научной конференции (Омск, 2729 апреля 2009 г.) / Под ред. О. С. Иссерс, Н. А. Кузьминой. Омск, 2009. С. 36–42.
3. *Богданова Н. В.* Действительно ли наша устная речь экономна в средствах? // Язык и речевая деятельность. 2010–2011. Том 10–11. В честь Н. Д. Светозаровой. СПб., 2011. С. 33–44.

4. *Богданова Н. В.* О проекте словаря дискурсивных единиц русской речи (на корпусном материале) // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Выпуск 11 (18) / Гл. ред. А. Е. Кибрик. М., 2012. С. 71–83.
5. *Богданова Н. В., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус русского языка: новый подход к исследованию речи // Труды международной конференции «Корпусная лингвистика-2011». 27–29 июня 2011 г., Санкт-Петербург. СПб., 2011. С. 98–103.
6. *Верхолетова Е. Ю.* Структурно-динамический подход к социальной стратификации устной речи. Дис. … канд. филол. наук. Пермь, 2010. 319 с. (машинопись).
7. *Frei H.* La Grammaire des Fautes. Paris, 1929. 317 p.

References

1. *Bart, R.* (1989) Drone of the Language. [Gul jazyka] R. Bart. Izbrannye raboty. Semiotika, Poetika (Selecta. Semiotics. Poetics). Moskva: 541–544.
2. *Bogdanova, N. V.* (2009) About One of the Ways of Russian Vocabulary Update (Creating New Learning Materials) [Ob odnom iz putej popolnenija russkogo leksikona (k formiorovaniju uchebnyx materialov novogo tipa)]. Rechevaja kommunikacyja v sovremennoj Rossii. Materialy I Mezhdunarodnoj nauchnoj konferencii (Speech Communication in Modern Russia: Material of the First International Scientific Conference). Omsk: 36–42.
3. *Bogdanova, N. V.* (2011) Is Our speech Really Prudent in Use of Resources? [Dejstvitel'no li nasha ustnaja rech' ekonomna v sredstvax?]. Jazyk i rechevaja dejatel'nost'. Tom 10–11. V chest' N. D. Svetozarovoj (The Language and Speech Activity. Vol. 10–11. In Honour of N. D. Svetozarova). Saint-Petersburg: 33–44.
4. *Bogdanova, N. V.* (2012) Dictionary of Discours Elements of Russian Speech; Project Description (based on Corpora Material) [O proekte slovar'a diskursivnyx edinic russkoj rechi (na korpusnom materiale)]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii «Dialog 2012» (11/18) (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference «Dialog 2012»). Moskva: 71–83.
5. *Bogdanova, N. V., Stepanova, S. B. & Sherstinova, T. Iu.* (2011). The Russian Speech Corpus: the New Approach to Speech Research [Zvukovoj Korpus Russkogo Iazyka: Novyj Podkhod k Issledovaniiu Rechi]. Trudy Mezhdunarodnoj Konferentsii «Korpusnaia Lingvistika-2011» (Proceedings of the International Conference «Corpora Linguistics-2011». Saint-Petersburg: 98–103.
6. *Verxoletova, E. Ju.* (2010) Dynamic Structure Approach to the Social Stratification of Speech. Candidate's Thesis [Strukturno-dinamicheskij podxod k socyal'noj stratifikacii ustnoj rechi. Kandidatskaja dissertacyja]. Perm': 319 p. (typing).
7. *Frei, H.* (1929) La Grammaire des Fautes. Paris : 317 p.

КОМПЬЮТЕРНЫЙ СЛОВАРЬ РУССКИХ ПАРОНИМОВ, ОСНОВАННЫЙ НА ФОРМАЛЬНОМ КРИТЕРИИ ПАРОНИМИИ

Большакова Е. И. (eibolshakova@gmail.com)
МГУ им. М. В. Ломоносова, Москва, Россия;
НИУ Высшая школа экономики, Москва, Россия

Большаков И. А. (iabolshakov@gmail.com)
Независимый исследователь, Москва, Россия

В результате исследования наиболее крупного печатного словаря паронимов русского языка предложен формальный критерий паронимии. Паронимами считаются те пары слов одного корня и одной части речи, у которых различия в аффиксах (раздельно в префиксах и суффиксах) находятся в фиксированных рамках. Согласно этому критерию построен компьютерный словарь русских паронимов, имеющий 21,8 тыс. статей с 192 тыс. паронимов и по объему превышающий все известные словари. В первую очередь словарь предназначен для исправления в текстах ошибочных замен слов их паронимами.

Ключевые слова: паронимы, паронимия, компьютерный словарь паронимов, паронимические ошибки, исправление ошибок в текстах

A COMPUTER DICTIONARY OF RUSSIAN PARONYMS BASED ON A FORMAL CRITERION OF PARONYMY

Bolshakova E. I. (eibolshakova@gmail.com),
Lomonosov Moscow State University, Moscow, Russia;
National Research University Higher School of Economics, Moscow, Russia

Bolshakov I. A. (iabolshakov@gmail.com),
Independent researcher, Moscow, Russia

We note that Western European lexicography has neither precise definition of paronymy nor dictionaries of paronyms. However, such dictionaries can help us correct malapropisms like *massive evacuation* or sensitive shoes. Although three comprehensive dictionaries of Russian paronyms have been published in the recent decades, it remains unclear what additional features of similarity of two words of the same root and the same POS are needed to consider the words paronymous. Based on the collected statistics of affix proximity of paronyms in the largest printed dictionary of Russian paronyms, we propose a formal criterion of paronymy. Two words of the same root and the same POS are considered formally paronymous if their affix differences (separately for suffices and prefixes) are limited to particular values. Affix difference equals the minimal number of editing operations on affixes (deletion, insertion or substitution) that transform an affix chain of one word into that of the other. Aiming to develop a computer dictionary of formal paronyms, we first compiled a computer dictionary of 23,000 Russian words divided into 2,400 same-root, same-POS groups. All words were split into morphs: prefixes, the root, suffixes, and the ending. Then affix distances between word pairs from the groups were automatically computed, and all formally paronymous pairs were selected. These pairs constitute the resulting computer dictionary of paronyms, which contains 21,800 word entries with their 190,000 paronyms, larger than all known dictionaries of paronyms.

Key words: paronyms, paronymy, computer dictionary of paronyms, paronymy errors, correction of malapropisms

1. Введение

Внешнее сходство слов является источником разнообразных ошибок, встречающихся в текстах и подлежащих исправлению. В лингвистике с внешним сходством слов связаны такие понятия, как паронимы и паронимия.

В английском языке слово *paronymous* известно с середины 17 века. Но если сейчас извлечь из интернетовских сайтов десяток англо- и франкоязычных определений паронимии, то единства мнения не обнаружится. В большинстве определений указывается совпадение корня (*wise — wisdom*), в других фигурирует совпадение звучания при различии смысла и орфографии (*hare — hair*). Упоминаются также никак не уточняемые совпадение деривации, единство происхождения, различие окончаний. В словаре [7] слово *paronymous* имеет два разных смысла, и лишь один из них имеет отношение к сходству слов. Найденные определения не указывают явно принадлежность паронимов к одной части речи.

При отсутствии единства в понимании паронимии становится понятным отсутствие в западноевропейской лексикографии словарей паронимов, содержащих описание различий их значений с приведением диагностирующих контекстов (синтаксически связанных слов). Некоторые сведения о паронимах содержатся лишь в словарях и пособиях по общему словоупотреблению, например, в [5] — для английского языка.

В то же время русская лексикография за последние десятилетия дала три содержательных словаря русских паронимов [1, 6, 8]. В них много

диагностирующих контекстов, а словарь [1] указывает смысловые различия паронимов особенно детально. В предисловиях содержатся некоторые уточнения понятия паронимии. Единым является требование **одинаковости корней и частей речи.**

Однако все три словаря русских паронимов не дают строгого определения паронимии, необходимого для построения компьютерного словаря паронимов. Неясно, какие именно дополнительные черты сходства слов необходимы, чтобы они считались паронимичными. Так, [8] требует от паронимов одинаковое место ударения (например, *сытый — сытный*) и, тем самым, одинаковое число слогов, а в [1] предлагается, например, паронимическая пара *показ — показание* с весьма разным числом слогов. Остается неясным и соотношение между паронимией и семантикой сопоставляемых корней и слов в целом. Например, в [6] не признаются полноправными паронимами синонимы типа *паронимический — паронимичный* и слова с омонимичными корнями типа *платный — платяной*.

В настоящей работе понятие паронимии уточняется в связи с определенным приложением словарей паронимов. Предлагается формальный, т. е. пригодный для проверки компьютером, критерий паронимии слов русского языка, и описывается построение компьютерного словаря паронимов, отвечающих этому критерию. Исследовав наиболее полный словарь русских паронимов, созданного В. Красных [6] (далее — **К-словарь**), мы нашли ту меру сходства слов в аффиксах (раздельно в префиксах и суффиксах), которая более чем в 99 % гарантирует паронимию в ее интуитивном лингвистическом понимании. Паронимами считаются те пары слов одного корня и одной части речи, у которых различия в аффиксах находятся в фиксированных рамках.

Компьютерный словарь русских паронимов построен на базе созданного нами ранее компьютерного словаря однокоренных слов (далее — **ОКЧ-словарь**), при этом для автоматического обнаружения паронимических пар применен предложенный формальный критерий. Построенный словарь (далее — **П-словарь**) превысил по объему все известные словари русских паронимов.

2. Необходимые соглашения и уточнения

Важным приложением словарей паронимов является подбор кандидатов на исправление тех ошибок, когда в тексте одно слово заменяется на другое существующее слово, на него похожее. Такие ошибки называются малапропизмами [3, 4]. Среди малапропизмов мы рассматриваем паронимические ошибки, т. е. неправомерные замены слов на слова с тем же корнем и той же части речи, как, например, при использовании словосочетания *массивный отъезд* вместо *массовый отъезд*.

Для указанного приложения важно, чтобы паронимы отвечали **принципу морфологической инвариантности контекста**: замена в тексте одного паронима другим без внесения каких-либо иных правок не нарушает морфологическую правильность текста, хотя может изменить его смысл.

К примеру, замена *отъезд* → *поезд* в сочетании *массовый отъезд* сохраняет **морфологическую** правильность в любом контексте. В то же время замена *отъезд* → *поездка* потребовала бы пересогласования прилагательного *массовый* по роду. Поэтому сформулированный выше принцип не допускает паронимию слов *отъезд* и *поездка*.

Принцип инвариантности облегчает исправление паронимических ошибок. Так, если в тексте встретилось словосочетание *экономическая эффектность,* и каким-то способом выявлена его ошибочность, то в построенном с учетом указанного принципа словаре паронимов будут сразу найдены всевозможные замены ошибочного слова и среди них *эффектность* → *эффективность* (но не *эффект*, так как это слово другого рода). Найденной заменой текст исправляется без какого-либо дополнительного его редактирования (см. подробнее в [3, 4]).

Для соблюдения принципа инвариантности нами были приняты следующие соглашения, уточняющие понятие части речи.

Расщепление существительных по числу. Формы единственного и множественного числа одного существительного будем считать разными существительными. Возникшие пары оказываются однокоренными и попадают в ОКЧ-словаре в одну группу. Группа обычно включает четыре подгруппы: муж. рода ед. числа, жен. рода ед. числа, сред. рода ед. числа и множ. числа (для множ. числа род считается нерелевантным). Согласно принципу инвариантности, представители разных подгрупп паронимами быть не могут.

Отделение причастий от глаголов. Причастия играют в текстах синтаксическую роль прилагательных, и у них та же морфопарадигма. Поэтому мы включаем далее причастия обоих видов и залогов в прилагательные. Однословные степени сравнения прилагательных считаются отдельными прилагательными. Все это позволяют искать паронимические пары в таких группах слов, как {***стар**ый, **стар**еющий, **стар**ейший, у**стар**евающий*...}.

Отделение деепричастий от глаголов. Русские деепричастия образуют при глаголах примерно те же зависимые обстоятельственные группы, что и наречия (*уйти* → *торопясь / торопливо*). Поэтому мы включаем все деепричастия в наречия.

Разделение глаголов и причастий по возвратности. Стоящая за окончанием глагола или причастия (у нас — прилагательного) возвратная частица *ся/сь* существенно меняет модель управления слова, тем самым делая его непохожим на все иные слова той же части речи без частицы. Поэтому наличие / отсутствие этой частицы делит глаголы и причастия на две несопоставляемые группы.

Расщепление глаголов по виду. Два вида русского глагола могут быть существенно разными морфологически. К тому же у них есть различия в комбинаторике, например, можно *делать прыжки*, но нельзя *сделать прыжки*. Поэтому мы считаем совершенный и несовершенный виды одного глагола различными глаголами.

Следующие решения, принятые нами ещё при создании ОКЧ-словаря, связаны с пониманием одинаковости корня.

Учет алломорфизма корня. Корень слов одной группы ОКЧ-словаря может иметь несколько алломорфов (***дух*** *—* ***душ****а;* ***лиц****о —* ***личина;*** ***отч****ество–* ***отеч****ество*). Лишь тогда, когда алломорфы корня оказывались слишком далекими по буквенному составу, мы формировали разные группы. Например, алломорфы ***лож/лаг*** Vs. ***клад/клас*** формируют группы {*по****лож****ить, на****лож****ить, по****лаг****ать…*} Vs. {***клас****ть, вы****клад****ывать, на****клад****ывать…*}.

Включение омонимичных корней. В одну группу мы включали слова с омонимичными корнями, например: {***бур****ый,* ***бур****ный,* ***бур****овой*}. Объединяли и однокоренные слова, омонимичные корни которых имеют хотя бы один одинаковый алломорф {***душ****а,* ***дух****ота +* ***душ***}, {*за****плак****анный,* ***плач****ущий +* ***плат****ный, у****плач****енный + платяной,* ***полот****няный*}. Лишь тогда, когда объединенная группа оказывалась слишком обширной, мы разбивали ее на 2–3 группы с неперекрывающимися алломорфами корня.

Включение заимствованных слов. У заимствованных слов вычленялись иноязычные аффиксы (***ин-****/****де-****/****ре-****/****про****-дукционный;* ***ак****-кредитация*), учитывались алломорфы заимствованных корней (***дубл****-ет —* ***дупл****-ет*). Изредка русский и заимствованный корень совпадают по смыслу, и тогда в одну группу попадают, например, *ин-****нов****-ация* и *об-****нов****-ление*, вместе с их аффиксами, русскими и заимствованными.

Исключение многокоренных слов, слов с префиксоидами и суффиксоидами. Мы не рассматриваем слова с префиксоидами типа *много, едино, мульти…* и суффиксоидами типа *летн, этажн…*, а также большинство многокоренных слов. Это значит, что не считаются паронимами слова *этажный* и *многоэтажный*, *летний* и *многолетний*, *законный* и *закономерный*, а также такие слова из К-словаря, как *зловредный*, *злокачественный*, *злонамеренный*…. Однако оставлены слова, имеющие два одинаковых склеенных корня, но разные аффиксы, например: *добровольный* и *добровольческий*.

Кроме рассмотренных уточнений наше формальное определение паронимии слов учитывает их сходство в аффиксах, причем отдельно — в префиксах и в суффиксах. Аффиксное сходство двух слов будем оценивать парой целых чисел (*Np*, *Ns*). Здесь *Np* — число различающихся префиксов, т. е. минимальное количество элементарных операций редактирования цепочки префиксов (их удаление, вставка или замена), переводящих цепочку префиксов одного слова в цепочку префиксов другого слова. Аналогично определяется число *Ns* для суффиксов. Отметим, что при аффиксном сравнении слов мы считаем нерелевантными их окончания, поскольку они определяются последним словообразовательным суффиксом или корнем слова.

Ограничения на значения (*Np*, *Ns*) для паронимов установлены нами в результате статистического обследования К-словаря, в ходе которого использовались данные о морфемном разборе слов ОКЧ-словаря.

3. Морфемный разбор и морфологический анализ слов ОКЧ-словаря

ОКЧ-словарь состоит из групп слов, имеющих одинаковый корень и относящихся к одной части речи (при уточненном их понимании, описанном выше).

Поскольку омонимы имеют одинаковую морфемную структуру, омонимия слов нами не учитывается, как если бы один омоним заменяет все остальные. С учетом этого упрощения ОКЧ-словарь имеет следующий состав (январь 2013 г.):

Количество слов	**23 054**
среди них, в процентах:	
существительных	42,2
глаголов	21,9
прилагательных	33,7
наречий	2,2
Число групп	2 426
Средний объем группы	9,5
Число однокоренных пар	301 074
Число сравниваемых пар	165 719

Слова ОКЧ-словаря были подвергнуты морфемному разбору вручную (за неимением программы автоматического выделения морфов), т. е. расчленены на префиксы, корень, суффиксы и окончание. Выделять суффиксы было особенно сложно. В частности, было не ясно, как задавать окончания в инфинитивах; присоединять ли так называемые тематические гласные *а, и, е* к суффиксам причастий *ющ* и *вш*. Было понятно, что склеивание некоторых морфов, различаемых лингвистами, отнесет к паронимам множество не слишком похожих слов, а расщепление морфов сильно отдалит в пространстве аффиксов даже похожие слова. Мы не учитывали аффиксный алломорфизм, и в ряде случаев склеивали смежные суффиксы.

В число префиксов включено слитное отрицание *не*, которое наряду с префиксами *а, анти, контра, против* формирует антонимы сравниваемых слов.

Ниже представлены примеры групп ОКЧ-словаря для существительных, глаголов и прилагательных после морфемного разбора (префиксам предшествует знак «–», корню «+», суффиксу «–», окончанию «*», перед частицей *сь/ся* также ставится «–»):

```
-АК+КРЕДИТ-АЦИ*Я
-АК+КРЕДИТ-ИВ*
-АК+КРЕДИТ-ИВ*Ы
+КРЕДИТ*
+КРЕДИТ-К*А
+КРЕДИТ-К*И
+КРЕДИТ-ОВ-АНИ*Е
+КРЕДИТ-ОР*
+КРЕДИТ-ОР-К*А
+КРЕДИТ-ОР*Ы
+КРЕДИТ*Ы
```

(1a)

```
+БЕД-Н*ЕТЬ
+БЕД-ОВ*АТЬ
+БЕД-СТВ-ОВ*АТЬ
-НА+БЕД-СТВ-ОВ*АТЬ-СЯ
-О+БЕД-Н*ЕТЬ
-О+БЕД-Н*ИТЬ
-О+БЕД-Н*ЯТЬ
-О+БЕД-Н*ЯТЬ-СЯ
-ПО+БЕД-СТВ-ОВ*АТЬ
-ПРИ+БЕД-Н*ИТЬ-СЯ
-ПРИ+БЕД-Н*ЯТЬ-СЯ
```

(1b)

```
-НЕ+СИСТЕМ-АТ-ИЗ-ИР-ОВ-АНН*ЫЙ
-НЕ+СИСТЕМ-АТ-ИЧ-ЕСК*ИЙ
+СИСТЕМ-АТ-ИЗ-ИР-ОВ-АНН*ЫЙ
+СИСТЕМ-АТ-ИЗ-ИР-УЮЩ*ИЙ
+СИСТЕМ-АТ-ИЧ-ЕСК*ИЙ
+СИСТЕМ-АТ-ИЧ-Н*ЫЙ
+СИСТЕМ-Н*ЫЙ
```

(1c)

В любом слове префиксов не более трех, суффиксов — не более шести, а корень, окончание и возвратная частица единственны.

Для автоматического формирования П-словаря специальная программа дополнительно определяет необходимые морфологические категории слов в ОКЧ-словаре. Для глаголов, прилагательных и наречий находится только часть речи, а для существительных — еще число и род.

4. Статистическое обследование К-словаря

Аффиксное сходство однокоренных слов изучалось на материале К-словаря [6], который фактически служил нам обучающим массивом, воплощающим лингвистическую интуицию. В К-словаре содержится 1100 так называемых паронимических рядов из 2–7 слов. Слова паронимического ряда относятся к одной части речи (существительные, глаголы или прилагательные) и упорядочены по алфавиту. Тем самым они неявно считаются равноправными внутри своего ряда, т. е. любое из них паронимично всем остальным.

Паронимические ряды К-словаря и пары слов из одного ряда обследовались визуально. При сравнении существительных одного ряда не учитывались пары, различные по роду и/или числу, но добавлялись множественные числа тех существительных, которые таковые имеют. В глагольные ряды добавлялись глаголы другого вида, если таковой у них существует.

Для всех рассмотренных таким образом пар слов их морфемный состав брался из ОКЧ-словаря, и подсчитывались расстояния *Np* и *Ns*.

Всего в К-словаре было насчитано 3297 пар. Статистика аффиксного расстояния представлена вторым и третьим столбцами Таблицы 1. Если построить ранговое распределение статистических данных, то первые два ранга займут пары, различающиеся только одним аффиксом, причем больше всего пар различается только одним суффиксом. Третий и четвертый ранг занимают пары, различающиеся двумя аффиксами.

Неожиданно большое количество пар слов (ранг 5 статистического распределения) оказалось на минимальном расстоянии (0, 0), т. е. когда слова

имеют одинаковый морфный состав. Сюда попали существительные с алломорфизмом корня (*отечество — отчество*), глаголы и прилагательные с алломорфными корнями и/или разными окончаниями (*воскресать — воскресить — воскрешать, временный — временной*).

Легко видеть, что набор из 7 расстояний: (0, 0), (0, 1), (1, 0), (0, 2), (1, 1), (1, 2), (0, 3), выделенных в таблице контрастом, покрывает 99,5% всех рассмотренных пар; его мы и берем в качестве **критерия аффиксного сходства**. Можно записать этот критерий в виде формулы:

$$(Np = 0)\ \&\ (Ns <= 3)\ V\ (Np = 1)\ \&\ (Ns <= 2).$$

Словесная формулировка критерия такова: либо префиксы в сравниваемой паре одинаковы, а различий в суффиксах не более трех, либо у них один различный префикс, а различий в суффиксах не более двух**.** Как видим, лингвистическая интуиция составителя К-словаря допускает у паронимов больше различий в суффиксах, чем в префиксах, и предпочитает считать паронимами слова с одинаковыми началами.

Таблица 1. Статистика аффиксного сходства

	К-словарь		**ОКЧ-словарь**		
Np, Ns	**Число**	**%**	**Число**	**%**	**Примеры**
0, 0	144	4,3	1152	1,3	*отечество — отчество, невежа — невежда, осветить — осветлить, заспавший — засыпавший*
0, 1	1034	31,4	7267	8,0	*корона — коронка, доносить — донашивать, маленький — малый, прогулы — прогулки, двигатель — движитель*
1, 0	990	30,1	35723	39,3	*вход — выход, входить — выходить, входной — выходной, ходить — сходить, выйти — пойти*
0, 2	535	16,6	6053	6,7	*манера — манерность, актировать — активизировать, стрелковый — стреляный*
1, 1	472	14,3	23470	25,8	*аккредитация — кредитка, проведать — выведывать, означенный — назначаемый*
2, 0	4	0,1	1264	1,4	*ход — перерасход, означить — переназначить, означенный — переназначенный*
0, 3	91	2,8	848	0,9	*акт — активатор, актерствовать — активизировать, актовый — активирующий*
1, 2	10	0,3	9875	10,9	*болезнь — заболевание, актировать — дезактивировать, актированный — дезактивизированный,*

Np, *Ns*	К-словарь		ОКЧ-словарь		Примеры
	Число	%	Число	%	
2, 1	0	0,0	1215	1,3	*запредельность — разделенность, ходули — перерасходы, надуманный — понапридумавший*
3, 0	0	0,0	29	0,0	*деление — перераспределение, задумывать — понапридумывать*
Проч.	14	0,4	4116	4,5	*мерзость — омерзительность, политизированность — аполитичность, опубликованный — публицистический*

Таким образом, в виду крайней редкости отбрасываемых нами случаев из К-словаря, мы считаем **формальными паронимами** слова одной части речи и единого корня, аффиксное расстояние между которыми удовлетворяет указанному выше критерию.

Представленные в таблице пары, не отнесенные к формальным паронимам, как правило, брались из ОКЧ-словаря. Обычно они внешне несходны, например, пара ***ход**ули — перерас**ход**ы*. Однако слова пары ***мерз**ость — о**мерз**ительность* кажутся схожими.

5. Формирование словаря паронимов

Для автоматического построения словаря паронимов был взят ОКЧ-словарь, подвергнутый морфемному разбору и морфологической категоризации. Каждая группа словаря из *M* слов преобразуется в *M* статей: одно слово исходной группы становится головным для статьи, а *M-1* остальных, подчиненных слов упорядоченно следуют за ним. Вычисляются значения *Np* и *Ns* для всех пар <головное слово, подчиненное слово>. После отсева подчиненных слов, не отвечающих формальному критерию паронимии с головным, все статьи, в которых осталось хотя бы одно подчиненное слово, включаются в П-словарь. Например, группа (1c) однокоренных слов из семи прилагательных переходит в следующие семь статей, где число слов, паронимичных головному, колеблется от одного до четырех:

```
НЕСИСТЕМАТИЗИРОВАННЫЙ
СИСТЕМАТИЗИРОВАННЫЙ =ANT

НЕСИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧЕСКИЙ =ANT
  СИСТЕМАТИЧНЫЙ ~ANT

СИСТЕМАТИЗИРОВАННЫЙ
  НЕСИСТЕМАТИЗИРОВАННЫЙ =ANT
  СИСТЕМАТИЗИРУЮЩИЙ

СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЗИРОВАННЫЙ
  СИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧНЫЙ

СИСТЕМАТИЧЕСКИЙ
  НЕСИСТЕМАТИЧЕСКИЙ =ANT
  СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЧНЫЙ ~SYN
  СИСТЕМНЫЙ

СИСТЕМАТИЧНЫЙ
  НЕСИСТЕМАТИЧЕСКИЙ ~ANT
  СИСТЕМАТИЗИРУЮЩИЙ
  СИСТЕМАТИЧЕСКИЙ ~SYN
  СИСТЕМНЫЙ

СИСТЕМНЫЙ
  СИСТЕМАТИЧЕСКИЙ
  СИСТЕМАТИЧНЫЙ (2)
```

Для существительных дополнительно учитывается род и число слов, и если подчиненное слово отличается от головного по этим параметрам, оно

автоматически исключается из статьи. Вот итоговые статьи, сформированные на основе группы (1а):

```
АККРЕДИТАЦИЯ        КРЕДИТ              КРЕДИТОР            КРЕДИТЫ
  КРЕДИТКА            АККРЕДИТИВ          АККРЕДИТИВ          АККРЕДИТИВЫ
  КРЕДИТОРКА          КРЕДИТОР            КРЕДИТ              КРЕДИТКИ
                                                              КРЕДИТОРЫ
АККРЕДИТИВ          КРЕДИТКА            КРЕДИТОРКА
  КРЕДИТ              АККРЕДИТАЦИЯ        АККРЕДИТАЦИЯ
  КРЕДИТОР            КРЕДИТОРКА          КРЕДИТКА

АККРЕДИТИВЫ         КРЕДИТКИ            КРЕДИТОРЫ
  КРЕДИТКИ            АККРЕДИТИВЫ         АККРЕДИТИВЫ
  КРЕДИТОРЫ           КРЕДИТОРЫ           КРЕДИТКИ
  КРЕДИТЫ             КРЕДИТЫ             КРЕДИТЫ
```

Существительное кредитование оказалось единственным в подгруппе слов среднего рода и поэтому паронимической статьи не породило. Приведенные примеры показывают, что число паронимов у разных слов из одной группы слов с одинаковым корнем и частью речи может существенно различаться.

Подсчитанная в ходе преобразования ОКЧ-словаря статистика аффиксного сходства пар слов представлена в четвертом и пятом столбцах Таблицы 1. Для сопоставимости с К-словарем были исключены наречия (их всего 2,2 %). Абсолютные цифры (четвертый столбец) значительно больше, чем в К-словаре, но процентные показатели (пятый столбец) в какой-то степени схожи. Косинус второго и четвертого столбцов, рассматриваемых как векторы, равен 0,79, что дополнительно подтверждает принятый нами формальный критерий паронимии.

Результирующий словарь характеризуется следующими параметрами:

Количество статей (= головных слов)	21 802
Количество подчиненных им паронимов	192 024
Среднее число паронимов в статье	8,8

Статей стало на 7 % меньше, чем слов в ОКЧ-словаре. При формировании статей произошел отсев многих однокоренных слов. Для существительных коэффициент отсева равен 3,73, для глаголов — 1,79, для прилагательных — 1,46. Тем самым, паронимия достаточно селективна, т. е. равенство корней и частей речи не гарантирует паронимию. Среднее число паронимов на статью оказалось довольно большим из-за объемных групп глаголов в ОКЧ-словаре.

Отметим, что наш критерий паронимии допускает, что паронимы могут быть синонимами (патетический — патетичный) или антонимами (типичный — атипичный), это не исключается и в К-словаре. В примере (2) показаны статьи П-словаря с автоматически размеченными (средствами компьютерного словаря [2]) синонимами и антонимами для головного слова. За исключением абсолютных синонимов, совокупности диагностирующих контекстов у таких пар слов чаще всего различны, и поэтому разумно хранить всех их в словаре паронимов. Это касается и абсолютных синонимов, имеющих разные паронимы. И только тогда, когда абсолютные синонимы образуют изолированную пару типа апельсиновый — апельсинный, их в паронимический словарь включать нецелесообразно.

6. Заключение

В результате обследования наиболее крупного печатного словаря русских паронимов мы нашли ту меру аффиксного сходства однокоренных слов, которая более чем в 99% случаев гарантирует паронимию в ее интуитивном понимании. Предложен формальный критерий паронимии: два слова паронимичны, если имеют одинаковый корень, принадлежат одной части речи (в уточненном ее понимании), и их аффиксные различия находятся в строго установленных рамках.

С использованием формального критерия на базе компьютерного словаря однокоренных слов автоматически построен словарь русских паронимов, по объему превышающий все известные словари. Главное приложение построенного словаря — подбор слов–замен для автоматизированного исправления паронимических ошибок в текстах.

Мы не исключаем дальнейшего уточнения предложенного критерия, чтобы допустить пары с одинаковыми конечными суффиксами (мерзость — омерзительность), и в то же время исключить внешне мало похожие пары (аккредитация — кредитка). Для других приложений может потребоваться несколько иное уточнение понятия паронимии, но в любом случае исходным ресурсом для построения словаря паронимов может браться все тот же словарь однокоренных слов.

Литература

1. *Belchikov, Yu. A., Panjusheva M. S.* (2004) Dictionary of Russian Paronyms [Slovar' paronimov russkogo jazyka] Moscow, Russkij Jazyk.
2. *Bolshakov, I. A.* CrossLexica: A large electronic dictionary of collocations and semantic links between words in Russian. [KrossLeksika — bolshoj èlektronnyj slovar' sochetanij i smyslovykh svjazei russkikh slov]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009" [Computational Linguistics and Intelligent Technologies: Proc. of the International Conference "Dialogue 2009"]. Moscow, 2009, pp. 45–50.
3. *Bolshakov, I. A., Gelbukh, A.* On Detection of Malapropisms by Multistage Collocation Testing. // A. Düsterhöft, B. Talheim (Eds.) Proc. 8th Intern. Conference on Applications of Natural Language to Information Systems NLDB´2003, Burg, Germany, GI-Edition, LNI V. P-29, Bonn, 2003, p. 28–41.
4. *Bolshakova, E. I., Bolshakov I. A.* (2007) Automatic detection and computer-aided correction of Russian malapropisms [Avtomaticheskoe obnaruzhenie i avtomatizirovannnoe ispravlenie russkikh malapropizmov] // Nauchnaya i Tekhnicheskaya Informatsiya. Ser. 2, No. 5, 2007, p. 8–13.
5. *Fowler, H. W.* (1994) Dictionary of Modern English Usage. Wordsworth Editions Ltd.
6. *Krasnykh, V. I.* Explanatory Dictionary of Russian Paronyms [Tolkovyj slovar' paronimov russkogo jazyka] Moscow, AST Astrel, 2007.
7. *Merriam* Webster's Collegiate Dictionary (1993) Merriam-Webster Inc.
8. *Vishnjakova, O. V.* (1984) Dictionary of Russian Paronyms [Slovar' paronimov russkogo jazyka] Moscow, Russkij Jazyk.

МОДЕЛИРОВАНИЕ НЕТРИВИАЛЬНЫХ УСЛОВИЙ ПОНИМАНИЯ СООБЩЕНИЯ (НА ПРИМЕРЕ ИРОНИИ)

Борисова Е. Г. (efcomconf@list.ru)
Московский городской педагогический университет, Москва, Россия

Пирогова Ю. К. (adv-pirogova@yandex.ru)
НИУ Высшая школа экономики, Москва, Россия

Рассматривается модель понимания сообщения, применимая для описания случаев, отклоняющихся от тривиальных, буквальных, («что сказано, то и имелось в виду»), и предполагающих дополнительные действия слушающего. В качестве примера выбрана ирония как наиболее сложный случай кодировки намерений говорящего. Отмечаются лингвистические средства маркировки таких случаев. Описание позволяет вскрыть различные фрагменты смысла высказывания и прагматических характеристик, которые должны приниматься во внимание при моделировании понимания как части динамической модели языка.

Ключевые слова: прагматика, ментальная активность, понимание, ирония

MODELING PECULIAR CONDITIONS OF UNDERSTANDING UTTERANCES (THE CASE OF IRONY)

Borisova E. G. (efcomconf@list.ru),
Moscow Teachers' Training University, Moscow, Russia

Pirogova Yu. K. (adv-pirogova@yandex.ru),
National Research University Higher School of Economics, Moscow, Russia

The article deals with modeling the understanding of natural language texts in special cases that differ from the trivial 'normal' condition 'what is said is what is meant' (literal understanding). This includes hints, metaphors etc. The article is focused on irony, which seems to be a paradox: 'what is meant' is different from 'what is said'. By thorough analysis of examples of irony both in the literature and in common usage (including texts of media and the Internet) we classify the cases of irony. The sense components of utterances which are to be understood in the opposite way have been identified. They are not only parts of the dictum but of the modal frame as well. The pragmatic analysis showed the intentions of the Speaker using irony, including the cases when the object of mockery is the Speaker himself, or the Hearer. Correlation of irony vs. mockery and irony vs. quotations is investigated. The results should be used by designing the model of natural text understanding.

Key-words: pragmatics, mental activity, understanding, sarcasm

1. Моделирование понимания вне тривиальных («буквальных») текстов

Моделирование понимания в рамках интерактивного подхода (учитывающего действия участников общения) не ограничивается распознаванием формы, дешифровкой (определением семантики частных значений лексенмы, так называемых лексико-семантических вариантов), «сложением», т. е. синтезом возможного смысла сообщения. Действия слушающего, как теперь очевидно, включают в себя и возможные выводы (импликатуры) из сказанного, и перебор возможных вариантов понимания с учетом «угадывания» намерений говорящего («если бы говорящий имел в виду X, он бы скорее сказал X, а не Y, как тут») и ряд других действий[1]. Такая более громоздкая модель может применяться не только для моделирования поведения участников общения в «нормальных», тривиальных условиях, когда говорящий стремится к наиболее полному и однозначному пониманию («что имел в виду, то и сказал», т. е. понимать надо буквально). Можно обратиться и к более сложным случаям понимания сообщений, когда буквальное понимание дает абсурдный результат, а постулаты Грайса не выполняются (заметим, что последнее не всегда имеет место, однако то, что такое общение тоже возможно, уже неоднократно замечалось), как, к примеру, при буквальном понимании фразы:

(1) *Он для меня готов горы свернуть. — Он что, взрывник?*
(пример взят из наблюдений авторов)

---

[1] *Борисова Е. Г.* Интерактивный подход в лингвистике: пределы применимости // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог — 2006» / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея — М.: Изд-во РГГУ, 2006 — с. 84–88.

Наиболее простой случай отклонения от буквального понимания — это, видимо, намек[2]. Это высказывание, которое содержит в себе некоторые инструкции по формированию импликатур, желательных для говорящего, например,

(2) *Что-то стало холодать.*

Отсутствие иллокутивной ценности высказывания («Зачем об этом говорить?») должно заставить адресата сделать возможные выводы, например, «Надо закрыть окно», «Ты, милый, должен принести мне плед» и т. п. Если говорящий хочет быть понят точнее, он выдаст больше информации, например, добавит:

(3) *А окошко все еще открыто* и т. п.

Как мы уже отмечали, импликатуры обязательно сопровождают понимание сообщения[3]. В случае намека говорящий так выстраивает свои высказывания, что они не имеют ценности сами по себе и этим стимулируется действие слушающего. Заметим, что для понимания намека необходим полный набор информации, сопровождающий каждый речевой акт: понимание языковых единиц, знание контекста и наличие общих сведений у участников общения, правила выводов, для намеков с перлокутивными целями — представления об иерархии, речевом этикете (какие-то просьбы неприлично высказывать в лоб) и т. п. Как отмечалось[4], информативность намека является следствием постулатов Грайса, а именно, принципа релевантности[5], хотя нарушаются многие другие постулаты. В той или иной степени, это можно отнести и к другим случаям нетривиальной подачи информации.

---

[2] *Баранов А. Н.* Намек как способ косвенной передачи смысла // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог — 2006» / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея — М.: Изд-во РГГУ, 2006 — с. 46–51.
Кобозева И. М., Лауфер Н. И. Об одном способе косвенного информирования // Известия АН СССР. Сер. лит. и яз. 1988. Т. 47, № 5. С. 462–470
Шатуновский И. Б. 6 способов косвенного выражения смысла // Труды международной конференции «Диалог — 2004». М., 2004.

[3] *Borisova E.* Special Entities Used for Governing the Processes of Understanding/Understanding by communication, eds. E. Borisova, O. Souleimanova. Cambridge Scholars Publishers, 2013 — pp. 95–103.

[4] *Кобозева И. М., Лауфер Н. И.* Об одном способе косвенного информирования // Известия АН СССР. Сер. лит. и яз. 1988. Т. 47, № 5. С. 462–470

[5] *Wilson Deirdre*. On Verbal Irony / Deirdre Wilson, Dan Sperber // Irony in Language and Thought: A Cognitive Science Reader. — New York Lawrence Erlbaum Associates. 2007, — pp. 35–55

2. Ирония как высказывание с намеренно «противоположным» пониманием

Остановимся на наиболее парадоксальном случае использования языковых средств — на иронии. По Квинтиллиану, ирония — это высказывание, которое надо понимать в противоположном смысле. Чистым случаем иронии можно считать высказывание:

(4) *Ну, такой герой добьется сокрушительной победы!* (пример взят из наблюдений авторов)

Однако в русском языке слово ирония, ироничный используется гораздо шире. Анализ бытовых употреблений слова ирония показывает, что в русском языковом сознании к иронии относится и насмешка, не связанная с «обратным» пониманием:

(5) *Когда я прохожу мимо фотографии Анисимова на филфаке, ирония прищуренных глаз напоминает мой провальный ответ и «трояк», который должен был незамедлительно перекочевать в зачетку, но почему-то превратился в «четыре» в последний момент.* [Нина Щербак. Роман с филфаком // «Звезда», 2010] [пример из НКРЯ]

Вряд ли взгляд способен передавать «обратное» понимание», отличающее иронию от насмешки. Насмешка же возможна и при буквальном употреблении слов:

(6) *Что-то ты у нас такой тихий всегда, когда доходит до дела! ?* (пример взят из наблюдений авторов)

Здесь действительно текст понимается буквально, что не мешает передавать насмешливое отношение к человеку. В русском языке прилагательное ироничный *(ироничный взгляд, тон)* может означать насмешливый, поскольку далеко не всегда сопровождает «обратное» понимание сказанного. Подобное расширенное употребление слова ирония и однокоренных встречается и в некоторых научных работах, что, видимо, указывает на то, что это слово не устоялось как термин. Аналогичное замечание можно сделать и относительно других языков, в частности, английского.

В большинстве случаев ирония, действительно, содержит и насмешку (которую можно интерпретировать как «я утверждаю Р и считаю, что Р может вызвать смех»). Однако, некоторые примеры — в основном, сопряженные с гневом, огорчением, злостью — вряд ли могут трактоваться таким образом:

(7) *Мы ничего не делали, ничем не рисковали, не смотрели из окопов на приближающиеся танки, зажав в руке бутылку с зажигательной смесью! ?* (пример взят из наблюдений авторов)

Говорящий в отрицательных предложениях явно перечисляет то, что на самом деле делал, в чем собеседник, видимо, усомнился, и это вызвало данное — ироническое, но отнюдь не насмешливое — высказывание. Заметим, что в случае отрицательного отношения к предмету сообщения обычно говорят не об иронии, а о сарказме. Не отрицая полезности этого термина для литературоведческих и других исследований, мы не будем его употреблять, считая, что все случаи сарказма подходят под определение иронии.

Во многих случаях ирония используется как средство передачи отрицательного отношения — к предмету обсуждения или к участникам общения (как данного акта общения, так и передаваемого). Однако степень осуждения может варьировать от мягкой или даже исчезающей, когда, к примеру, взрослые иронизируют над детьми, до очень резкой, как в примере (7) по отношению к адресату.

Таким образом, при принятом нами (с опорой на классическое) определении иронии как «использования слова, фрагмента или высказывания целиком для передачи прямо противоположного смысла» нам приходится оставить за рамками рассмотрения немало примеров, которые в русском языковом сознании относятся к проявлениям иронии, определив их как насмешку. Однако и оставшиеся случаи демонстрируют большое разнообразие. Поэтому моделирование понимания при использовании иронии должно учитывать большое разнообразие эффектов в зависимости от адресации иронии, месте «переворачиваемого» смысла в содержании сообщения, а также специальных маркеров — языковых единиц, помогающих адресату распознать иронию.

3. Семантика смысла, подвергаемого иронической трансформации

Наиболее очевидным случаем иронии является использование в противоположном смысле основного содержания (диктума) сообщения:

(8) *Он такой маленький, в толпе потеряется* (пример взят из наблюдений авторов)

— если известно, что человек высок ростом;

(9) *Погодка — мечта* (пример взят из наблюдений авторов)

(в ситуации, когда погода плохая).

Однако «обратное» понимание возможно и относительно других фрагментов смысла сообщения. В частности, иронически может пониматься часть сообщения, содержащая в себе оценку:

(10) *И он притащил свой «восхитительный» аппарат.* (пример взят из наблюдений авторов)

(11) *Отколе, **умная**, бредешь ты голова? -*
Лисица, встретившись с ослом, его спросила.
(И. А. Крылов, Лисица и Осел»)

Эту цитату часто приводят как пример иронии. Здесь действительно значение слово «умный» должно пониматься в противоположном смысле. Однако это словосочетание не меняет понимание высказывания в целом — Лисица действительно спрашивает Осла, а «обратное» понимание относится только к дополнительной характеристике.

Ироническая оценка может выражаться и еще менее эксплицитно:

(12) *что же сделал я за пакость,*
Я, убийца и злодей? (Б. Пастернак. Нобелевская премия).

Поэт описывает гонения после присуждения Нобелевской премии. Самоосуждающие характеристики явно выглядят цитированием гонителей. Вообще иронические высказывания часто понимаются как отсылка к чужим слова, цитирование или хотя бы цитация (использование неавторских средств номинации[6]). Однако в целом это качество нельзя считать обязательным — в большом количестве примеров: (8), (9), (11) и др. говорящий сам дает иронические номинации, и источник цитирования не просматривается.

Иронической может быть модальность:

(13) *Вот бы мне туда!* (пример взят из наблюдений авторов)

В ситуации, когда описывается неприятное место (в данном контексте описывались некоторые сложности службы в армии), модальность «я хочу туда» распознается как ироническая, понимаемая в обратном смысле.

Еще более тонкий (но нередкий!) вариант иронии: «перевернутым», неправильным оказывается часть смысла, отражающая степень проявления какого-либо признака

(14) *В дворницкой стоял запах гнющего навоза, распространяемый новыми валенками Тихона. Старые валенки стояли в углу и воздуха тоже не озонировали.* (Ильф и Петров «Двенадцать стульев»)

---

[6] *Зализняк А. А.* Семантика кавычек// Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог — 2000» / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея — М.: Изд-во РГГУ, 2007 — с. 188–193.
Шилихина К. М. Ирония в политическом диалоге // Политическая лингвистика. № 4 (38) 2011. Стр. 177–182.
Wilson Deirdre. On Verbal Irony / Deirdre Wilson, Dan Sperber // Irony in Language and Thought: A Cognitive Science Reader. — New York Lawrence Erlbaum Associates. 2007, — pp. 35–55

Здесь сообщение, что валенки не способствовали улучшению состава воздуха, должно пониматься вполне в прямом смысле. Однако использованный оборот явно в недостаточной степени отражает, насколько на самом деле воздух был испорчен запахом прелой обуви. Поэтому иронически воспринимается часть смысла «не улучшали состояния воздуха, создавая некоторую степень неудобства», в противоположном смысле понимается «некоторая степень». — степень была не неопределенной, а значительной.

Наконец, часть случаев, обычно называемых иронией, заключаются в «обратном» понимании информации, передаваемой формой, стилем, штампом[7]. Иронией будет использование газетного штампа в примере:

(15) *Тогда слуги народа будут думать не о том, как содрать взятку с собственника, а как заставить его исправно платить все положенные налоги.* [Николай Анисин. Кабала хапуг (к проблемам нынешних отношений власти и бизнеса в России) (2003) // «Завтра», 2003.03.26] [пример из НКРЯ]

Здесь штамп *слуги народа*, характерный для «высокого», пафосного стиля, употреблен в другой ситуации, которую можно рассматривать как противоположную. Т.е. именно компонент стилевой окраски должен пониматься в противоположном смысле. Само же сочетание *слуга народа* (описательное выражение для «депутат» и «чиновник») не меняет свой смысл — имеются в виду те же люди. (Впрочем, ироническое значение этого штампа начинает вытеснять исходное, чего десять лет назад, когда писалось приведенное высказывание, видимо, еще не было).

Подобные случаи иронического употребления штампов широко распространены и стали частью молодежной субкультуры в конце 80-х годов под названием *стёб* (носители интерпретируют это сленговое слово как «издевательство»). Он по-прежнему распространен в языке некоторых СМИ.

4. Прагматические аспекты употребления иронии

Поскольку ирония была «открыта» как один из тропов, то есть выразительных приемов, то по большей части ее изучали в рамках анализа художественного текста. Впрочем, с самого начала была очевидна и другая сфера ее применения — это риторическая фигура, активно использовавшаяся в публичных выступлениях. В этом смысле роль иронии в СМИ, можно сказать, наследуется из риторики. Но, как отмечают наиболее проницательные исследователи[8], иро-

---

[7] Шилихина К. М. Ирония как эффект языковой игры // Язык, коммуникация и социальная среда. Выпуск 8. Воронеж: ВГУ, 2010. С. 37–45.

[8] *Шилихина К. М.* Ирония в политическом диалоге // Политическая лингвистика. № 4 (38) 2011. Стр. 177–182

ния является важной и весьма распространенной частью повседневного общения: устной беседы, письма и т. п. При этом иронический способ выражения позволяет передать ряд дополнительных смыслов, актуальных в общении.

Одним из важнейших можно считать, пожалуй, «присоединение»: говорящий, употребляя нетривиальный способ выражения, показывает адресату, что рассчитывает на его понимание. Это можно рассматривать как знак приобщения к «своим», «посвященным», что является и ‚в некотором роде» лестью говорящего адресату. Это весьма сильное средство воздействия, поскольку адресат, которого уже объявили таким образом «приобщенным», уже не захочет отвергать содержания информации. Наверное, именно этим можно объяснить широкое распространение в печати так называемого стёба: иронической подачи позиций противника.

Несомненно, ирония достаточно часто используется для передачи отрицательной оценки содержания сообщения, см. пример (7). Возможно, это связано с тем, что ирония часто представляет собой цитирование тех, кого подвергают осмеянию[9]. Говорящий не просто заявляет эксплицитно, что некоторое выказывание неверно, но предлагает лично убедиться адресату очевидности неправоты.

Отрицательное отношение такого же источника имеет место даже в еще большей степени, когда место насмешки заменяет гнев, как в примере (7).

Вообще наличие юмористической, игровой компоненты — достаточно частый спутник иронического употребления[10], но необязательный. Вряд ли можно говорить об игре и юморе в примере (12) Б. Пастернака.

Намерения говорящего во многом определяются тем, насколько роль «сообщника» и «объекта осуждения» распределены среди участников общения. Классическим случаем является осуждение чего-либо внешнего по отношению к участникам общения (говорящему и адресату). Однако объектом осуждения может стать и сам говорящий (или нечто, с ним связанное) — тогда мы говорим о самоиронии:

(16) *Нда. Я, как всегда, создаю общую теорию всего. В результате ни одной статьи не могу подать вовремя.* (пример взят из наблюдений авторов)

Заметим, что этот пример отличается от пастернаковского (12), самоиронией не являющегося, тем, что здесь говорящий осуждает себя, тогда как Б. Пастернак осуждает авторов цитируемых характеристик, и иронию к себе не прилагает.

---

[9] *Wilson Deirdre.* On Verbal Irony / Deirdre Wilson, Dan Sperber // Irony in Language and Thought: A Cognitive Science Reader. — New York Lawrence Erlbaum Associates. 2007, — pp. 35–55.
Зализняк А. А. Семантика кавычек// Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог — 2000» / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея — М.: Изд-во РГГУ, 2007 — с. 188–193.

[10] *Санников В. З.* Русская языковая шутка: От Пушкина до наших дней / В. З. Санников. — М.: Аграф, 2003. — 556 с.
Crystal, D. Language Play / D. Crystal. — Chicago: The University of Chicago Press, 2001. — 248 p.

Объектом иронии бывает адресат. Этот случай может распадаться на несколько частных. Собеседник может быть открыто сделан объектом иронии:

(17) *Ты ж у нас вундеркинд, что тебе какие-то лабораторные (работы).*
(пример взят из наблюдений авторов)

Однако встречается и скрытое –для собеседника — иронизирование над ним, возможно, для него непонятное, адресованное другим — читателям описываемого случая, свидетелям происшествия.. Именно так можно расценивать эпизод из лермонтовского «Героя нашего времени»

(18) *Ты, говорят, эти дни ужасно волочился за моей княжной? — сказал он довольно небрежно и не глядя на меня.*

— Где нам, дуракам, чай пить! — отвечал я ему, повторяя любимую поговорку одного из самых ловких повес прошлого времени, воспетого некогда Пушкиным.

Ирония не была воспринята адресатом, но читателями романа — вполне.

Мы здесь не можем привести полный перечень возможных объектов иронии. Отметим только, что в зависимости от роли объекта в речевом акте наблюдаются весьма разнообразные намерения говорящего, решившего использовать прием иронии. Таким образом, ирония позволяет вскрыть (или подтвердить) различные типы прагматики речевого поведения — более сложные, чем просто понимание сообщения адресатом.

5. Каким образом слушающий понимает говорящего?

Если исходить из того, что говорящий надеется на понимание (а это так, по крайней мере, для части адресатов сообщения, за исключением редких случаев, которые будут рассмотрены позднее), следует выделить условия, обеспечивающие все те выводы, которые необходимо сделать для понимания сообщения. Определение этих условий исключительно важно для автоматического анализа текста, к примеру, при автоматическом реферировании СМИ. В этом направлении ведутся активные работы[11], результаты которых мы учитываем в предлагаемом перечне.

Одним из важнейших факторов является контекст — как широкий (общие знания об устройстве мира, о взаимоотношениях актантов описываемой ситуации, о семантико-прагматических характеристиках употребляемых знаков), так и узкий — данный текст (монологического или диалогического характера)

---

[11] *Dmitry Davidov, Oren Tsur, and Ari Rappoport.* Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In Proceeding of Computational Natural Language Learning (CoNLL 2010), Uppsala, Sweden, July. 2010.

и характеристики акта общения (например, представления о наличии здравого смысле и чувства юмора у говорящего). Осознание того, что буквальное понимание невозможно, нередко может быть спусковым крючком для включения механизма выводов при различных употреблениях тропов. (Сюда же следует отнести и представление о жанрах, например, при чтении памфлета логично искать иронию).

Рассмотрим пример из обсуждения статьи на ресурсе mirtesen. Vybor 2012 (пример взят 27.01.13). В материале для обсуждения приводилось мнение Дж. Сороса о проблемах в российской экономике (орфография и пунктуация сохраняются):.

(19) *Илья* 10928 *да кто такой сорос? не знаем мы никакого сороса! сорос в америке и ниче про нас не знает. мы верим тока нашему премьеру медведеву. он сказал, все ништяк!!! а он явно в теме: у него айфон есть и он умеет твиттером пользоваться*

Валерий Ковальчук 12950 *Илья, кто то минус ставил вам, не поняли что это ирония.*

Начало формы

Конец формы

25 янв, 22:06 ответить

Илья 10928 *валере. ну наконец-то кто-то догадался. я уж думал совсем плохо с юмором.*

Обсуждение на этом ресурсе (в отличие от форумов и групп, объединяющих более или менее схожих по воззрениям пользователей) привлекает людей (а возможно, и боты) очень разных взглядов и позиций. Для понимания сообщения Ильи как иронического необходимы такие фрагменты картины мира, как представление об известности Дж.Сороса, о том, что умение пользоваться современными средствами связи вовсе не является доказательством высокой компетентности. Наконец, и то, что премьер может быть объектом насмешек. Видимо, далеко не все участники обсуждения были убеждены, что у автора записи есть эти фрагменты в картине мира, почему и ставили ему минусы в знак неодобрения высказывания, которое они поняли буквально.

Заметим, что при выявлении иронических употреблений в политических коммуникациях, что важно для задач автоматического реферирования СМИ, роль контекста в широком смысле играет тип политического дискурса: слова, характерные для, к примеру, коммунистичесгого дискурса — манифестативные (manifestive[12]) трудящиеся,на благо Родины и т. п., в либеральном дискурсе будут употребляться или в составе цитаты, или иронически, ср. заголовок

[12] *Borisova, E.* “Opposition Discourse in Russia: Political Pamphlets 1989–1991.” In Political Discourse in Transition in Europe 1989–1991, edited by P. Chilton, M. Ilyin and J. Mey, 111–130. Amsterdam/ Philadelphia: Jon Benjamins, 1998.

М. Ганапольского «На благо Родины и ее бабла». *(«Эхо Москвы», 12.02.2013, www.echo.msk.ru/blog/ganapolsky/1010574-echo/)*

Второй важный фактор — специальные маркеры «иронического понимания». Они могут носить фонетический (интонационный, фонический), графический, лексический, стилистико-прагматический характер. Причем встречаются и знаки, которые специально предназначены для выражения иронии — утрированная интонация, слова и сочетания «Щазз!», «Бегу и падаю», «Дадут — догонят и еще добавят» и в этих случаях феномен иронии несколько смывается. Чаще всего роль маркеров иронии играют слова и выражения, которые представляют собой преувеличение каких-то характеристик слова: эмоциональной окраски, принадлежности к определенному языку вплоть до штампов и шаблонов. Именно такими маркерами являются наименования «отец народов», «лучший друг физкультурников» по отношению в И. В. Сталину. Фактически утрируются те характеристики языковых единиц, которые позволяют увидеть несоответствие их употребления замыслу говорящего, например, неуместный пафос, ср. «нажитое непосильным трудом».

Важно отметить, что понимание наступает не всегда. Даже если не считать случаев коммуникативной неудачи, когда задуманный смысл, переданный иронически, был понят буквально, есть немало вариантов того, что автор сообщения и рассчитывал на то, что часть адресатов поймет текст в буквальном смысле, не разглядит иронии. Это может иметь место, если автор хочет обезопасить себя с точки зрения правовых норм: он может утверждать, что даже не подозревал о скрытой иронии в собственных словах:

(20) *В прошлом Ольга Юрьевна прошла суровую школу дикого капитализма в ОНЭКСИМе, «Норникеле», РСПП. Теперь внедряет государственные концепции, ориентированные на благо простого человека….*
Ну не представляет же она в конце концов интересы олигархов в правительстве? (ЛГ. № 6, 2013, с 10)

Если бы последнее было заявлено как прямое утверждение, автору грозило бы преследование за клевету или нанесение ущерба деловой репутации Ольги Юрьевны.

А «ирония не может быть предметом судебного разбирательства, так как связана с восприятием; одно и то же ироническое высказывание разные люди воспринимают по-разному: одни как положительную, добрую иронию, другие — как отрицательную, злую, а третьи вообще не воспринимают ироническую окраску[13]».

Если в медиа ирония распространена исключительно широко, то в казалось бы близком дискурсе — рекламном[14] — иронию обнаружить непросто.

---

[13] *Как провести лингвистическую экспертизу спорного текста?* Памятка для судей, юристов СМИ, адвокатов, прокуроров, следователей, дознавателей и экспертов / Под ред. проф. М. В. Горбаневского. — 2-е изд., испр. и доп. — М.: Юридический Мир 2006, — 112 с.

[14] *Pirogova Yulia.* Message perception and comprehension in Marketing communication discourse/ Understanding by communication, eds. E. Borisova, O. Souleimanova. Cambridge Scholars Publishers, 2013 — pp. 176–210.

Большинство случаев, интерпретируемых как ирония, должны быть отнесены к насмешке, т. к. в них нет «переворота» смысла. Те случаи, что все-таки предполагают подобное мыслительное действие, чаще всего содержат подсказку в виде картинки, дальнейшего развертывания сюжета и т. п. Так, несомненно, иронической можно считать рекламу зубных палочек для собак. Этот ролик после информации о вреде зубного налета демонстрирует кадры, «рекламирующие» зубные протезы (причем собаки представлены улыбающимися, с человеческими вставными челюстями, что явно относит кадр к соответствующим рекламным сообщениям услуг стоматологов). Данная «реклама» на самом деле рекламой не является, что легко понимают собаководы. В «обратном» смысле здесь понимается важнейшая часть рекламы — предложение товара. Однако это только часть рекламного сообщения. Данный кадр сменяется следующим, где собака, грызя рекламируемый продукт — зубную палочку, ворчит: «Что за ерунда! На самом деле нужна зубная палочка». Т. е. авторы сообщения не оставляют зрителю необходимости понять иронию в предыдущих кадрах, а тут же все объясняют.

Различия в частоте использования иронии в текстах разных типов заставляет предположить, что при восприятии сообщений в различных типах дискурсах предполагается различная активность адресата. И в рекламном дискурсе она одна из самых низких.

Выводы

Таким образом, ирония оказывается не тривиальным, но достаточно широко распространенным способом передачи информации слушающему. При этом от слушающего требуются некоторые дополнительные усилия, иногда значительные, что кажется противоречащим максимам Грайса о дружелюбии, количестве. Однако при использовании иронии говорящий оказывается в состоянии передать дополнительные фрагменты смысла:

А) говорящий вызывает в адресате чувство приобщенности к сокровенному знанию (поскольку понимает сказанное), чувство принадлежности к одному сообществу, что способствует некритичному восприятию сказанного,

Б) передаваемая информация восстанавливается адресатом, что способствует ее восприятию как собственный вывод и в меньшей степени отвергается адресатом,

В) говорящий обеспечивает правильное понимание путем дополнительных маркеров (ироническая интонация, «кавычки» и т. п.) и учитывает представления о мире и возможности адресата провести необходимые процедуры, поэтому в определенных типах коммуникации, где эти условия соблюсти трудно, ирония употребляется редко,

Г) действия по восстановлению иронии обычно бывают эмоционально насыщенными, причем чаще всего переживания определяются юмористической составляющей и игровой — радостью адресата при восстановлении замысла говорящего.

Все перечисленное делает иронию сильным инструментом воздействия говорящего на адресата.

Моделирование действий адресата по пониманию иронии заставляют учитывать такие компоненты смысла сообщения как диктум, модальная рамка, оценка, эмоции, жанровая и стилистическая принадлежность. А в действия приходится включать не только понимание (прямое) сообщения, но и сопоставление с картиной мира, выводы адресата о намерениях говорящего (причем разнородные и неодновременные), а также восприятие оценки и эмоций.

Литература

1. *Баранов А. Н.* Намек как способ косвенной передачи смысла // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог — 2006» / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея — М.: Изд-во РГГУ, 2006 — с. 46–51.
2. *Борисова Е. Г.* Интерактивный подход в лингвистике: пределы применимости // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог — 2006» / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея — М.: Изд-во РГГУ, 2006 — с. 84–88.
3. *Зализняк А. А.* Семантика кавычек// Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог — 2007» / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея — М.: Изд-во РГГУ, 2007 — с. 188–193.
4. *Как провести лингвистическую экспертизу спорного текста?* Памятка для судей, юристов СМИ, адвокатов, прокуроров, следователей, дознавателей и экспертов / Под ред. проф. М. В. Горбаневского. — 2-е изд., испр. и доп. — М.: Юридический Мир, 2006. — 112 с.
5. *Кобозева И. М., Лауфер Н. И.* Об одном способе косвенного информирования // Известия АН СССР. Сер. лит. и яз. 1988. Т. 47, № 5. С. 462–470.
6. *Походня, С. И.* Языковые виды и средства реализации иронии / С. И. Походня — Киев: Наук. думка, 1989. — 126 с.
7. *Санников В. З.* Русская языковая шутка: От Пушкина до наших дней / В. З. Санников. — М.: Аграф, 2003. — 556 с.
8. *Шатуновский И. Б.* 6 способов косвенного выражения смысла // Труды международной конференции «Диалог — 2004». М., 2004.
9. *Шилихина К. М.* Ирония в политическом диалоге // Политическая лингвистика. № 4 (38) 2011. С. 177–182.
10. *Шилихина К. М.* Ирония как эффект языковой игры //Язык, коммуникация и социальная среда. Выпуск 8. Воронеж: ВГУ, 2010. С. 37–45.

References

1. *Baranov A. N.* (2006) Hint as an Instrument of Inderect Communication [Namëk kak sposob kosvennoj peredachi smysla] in Computational Linguistics and Intellectual Technologies in International Conference "Dialogue 2006" Proceedings. (Bekasovo 31 May — 4 June 2006) — IPPI RAN, 2006 — pp. 46–51.
2. *Borisova, E.* (1998) "Opposition Discourse in Russia: Political Pamphlets 1989–1991." In Political Discourse in Transition in Europe 1989–1991, edited by P. Chilton, M. Ilyin and J. Mey. Amsterdam/ Philadelphia: Jon Benjamins. — pp. 111–130
3. *Borisova, E. G.* (2006) The interactive Approach in Linguistics: Limits of Application [Interaktivnyj podhod v lingvistike: predely primenimosti] in Computational Linguistics and Intellectual Technologies in International Conference "Dialogue 2006" Proceedings . (Bekasovo 31 May — 4 June 2006) — IPPI RAN, 2006 — pp. 84–88.
4. *Borisova E.* (2013) Special Entities Used for Governing the Processes of Understanding in Understanding by communication, eds. E. Borisova, O/Souleimanova. Cambridge Scholars Publishers, — pp. 95–103.
5. *Crystal, D.* (2001) Language Play / D. Crystal. in Chicago: The University of Chicago Press. — 248 p.
6. *Davidov D.* (2010) / Dmitry Davidov, Oren Tsur, and Ari RappoportSemi-supervised recognition of sarcastic sentences in Twitter and Amazon. In Proceeding of ComputationalNatural Language Learning (CoNLL 2010), Uppsala, Sweden, July, 2010.
7. *Gorbanevsky M. V.* (2006) (ed). How the Linguistic Examination of a Text is to be Held? For lawyers, judges, experts [Kak provesti lingvisticheskuju ekspertizu spornogo teksta? Dl'a sudej, advokatov, prokurorov, sledovatelej, doznavatelej I ekspertov] — Juridicheskij mir. — 112 pp.
8. *Hutcheon, L.* (1995) Irony's Edge. The Theory and Politics of Irony / L. Hutcheon. — New York: Routledge, 1995. — 248 p.
9. *Kobozeva I. M., Laufer N. I.* (1986) One Way of Indirect Informing [Ob odnom sposobe kosvennogo informirovanija] in Izvestija AN SSSR v. 47 (5) — pp. 462–470
10. *Pirogova Yulia* (2013). Message perception and comprehension in Marketing communication discourse, Understanding by communication, eds. E. Borisova, O. Souleimanova. Cambridge Scholars Publishers, — pp. 176–210.
11. *Pohodnja S. I.* (1989) Language Means of Implementation of Irony [Jazykovye sredstva realizatsii ironii] Kiev, Naukova dumka.
12. *Sannikov V. Z.* (2003) Russian Language Joke.From Pushkin till Nowadays [Russkaja jazykovaja shutka. Ot Pushkina do nashih dnej] Moscow, Agraf.
13. *Shatunovskij I. B.* (2004) 6 Ways of the Inderect Expression of Sense [6 sposobov kosvennogo vyrazhenija smysla], in Computational Linguistics and Intellectual Technologies in International Conference "Dialogue 2004" Proceedings , 2004
14. *Shilikhina K. M.* (2010) Irony as an Effect of the Language Play [Ironija kak effect jazykovoj igry], Language, Communication and Social Environment. Issue 8. Voronezh, pp. 37–45.

15. *Shilikhina K. M.* (2011) Irony in Political Dialogue [Ironija v politicheskom dialoge], Politicheskaja Lingvistika 4 (38). — pp. 177–182
16. *Wilson Deirdre* (2007) On Verbal Irony, Deirdre Wilson, Dan Sperber, Irony in Language and Thought: A Cognitive Science Reader. — New York Lawrence Erlbaum Associates. 2007, — pp. 35–55
17. *Zalizniak Anna. A.* (2007) The Semantics of Inverted Commas [Semantika kavychek]. in Computational Linguistics and Intellectual Technologies in International Conference “Dialogue 2007” Proceedings. (Bekasovo 31 May — 4 June 2007) — IPPI RAN, 2006, pp. 188–193

СЛОВАРНЫЙ ПОДХОД К РАЗРЕШЕНИЮ ОМОНИМИИ ПРИ ВЫДЕЛЕНИИ ИМЕНОВАННЫХ СУЩНОСТЕЙ В РУССКОМ ЯЗЫКЕ

DICTIONARY-BASED AMBIGUITY RESOLUTION IN RUSSIAN NAMED ENTITIES RECOGNITION. A CASE STUDY[1]

Brykina M. M. (m.brykina@gmail.com)
ZAO Eventos, Moscow, Russia;
Lomonosov Moscow State University, Moscow, Russia

Faynveyts A. V. (fainalex@lyandex.com)
Freie Universität Berlin, Berlin, Germany;
ZAO Eventos, Moscow, Russia

Toldova S. Yu. (toldova@yandex.ru)
Lomonosov Moscow State University, Moscow, Russia;
The Center for Semantic Technologies,
Higher School of Economics, Moscow, Russia

The Information Extraction task and the task of Named Entities recognition (NER) in unstructured texts in particular, are essential for modern Mass Media systems. The paper presents a case study of NER system for Russian. The system was built and tested on the Russian news texts. The method of ambiguity resolution under discussion is based on dictionaries and heuristic rules. The dictionary-oriented approach is motivated by the set of strict initial requirements. First, the target set of Named Entities should be extracted with very high precision; second, the system should be easily adapted to a new domain by non-specialists; and third, these updates should result in the same high precision. We focus on the architecture of the dictionaries and on the properties that the dictionaries should have for each class of Named Entities in order to resolve ambiguous situations. The five classes under consideration are Person, Location, Organization, Product and Named Event. The properties and structure of synonyms and context words, expressions and entities necessary for disambiguation are discussed.Key words: Named Entities Recognition, Named Entities ambiguity, Named Entities disambiguation, rule-based approach.

Key words: named entities recognition, named entities disambiguation, dcitionary-based approach

[1] We would like to thank our colleagues Anastasia Bonch-Osmolovskaya, Andrei Idiatullin, Evgeny Fedko, Yulia Zinova, Alexandre Arkhipov, our programmer Petr Zhalybin, and RIA Novosti project manager Philip Dudchuk for their valuable advice, support and collaboration.

1. Introduction

The task of Named Entities recognition (NER) in unstructured texts is essential for modern Mass Media systems. Research in this area has been conducted for more than 20 years (cf. the report on the state of the art in [1], [2], [3], [8]). However the majority of works deal with English or other well-studied European languages. Some systems for Russian are discussed in [4], [5], [6].

There are two basic approaches to the NER task: handmade rule-based systems and machine learning-based systems.In this paper we discuss the architecture of a rule-based system within the task of extraction of a predefined set of Named Entities (NEs). Advantages of the suggested approach include a predictable system behavior, high precision for the user-specified list of NEs, the possibility for the user to update the system as well as to control the effects of NEs database extension. We will mainly focus on the structure of the dictionary where the NEs synonyms are stored.

Since all the entities and their synonyms within the task defined above are enumerated in the dictionary, the focus of the development shifts to the task of ambiguity resolution. There are three commonly known types of ambiguity:

- the ambiguity between a NE and a common noun (cf. *Rubin*, a football club vs. *rubin*, a jewel);
- the ambiguity between NE classes (cf. *Vladimir* as a town vs. *Vladimir* as a personal name);
- the ontology ambiguity between two entities with the same name (*Sergei Ivanov*, a politician vs. *Sergei Ivanov*, a scientist).

Some cases of NE overlapping are not discussed here, for in our case it is the user who should decide how to treat them. E.g. in (1) the user may or may not be interested in extracting Moscow as a Location while it is a part of an Organization.

(1) *Tverskoj sud Moskvy*
Tverskoj Court of Moscow

In the following sections we discuss the ways we deal with above-mentioned ambiguity cases for several classes of NEs. Afterwards we summarize the means used in the dictionaries to resolve ambiguity of a dictionary entry, and finally we present the results of our system's evaluation.

2. PLO and Events Disambiguation

2.1. System overview and task specification

Our objective was to develop a system that extracts NEs of a predefined ontology from news texts. Initial lists of NEs to be recognized were provided by the user (their size is given below); the important requirement was the following: it should the easy to add new NE and new types of NE to these lists by the user.

- Persons ≈ 5000;
- Locations ≈ 10000;
- Organizations ≈ 4000;
- Products ≈ 1000;
- Events ≈ 1000.

The program we developed is based on OntosMiner technology [7] and partially on GATE software (http://gate.ac.uk/). The main function of the OntosMiner processor is extracting semantic information from unstructured texts and presenting in triples of a specific format Turtle (format for expressing data in the RDF data model). Thus it is possible to map the system output onto a user-defined Domain Model stored in ontology (see Fig. 1.).

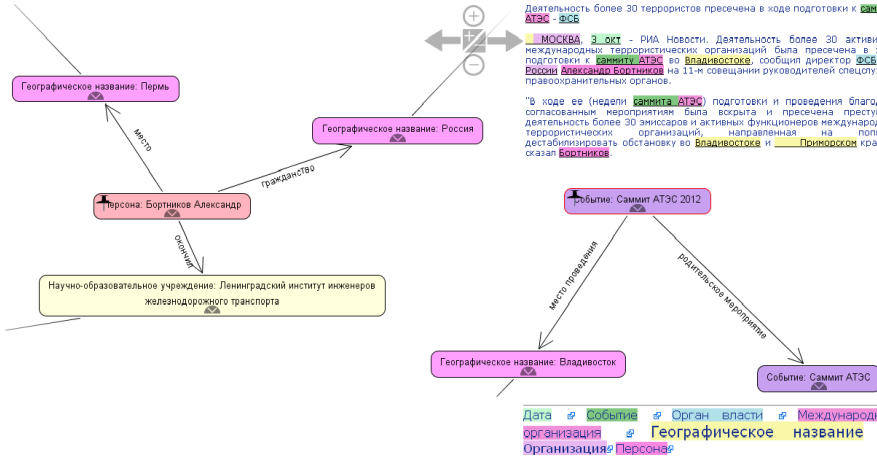


Fig. 1. Visualization of OntosMiner output

Below we discuss the dictionary-oriented NER subsystem of OntosMiner which includes the NE ontology and linked synonyms dictionaries.

One way to present the synonyms for certain NE is to have all the synonymous full noun phrases as entries of a synonym dictionary (further on we would use the term “lookup entry,” or just a “lookup”). Our approach is quite different: we use a minimum common text material for a set of synonyms, and then expand lookup boundaries with certain rules if necessary. For instance, we need the only one dictionary entry Microsoft for three names Microsoft, Microsoft Corp., Microsoft Company etc. This approach requires less human effort to fill dictionaries and also makes it easier to treat conjunction.

In our system we distinguish between unambiguous and ambiguous lookup entries. Unambiguous lookups (=ULs) immediately identify an object, while ambiguous ones (=ALs) need to be verified through the context. Verification is performed by (manually) assigning attributes to lookups that specify the additional information needed to verify an AL (turning it into an UL). These attributes are processed by heuristic-based rules.

Unlike verification, rejection of lookups can be applied both to ULs and ALs. Lookups are rejected when their context changes the class of an object. E.g., location lookups can be rejected when adjoined to currency names (dollar SSHA — US dollar) or other entities class keywords (mul'tfil'm Madagaskar — "Madagascar" film). Company names may be ignored when used in product names (utjug Philips — a Philips iron).

2.2. Location lookups

Geographical names, their variants and abbreviations are the most common synonyms for Locations (Moskovskaja oblast', Podmoskov'e, MO — Moscow Region). According to our principle of minimal lookups we exclude locative keywords from the synonym: Moskovskaja would be the lookup entry for Moskovskaja oblast' (Moscow Region) entity. Instead we use an attribute which signals that a locative keyword is required. The algorithm requires an appropriate keyword to adjoin the lookup or to be the head of a conjunction phrase (v Moskovskoj i Ivanovskoj oblastjah — in Moscow and Ivanovo Regions). Locative keywords are stored in controlled vocabularies (CV) and can have their own synonyms.

2.2.1. Verification attributes

i. Location vs. Location ambiguity

To disambiguate homonymous Locations, such as the eight Soviet Regions (Sovetskij rajon) in Russia, ontological information about the hierarchy of administrative division is required. A specific algorithm searching for parent, ancestor or sibling locations has been developed to resolve this type of ambiguity.

Another case is the ambiguity between different types of Locations with the same name: gorod Tunis (Tunis) vs. strana Tunis (Tunisia). Here an appropriate keyword must be found to verify a lookup.

ii. Location vs. Person (family) names

Some city names are ambiguous with family names (Mogilev — Mogilev City) or even first names (Vladimir — the city of Vladimir). The solution is to have a heuristic module which extracts Person entities including those not present in a given ontology. Afterwards we use a minimization rule: when a Location is embedded into a Person, this Location is removed. It is indeed important to have at least a basic PLO-extracting module (which extracts all, including non-dictionary, mentions of PLO in the text), because one can never put all possible ambiguity cases in the dictionary.

iii. Location vs. common words

To distinguish Locations vs. common words ambiguity (g. Nahodka — lit. (the city of) Find) we use one of the verification attributes mentioned above (either keyword or region verification). Another supporting context for such lookups would be their position in a prepositional phrase (v Nahodke — in Find).

Some ALs, including ambiguous abbreviations, may have no specific clues and can only be verified if a corresponding entity has already been found in the text.

2.2.2. Location attributes: overview

Type of attribute	Scope
Requires a locative keyword	Adjoined
Requires region verification	Document
Requires PP-verification	Left-adjoined
Requires entity to be already present in the text	Document

2.3. Organization lookups

Several types of organizations have been extracted: commercial organizations, government organizations, academic institutions, international organizations, and political parties. The names of organizations, abbreviations, or the shortest names without surrounding quotation marks (both in Cyrillic and Latin alphabet) were used as lookup entries.

2.3.1. Verification attributes

i. Organization vs. other entity ambiguity

First of all punctuation can be very helpful in this case, i. e. a lookup can become unambiguous if it occurs in quotation marks. However, quotation marks can be omitted, that's why punctuation cannot be the only means to rely on.Left-adjoined word or phrase can be useful, either the name of the organization type (e. g. factory, company, foundation) or other keyword(s) (director, CEO, chief accountant). These keywords are organized as in which can be expanded by the user.

To parse sentences (2b-d) correctly one needs to mark a lookup entry for the organization whose name is ambiguous like in (2a). If one of the distinguishing criteria occurs, the correspondence with a proper entity is not called in question any more.

(2) a.

Lookup entry	Attribute	Entity
Mir dereva Mir dereva (lit.: the world of wood)	*needs quotation marks; needs left-adjoined word (sequence) from the list*	*Derevoobrabatyvajushchaja kompanija «Mir dereva»* Woodworking company "Mir dereva"

b. *«Mir dereva» javljaetsja krupnym proizvoditelem izdelij iz dereva.*
"Mir dereva" is a big manufacture for wooden products.

c. *Kompanija Mir dereva — odin iz organizatorov ètoj vystavki.*
"Mir dereva" company is one of the organizers of this exhibition.

d. *Tema segodnjashnego zanjatija v detskom sadu «Mir dereva i metalla»*
Theme of the today's lesson in the kindergarten is "The world of wood and metal"

ii. Organization vs. Organization ambiguity

Different companies that have nothing in common can bear the same name. Quotation marks and common words or phrases like “organization” or “CEO” in left context are helpless here. For the cases such as (3b)–(3c) besides the left context we should take into account the industry to which a corresponding NE refer. We need to create two lookup entries for two different Rubin organizations and specify context words needed for each of them.

(3) a.

Lookup entry	**Attribute**	**Entity**
Rubin Rubin (lit.: ruby)	needs left-adjoined word from the list for the specific industry	*futbol’nyj klub «Rubin»* FC Rubin
Rubin Rubin (lit.: ruby)	needs left-adjoined word from the list for the specific industry	*Konstruktorskoe bjuro «Rubin»* design office Rubin

b. *FK “Rubin” vozglavljaet turnirnuju tablitsu.*
FC Rubin leads the table.

c. *KB “Rubin” zanimaetsja proektirovaniem.*
Design office Rubin does projecting.

Our CV gets a sort of hierarchical structure: we supplement each entry of a CV either with the name of industry or with a special attribute value which means that it is impossible to determine the industry.

The other case of organization ontological ambiguity is the organizations with the identical names situated in different regions (4a, b). To be able to link such ALs with proper entities we need to add an attribute pointing to the corresponding region. The attribute value is implemented as a link to the object dictionary entry for the location in question.

Once such a lookup is found, the nearest mention of location in this document is being looked for. If no appropriate location is mentioned in the document, the default location — Russian Federation — is set up, since we deal with Russian news agencies’ texts. Thus both in (4a) and in (4c) the Ministry of Foreign Affairs of Russian Federation should be recognized, although there is no explicit mention of Russian Federation in (4c).

(4) a. *Ministerstvo inostrannyh del RF otvetilo na pros’bu Gennadija Onishchenko*
Ministry of Foreign Affairs of Russian Federation answered to the request of Gennadij Onishchenko.

b. *Vladimir Putin posetil Parizh i provel vstrechu s glavoj MIDa.*
Vladimir Putin visited Paris and hold a meeting with the head of the Ministry of Foreign Affairs.

c. *Ministerstvo inostrannyh del otvetilo na pros’bu Gennadija Onishchenko*
Ministry of Foreign Affairs answered to the request of Gennadij Onishchenko.

2.3.2. Types of controlled vocabulary entries

Controlled vocabularies for keywords use two obligatory attributes for each CV entry: type and industry. Entries of different type are processed differently:

- A *prefix* ("organization", "joint-stock company", "factory") designates the following AL as an organization. A *general prefix* ("company", "society") doesn't imply a specific industry, while a *specific prefix* ("bank") does.
- A *keyword* ("director", "CEO") means that the AL is *probably* an organization.
- A *key adjective* ("cosmic", "aircraft") forms a specific prefix when combined with a general prefix.
- A *postfix* ("Limited", "& Co") designates the previous AL as an organization.

2.3.3. Organization attributes: overview

Type of attribute	Scope
Requires a specific location	Document
Requires a general left-adjoined keyword	Left-adjoined
Requires a specific left-adjoined keyword	Left-adjoined
Requires quotation marks	Surrounding

2.4. Person lookups

In Russian news texts a person is usually introduced by calling him/her with first name, family name, and optionally patronymic. Further mentions use a family name only. We normally use family names as lookup entries. An obligatory check for these lookups is name verification: somewhere in the text the family name must be used adjoined to appropriate first name(s) or initials. Exceptions are allowed for a short list of very famous people (e. g. presidents) who can be referred to by family names alone. Other lookups for Persons which do not require name-verification can be:

- nickname or stage name: *Boris Akunin (Boris Akunin), Vitas (Vitas)*;
- first name(s) plus family name if there are several variants of name components combinations (for example, Arab person names);
- person name plus person's status-role: *Koroleva Elizaveta (Queen Elisabeth II)*;
- person name written in English or her/his native language, which can be useful if Russian transcription is not consistent.

2.4.1. Verification attributes

i. Person vs. Location ambiguity

Proper names designating first names, and more often family names can also stand for Locations: *Anton Chekhov* vs. *Chekhov* (city); *Lion Izmajlov* vs. *Lion* (city). As mentioned, this ambiguity is resolved by means of minimization: when a possible Location is embedded into Person, this Location is removed.

ii. Person vs. Person ambiguity

Ambiguous Person names are not often found in top-news texts in Russian, though the more widespread name a Person has, the more often we expect to find his/her namesake. Wikipedia lists 25 persons for *Sergej Ivanov*, most of which can be expected in news texts (though only one of them is being mentioned constantly through the past years, and one more was several years ago). To disambiguate homonymous person names the following ontology information is used:

- (a) profession, academic rank, title (some constant status of a Person);
- (b) place of employment (this can be an Organization or a Location);
- (c) position at the place of employment.

We use CVs for (a) and (c), so that synonyms of one concept form a group, and each of the words could be used as verification marker. The search scope is the paragraph containing the AL.

iii. Person vs. other entity ambiguity

We don't use specific attributes to distinguish between a Person and another entity, although family names can be ambiguous with common words. It seems sufficient to have one verified mention of a person in the text to consider all the other occurrences verified (except for when there is more than one person with the same family name).

2.4.2. Person attributes: overview

Type of attribute	Scope
Requires name verification	Adjoined
Requires title/employment/position verification	Paragraph

2.5. Event lookups

Some cultural, governmental and other actions are described as Events (exhibitions, film festivals, messages of the President to the Federal Assembly, United Nations summit). Shortest versions of names, types of events without quotes are used as lookup entries.

2.5.1. Verification attributes

i. Event vs. another entity ambiguity

Event names are not always unique and could mean something totally different in other contexts. Quotation marks are not regularly used with Event names (even more rarely than with organizations), that's why the only reliable clue is a left-adjoined word or phrase of a proper class.

(5) a.

Lookup entry	Attributes	Entity
Zolotoj lev	needs left-adjoined word (sequence) from the list for the theater festival	Theater festival "Zolotoj Lev" (lit.: Golden lion)

b. *Ivan Ivanov — glavnyj redaktor zhurnala «Zolotoj Lev».*
Ivan Ivanov is the chief editor of the "Zolotoj Lev" (lit.: Golden lion) magazine.

c. *Teatral'nyj festival' «Zolotoj lev» proshel v Permi.*
Theater festival "Zolotoj Lev" took place in Perm.

d. *Priz kinofestivalja v Venetsii — statuètka zolotogo l'va.*
The prize of the film festival in Venice is a golden lion statuette.

Among the entities named Zolotoj Lev in (5), only in (5c). the Event should be recognized. To ensure this, the lookup entry Zolotoj Lev needs an attribute saying that an appropriate keyword is required (5a). Possible keywords are also organized in a CV.

ii. Parent Event vs. Child Event

Many Events like festivals are periodical. For instance, in (6) we would like to see three different events. In (6a), a class of periodical song contests called Eurovision (parent Event); in (6b) and (6c), specific child Events should be recognized.

(6) a. *Evrovidenie — vazhnoe sobytie v mire pop-muzyki.*
Eurovision is an important event in pop music world.

b. Poslednee Evrovidenie proshlo v Moskve.
The last Eurovision took place in Moscow.

c. Na Evrovidenii v 2009 godu pobedil Dima Bilan.
Dima Bilan won Eurovision contest in 2009.

A periodical Event like in (7a) must be associated with a unique place, number or year to be recognized as a specific child Event:

(7) a. *Evrovidenie v Baku* *OR* *Evrovidenie-2012*
Eurovision in Baku OR Eurovision-2012

b.

Lookup entry	Attributes	Entity
Evrovidenie Eurovision	needs a time-marker: year (2012) OR needs a place-marker (Baku)	*Evrovidenie-2012* Eurovision-2012

For the cases like in (7), the system should be aware that: (i) both events took place once in 2012; (ii) they are the same; (iii) they did not take place in 2011 or in 2010. Thus, lookup entries must have an attribute requiring a mention of a year,

a number or a place. In (7b) it is essential to point out that we need a time-marker OR a place-marker: both at once are also helpful but not necessary.

For some other types of Events, attributes should rather be combined with AND. Consider (8):

(8) a. *avtosalon v Toronto*
automobile show in Toronto

b. *avtosalon, otkryvshijsja vchera v Toronto*
automobile show (that) opened yesterday in Toronto

c.

Lookup entry	Attribute	(Hypothetical) proper entity
Avtosalon automobile show	needs a place-marker (Toronto)	*Avtosalon v Toronto* Automobile show in Toronto

The events in (8) must both correspond to the same Event. It may be impossible to estimate all the language constructions that describe the connection between this type of events — automobile show — and its location — Toronto. Instead we use an attribute signaling that a place-marker is needed (8c).

If such an event is periodical, a time-marker (number or year) is also to be found. Thus, to determine a correspondence with a proper entity both types of markers (time AND place) should be present in the document (9).

(9)

Lookup entry	Attributes	(Hypothetical) proper entity
avtosalon automobile show	needs a time-marker: year (2011) AND needs a place-marker (Toronto)	*Avtosalon v Toronto 2011* Automobile show in Toronto — 2011
avtosalon automobile show	needs a time-marker: year (2012) AND needs a place-marker (Toronto)	*Avtosalon v Toronto 2012* Automobile show in Toronto — 2012

2.5.2. Event attributes: overview

Type of attribute	Scope
Needs a place-marker	Sentence
Needs a time-marker	Sentence
Needs a left-adjoined keyword	left-adjoined
Needs both place and time markers	Sentence
Needs either place or time marker	Sentence

3. Unified Template for an Arbitrary Object Type (NE)

All the attribute systems described in previous sections were designed for specific object types. But the list of object types was not fixed and could be modified by the user. We have then created a unified template of preset attributes through which ALs can be marked and then processed by disambiguating rules.

3.1. Components of a unified attributes template:

i. User can operate both substrings and ontology entities. E.g. if a mention of Russian Federation in the context is needed to disambiguate an AL, the user can just give a link to the corresponding entity. For example, if we want to extract Russian Football National Team, we can create an AL *national team*, requiring *Russian Federation* or its synonyms to verify this entry. An entity from any dictionary (also CV) can be chosen. An ontological feature that contains a link to an object and requires verification can also be used.

ii. Key-words/entities or stop-words/entities can be set. This means that user can define not only obligatory context but also a context that blocks the association of an AL to an entity.

iii. Users can choose the scope of each key- or stop-parameter: immediate left or right context, sentence, paragraph, or document.

iv. Key- and stop-parameters can be combined with each other with OR or AND operators.

v. Users can set whether quotation marks can be used or are required to identify an object.

vi. Users can add ALs that are verified only if a corresponding entity has already been found in the text

This unified template was used to make a lookup dictionary for Product entities. The results can be seen in *Table 1*.

4. Evaluation and Discussion

The general assessment of the system is based on a procedure similar to one described in the MUC[2] conferences. A random set of 300 texts from RIA Novosti news agency was chosen as test corpus. It was manually annotated in GATE[3] by two persons[4], the cases of annotators' disagreement were further revised. In the evaluation procedure only entities from the Ontology were taken into consideration. The results are shown in Table 1.

---

[2] http://www-nlpir.nist.gov/related_projects/muc/

[3] http://gate.ac.uk/

[4] We are grateful to our colleagues E. Tarasov and I. Novikov who have helped us to annotate the corpus.

Table 1. Evaluation results

NE	Number of cases	Recall	Precision	F-measure
Location	2,270	0.98	0.99	0.98
Organization	1,654	0.93	0.95	0.94
Person	453	0.94	0.99	0.96
Event	55	0.98	0.85	0.91
Product	138	0.96	0.98	0.97

Three notes are in order concerning the results above:

- we only provide counts for NPs that correspond to the NE from the Domain ontology (we've also used heuristic mechanisms to extract NEs, but these are not counted);
- the quality of the system depends crucially on the quality of manually filled dictionaries, and thus these results cannot be reproduced unless based on the same dictionaries;
- the system gives its user control over any particular object; the metric based on the quality evaluation for each ontology entity is out of discussion in present paper.

The evaluation however demonstrates that the lightweight system based on user dictionaries and rather simple rules could be quite helpful if one's goal is extracting a limited number of NEs, and that the cases of ontology ambiguity are not too frequent in general news texts to influence significantly the performance of the system based on predefined high-quality NEs dictionaries.

5. Conclusion

We have presented a system for NEs extraction and disambiguation based upon manually created dictionaries[5]. Ambiguous lookups can be assigned multiple attributes depending on the contextual information needed for disambiguation. Verification is performed by heuristics-based rules. Thus, our system doesn't require train corpus, it has predictable behavior, and is extensible to other types of objects. The user has access to every object's properties and a possibility to use a synonymous set of keywords.

We have discussed clues that help disambiguate NEs in Russian news texts, such as:

- text properties (quotation marks);
- ontological properties:
- object class hierarchy;
- location associated with an object;
- class-specific features (status-role of a Person, organization industry etc.)[6].

[5] Automatic synonyms detection and attributes filling would be more appropriate, but has not been implemented yet.

[6] The frequency of different ambiguity situations and the impact of attributes on the final result need further quantitative analysis and evaluation methods.

Clues may have different scope, e. g. a sentence, a paragraph, or even the whole text (with special search algorithm). In some cases adposition of a clue is required.

We argue that for such systems:

- it is possible to suggest a unified template allowing to add attributes that serve to disambiguate different types of objects (other than PLO);
- along with dictionary-based extraction, it can be useful to have a guessing-module to help resolve some types of ambiguity and thus ascribe less attributes to lookup entries;
- there can be specific rules based on metatextual information, which recover implicit attributes (for example, Russia can be set as default location for Russian news agencies texts).

References

1. *Antonov E.* (2012) Models of automatic disambiguation of ontologic homonymy. [Modeli avtomaticheskogo razreshenija ontologicheskoj omonimii], PhD thesis, manuscript.
2. *Bunescu R. and Paska M.* (2006), Using Encyclopedic Knowledge for Named Entity Disambiguation, Proceedings of of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), Trento, Italy: Association for Computational Linguistics, pp. 9–16.
3. *Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol M., Taneva B., Thater S., Weikum, G.* (2011), Robust Disambiguation of Named Entities in Text, Proceedings of Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 782–792.
4. *Kiselev S., Ermakov A., Pleshko V.* (2004), A search of facts in texts on natural language on the base of net descriptions, [Poisk faktov v tekste estestvennogo jazyka na osnove setevyh opisanij], Computational linguistics and intellectual technologies ("Dialog-2004"). Proceedings of the International Conference [Kompjuternaja lingvistika i intellektual'nye tehnologii ("Dialog-2004"). Trudy Mezhdunarodnoj Konferentsii]. Zvenigorod, pp. 282–285.
5. *Krejdlin L.* (2006), TagLite: Programma vydelenija russkih individualizirovannyh imennyh grupp TagLite [TagLite: The Program of Identification of Russian Individualized NPs]// Kompjuternaja lingvistika i intellektual'nye tehnologii ("Dialog-2006"). Trudy Mezhdunarodnoj Konferentsii [Computational linguistics and intellectual technologies ("Dialog-2006"). Proceedings of the International Conference]. Zvenigorod, pp. 292–297.
6. *Kuznetsov I.* (2012), The methods of discovery of objects and their links presented implicitly in text. [Metodiki vyjavlienija objektov i zvjazej, zadannyh v nejavnom vide], available at: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/%D0%9A%D1%83%D0%B7%D0%BD%D0%B5%D1%86%D0%BE%D0%B2_%D0%98_%D0%9F.pdf

7. *Minor S. Starostin. A.* (2007), Ontos: Technology of knowledge extraction from unstructured texts and semantic indexing, [Ontos: Tehnologija izvlechenija znanij iz nestrukturirovannyh tekstov i semanticheskoe indeksirovanie], Computational linguistics and intellectual technologies ("Dialog-2007"). Proceedings of the International Conference [Kompjuternaja lingvistika i intellektual'nye tehnologii ("Dialog-2007"). Trudy Mezhdunarodnoj Konferentsii],. Bekasovo.
8. *Nadeau D. and Sekine S.* A survey of named entity recognition and classification, Linguisticae Investigationes, Amsterdam, Netherlands: John Benjamins Publishing Company, 1: Vol. 30. pp. 326.

АВТОМАТИЧЕСКОЕ ДОСТРАИВАНИЕ ТАКСОНОМИИ НА РУССКОМ ЯЗЫКЕ НА ОСНОВЕ РЕСУРСОВ ВИКИПЕДИИ

Черняк Е. Л. (echernyak@hse.ru),
Миркин Б. Г. (bmirkin@hse.ru)

Отделение прикладной математики и информатики,
НИУ Высшая Школа Экономики, Москва, Россия

Ключевые слова: достраивание таксономии, близость между строкой и текстом, использование Википедии, суффиксные деревья

COMPUTATIONAL REFINING OF A RUSSIAN-LANGUAGE TAXONOMY USING WIKIPEDIA

Chernyak E. L. (echernyak@hse.ru),
Mirkin B. G. (bmirkin@hse.ru)

Department of Applied Mathematics and Informatics, National Research University, Higher School of Economics, Moscow, Russia

A two-step approach to devising a hierarchical taxonomy of a domain is outlined. As the first step, a coarse “high-rank” taxonomy frame is built manually using the materials of the government and other representative sites. As the second step, the frame is refined topic-by-topic using the Russian Wikipedia category tree and articles filtered of “noise”. A topic-to-text similarity score, based on annotated suffix trees, is used throughout. The method consists of three main stages: 1) clearing Wikipedia data of noise, such as irrelevant articles and categories; 2) refining the taxonomy frame with the remaining relevant Wikipedia categories and articles; 3) extracting key words and phrases from Wikipedia articles. Also, a set of so-called descriptors is assigned to every leaf; these are phrases explaining aspects of the leaf topic. In contrast to many existing taxonomies, our resulting taxonomy is balanced so that all the branches are of similar depths and similar numbers of leaves. The method is illustrated by its application to a mathematics domain, “Probability theory and mathematical statistics”.

Keywords: taxonomy refinement, string-to-text similarity, utilizing Wikipedia, suffix trees

1. Introduction

Taxonomy, or hierarchical ontology, is a popular computational instrument for representation, maintaining and usage of domain knowledge [10, 13]. A taxonomy is a rooted tree formalizing a hierarchy of subjects in an applied domain. Such a tree corresponds to a generalizing relation between the subjects such as "B is part of A" or "A is more general than B". Automating the process of taxonomy building is important for further progress of computational text processing and information retrieval [14, 17]. The mainstream work for advancing into the problem assumes usage of a large collection of unstructured texts related to the domain. These are used to find a set of keywords/keyphrases with a clear cut relation of "inheritance" between them so that the set of keywords and the relation are output as the taxonomy that has been looked for. Drawbacks of this approach are well-known: (a) not every domain can be supplied with a representative large corpus of unstructured text documents, and (b) methods for finding semantic relations between words are not that perfect currently so that both the vocabulary and structure of a found taxonomy are less than satisfactory, as a rule [9]. Therefore, the idea of using an Internet resource, such as Wikipedia, instead seems quite natural [5]. Moreover, one should expect that Wikipedia would supply the taxonomist with a set of subjects and a hierarchic relation over them because of its very nature. Yet one cannot expect that the subjects and the hierarchy can be transferred for the task as easy as it seems to be. The issue is that Wikipedia writers are more enthusiastic than professional. Therefore, one should expect that either the set of subjects or the hierarchy or even some articles or all of those — no one can say what — may be flawed.

In the remainder, we describe a semi-automatic method for deriving a domain taxonomy in two steps. First step, manually building a "coarse", top level, taxonomy, usually of one or two layers only, by taking them from the official documents and definitions.

Second step is of step-by-step refining the taxonomy topics by adding fragments of the Russian Wikipedia category tree, and articles in the categories, both pre-filtered of "noise items". A topic-to-text similarity score, based on annotated suffix trees, is used throughout. Our method for refining of a taxonomy leaf, after the relevant materials from Wikipedia are downloaded, involves removing "irrelevant" subjects and articles. The method is illustrated by its application to a mathematics domain, "Probability theory and mathematical statistics" (in Russian), which highlights both advantages and drawbacks of the method.

This application is relevant to our work on using taxonomies for computational visualization and interpretation of published paper abstracts and university course syllabuses in the field of applied mathematics and informatics. In Russian, the only publicly available taxonomy of Mathematics and related areas is the classification for the government-sponsored Abstracting Journal of Mathematics [15] developed in 1999. This is somewhat outdated and unbalanced. Fortunately, in Russia one can find a live and frequently updated classification of sciences maintained by the High Attestation Committee (HAC) of Russia supervising the national system of PhD and ScD theses [7]. It is not quite deep; it covers just two layers of the body of science. Two or three more layers can be derived from the so-called HAC specialty passports

available for each of the classification leaves. Yet all these layers are of rather coarse granularity. To reach the base granularity concepts, such as the concept of derivative in mathematics, one needs two to four layers of more and more refined concepts.

This specifies the problem. We need a method to refine a coarse taxonomy by using Wikipedia (ru.wikipedia.org). The method should allow us to produce a more or less balanced tree structure. One more requirement to the refinement method is that every refined leaf in its output is to be assigned with a number of keywords or key phrases clarifying the contents of the corresponding concept. Such is the ACM Computing Classification System [1], one of the most advanced domain taxonomies, so that we refer to the required balance properties and clarifying labels as the ACM CCS gold standard.

The problem of refinement of a taxonomy has received some attention in the literature. A big question arising before starting any refinement steps is about the sources for generating new topics. Usually the results of a search engine query, such as "A consists of...", where A is an existing taxonomy topic, are analyzed [16]. Such a query would lead to a set of concepts that can be considered as potential subtopics for topic A. This works especially easy if the ontology is represented by means of a formal language, such as OWL, by introducing new logical relations [4]. On the whole, not only fully unstructured sources like collections or corpora of text may work well in this situation, but also sources such as other taxonomies or ontologies can be used. Another approach, becoming much popular, is using the Wikipedia as a major source of new topics [12,16,18]. Wikipedia offers a lot of data types, such as unstructured texts, images, the category trees, revision history, redirect pages and covers many specific knowledge domains. Reference [5] lists these advantages of using Wikipedia in taxonomy building:

- Wikipedia is consistently updated, thus Wikipedia-based taxonomies can be easily maintained.
- Wikipedia is multilingual, so any method developed for one language can be extended to another.

In papers [12,16,18] different approaches for constructing or refining ontologies and taxonomies by using Wikipedia article data are presented. In [12] the Wikipedia articles, in [20], the Wikipedia category tree, and in [18], the Wikipedia infoboxes, are utilized. Our approach to refining taxonomies is somewhat different. We extract topics both from the Wikipedia category tree and from the articles, and moreover, we score the extent of relevance of those to the parental category. This allows us to follow the ACM-CCS gold standard of taxonomy. By restricting the domain of the taxonomy to smaller topics such as the probability theory and mathematical statistics, we avoid the issue of big Wikipedia data and, also, get the possibility to manually examine the results.

2. Our approach to taxonomy refinement using Wikipedia

We specify the taxonomy frame manually by extracting basic topics from the publicly available instruction materials of the Higher Attestation Commission of Russia [7]. The HAC materials are reflected in a three-level rooted tree of the main topics of probability theory and mathematical statistics (see Table 1).

Table 1. HAC based "Probability theory and mathematical statistics" taxonomy frame

Probability theory and mathematical statistics		
1	**Probability theory**	
	1.01	Models and characteristics of random events
	1.02	Probability distributions and limit theorems
	1.03	Combinatory and geometrical probability problems
	1.04	Random processes and fields
	1.05	Optimization and algorithmic probability problems
2	**Mathematical statistics**	
	2.01	Methods of statistical analysis and inference
	2.02	Statistical estimators and estimating parameters
	2.03	Test statistics and statistical hypothesis testing
	2.04	Time series and random processes
	2.05	Machine learning
	2.06	Multivariate statistics and data analysis

We use the corresponding Wikipedia category, that is, "The Probability Theory and Mathematical Statistics", as the only source for new topics. Luckily, the topic of our interest is a category in Wikipedia, so there is no need to address any other categories. For our purposes, it is useful to distinguish between two Wikipedia data types:

1. The hierarchical structure of Wikipedia category tree
2. The collection of unstructured Wikipedia articles.

Hereafter we are going to use the Wikipedia category tree for extending our taxonomy tree, whereas the articles are used as the source of keywords. We try to assign Wikipedia categories and the underlying subcategories to every taxonomy topic of the first and second levels. First, we find those Wikipedia categories that correspond to our taxonomy topics — they should be subdivisions of the topics. Each subdivision is further divided according to the Wikipedia articles in that, so that the titles of the articles are leaves of the final taxonomy tree. Then, we extract keywords representing the content of each Wikipedia article. These keywords are used then as leaf descriptors.

Therefore, each topic is refined in a two-level subtree, which consists of a Wikipedia category and corresponding Wikipedia articles.

Unfortunately, the structure of the Russian Wikipedia categories is rather noisy. Some categories semantically have nothing to do with their parental categories. For example, in the Russian Wikipedia category tree, the Optimization category lies under the Machine learning category, which itself falls in the Mathematical Statistics category (accessed December 2012). Moreover, the category tree in some places loses its tree-like format and gets cycles within it. One of the explanations of this phenomenon is given in [8]: Wikipedia users' passion to category assignment.

To be used for taxonomy refining, the relevant part of the category tree should be first cleared from all irrelevant subcategories and articles: the clearing action appears to be necessary for obtaining meaningful results.

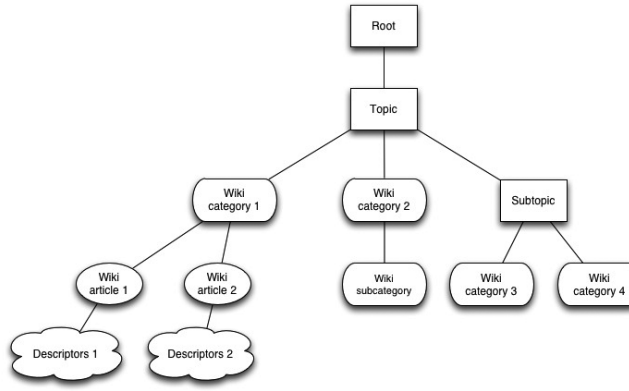


Fig. 1. Our refining scheme: Initial taxonomy topics are in rectangles, the Wikipedia categories and subcategories are in rounded rectangles, the Wikipedia articles are in the ellipses, and the leaf descriptors are in the clouds

Here are the main steps of our approach to taxonomy refining:

1. Specify a taxonomy topic to be refined.
2. Download the related Wikipedia category subtree and the articles from the taxonomy topic under consideration.
3. Clear the category subtree of irrelevant articles.
4. Clear the category subtree of irrelevant subcategories.
5. Extend the taxonomy tree in the specified topic node with the cleared Wikipedia subtree.
6. Put Wikipedia articles in each added category node as the leaves.
7. Extract keywords from Wikipedia articles and use them as leaf descriptors.

Let us illustrate these steps using the following manually made example.

2.1. Specifying taxonomy topic:

Probability Theory and Mathematical Statistics (PTMS).

2.2. Downloading the contents of the topic according to Wikipedia

Download from Wikipedia the subtree of concepts rooted at PTMS. There were 640 Wikipedia articles assigned in 48 categories.

2.3. Clearing the category tree of irrelevant articles

Some of the nodes in the downloaded tree have obviously nothing to do with the PTMS subjects, such as “Software optimization” and “Natural language toolkit”. We consider that an article is irrelevant if the similarity between the parent category title ant the text of the article is low; setting a threshold value is described in section 3. The similarity value follows from the annotated suffix tree (described later) and ranges from 0 to 1. It expresses the average level of conditional probability of a symbol in a string to appear after the string’s prefix.

2.4. Clearing the category tree of irrelevant subcategories

We declare that a subcategory is irrelevant if the similarity between its parent category title and the text obtained by merging all the articles in the subcategory is low; setting a threshold value is described in section 3. Unfortunately, this approach may fail sometimes. For example, the Decision Tree subcategory is irrelevant to the Machine Learning according to our rule, which is obviously wrong. The cause: none of the four articles in the category Decision Tree contain phrase "Machine Learning" or any of its substrings.

2.5. Extending the taxonomy tree by Wikipedia categories

After clearing the category tree from irrelevant categories and articles, we assign each of the remaining Wikipedia categories to a corresponding topic in the current fragment of taxonomy using, again, the AST similarity between the taxonomy topics and the categories represented by all their articles merged.

2.6. Putting Wikipedia articles as the taxonomy leaves to the taxonomy tree

If a Wikipedia category is assigned to a taxonomy topic, all the articles left in it after clearing procedures are put as new leaves descending from the topic.

2.7. Extracting keywords from Wikipedia articles and using them as descriptors to leaves

A leaf taxonomy topic can be assigned with a set of phrases falling in it, as is the case of ACM-CCS. To extract keywords and key-phrases, we don't employ any sophisticated techniques and take the most frequent nouns and the most frequent collocations, respectively. Of course, a key phrase is looked for as a grammar pattern, such as adjective + noun or noun + noun.

3. AST method

The suffix tree is a data structure used for storing of and searching for symbolic strings and their fragments [6]. In a sense, the suffix tree model is an alternative to the Vector Space Model (VSM), arguably the most popular model for text representation [19]. When the suffix tree representation is used, the text is considered as a set of strings, that is, any semantically significant parts of text, like a word, a phrase or even a whole sentence.

An annotated suffix tree (AST) is a suffix tree whose nodes (not edges!) are annotated by the frequencies of the strings fragments. An algorithm for the construction and the usage of AST for spam-filtering is described in [11], and some other applications — in [2, 3].

In our computations, we consider a Wikipedia article to be a set of three-word strings. The titles of the Wikipedia categories and articles are also considered as strings in the set. To estimate the similarity between a standalone string and a collection of strings, we build an AST for the set of strings and then find all the matches

between the AST and fragments of the given string. For every match we compute the score as the average frequency of a symbol in it related to the frequency of its prefix. Then the total score is calculated as the average score of all the matches. Obviously, the final value has a flavor of the conditional probability and lies between 0 and 1. In contrast to similarity measures used in [2,3,11], this one has a natural interpretation and, moreover, does not depend on the text length explicitly, and, as our experiments show, implicitly. To specify an "irrelevance" threshold for the similarity between a category and a text, we take the threshold of 0.2, which amounts to 1/3 of the maximum similarity value and, in our experiments, works well.

4. Results

For the taxonomy in Table 1 the resulting taxonomy tree has 7 levels, with its depth varying from 4 to 7. A fragment of the tree is presented on Figure 2. At the clearing steps a hundred irrelevant articles and two irrelevant categories were removed from the Wikipedia category subtree. Some of the taxonomy topics remain untouched as, for example, "Methods for Statistical Analysis and Inference".

There is a problem with the obtained taxonomy tree: the position of the topic "Decision Trees". According to our method, this topic should be placed under "Multivariate Statistics and Data Analysis" and be, thus, a sibling of the "Machine Learning" topic. Moreover, as mentioned above, the "Decision Trees" has a very low similarity to "Machine Learning".

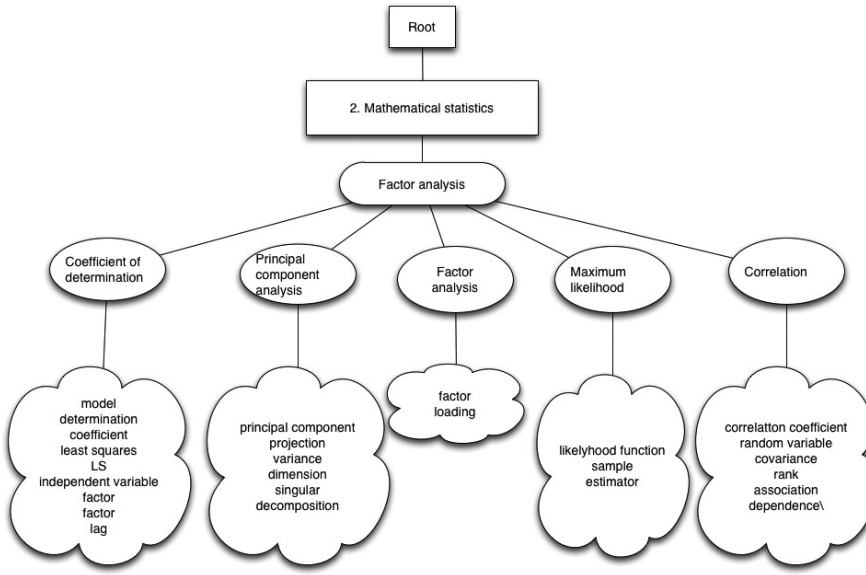


Fig. 2. A fragment of the refined taxonomy tree: the "Factor Analysis" branch

To refine a taxonomy at a given topic, the AST method works three times:

1. Clear the Wikipedia category subtree of irrelevant articles;
2. Clear the category subtree of irrelevant categories;
3. Relate taxonomy topics to Wikipedia categories.

5. Conclusion

The approach of automated refinement is part of a two-step approach to taxonomy building. First step: an expert sets a frame of the taxonomy. Second step: this frame is refined topic-by-topic until an appropriate level of granularity is reached. This approach allows protecting the taxonomy being built from noise, such as irrelevant or too detailed topics. Wikipedia is a good source for new taxonomy topics, because it contains both structured (the category tree) and unstructured (articles) data.

The presented implementation of the approach, by using an AST based similarity estimates, bears both positive and negative effects. The positive relates to the independence on the language and its grammar; and the negative, with the lack of tools for capturing synonymy and near-synonymy. This method is of little help when there is no word by word coincidence, which should be one of the main subjects for the further developments.

References

1. *ACM* Computing Classification System (ACM CCS), (1998), available at: http://www.acm.org/about/class/ccs98-html
2. *Chernyak E. L., Chugunova O. N., Mirkin B. G.* (2012), Annotated suffix tree method for measuring degree of string to text belongingness [Metod annotirovannogo suffiksnogo dereva dlja otsenki stepeni vhozhdenija strok v tekstovie dokument], Biznes-Informatika [Business Informatics], no.3, pp. 31–41.
3. *Chernyak E. L., Chugunova O. N., Askarova J. A., Nascimento S., Mirkin B. G.* Abstracting concepts from text documents by using an ontology. Proceedings of the 1st International Workshop on Concept Discovery in Unstructured Data. Moscow, 2011, pp. 21–31.
4. *Grau B. C., Parsia B., Sirin E.* Working with Multiple Ontologies on the Semantic Web. In Proceedings of the 3d International Semantic Web Conference, Hiroshima, Japan 2004, pp. 620–634.
5. *Grineva M., Grinev M., Lizorkin D.* (2009), Text documents analysis for thematically grouped key terms extraction [Analis tekstovih dokumentov dlja isvlechenenija tematicheski sgruppirovannih kljuchevih terminov], in Trudy Instituta sistemnogo programmirovanija RAN [Works of Institute for System Programming of the RAS], Institute for System Programming, pp. 155–156.
6. *Gusfield D.* (1997), Algorithms on Strings, Trees, and Sequences, Cambridge University Press.
7. *Higher Attestation Commission of RF Reference*, (2009), available at: http://vak.ed.gov.ru/ru/help_desk/

8. *Kittur A., Chi E. H., Suh B.* What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, USA, 2009, pp. 1509–1512.
9. *Liu X., Song, Y., Liu S., Wang H.* Automatic Taxonomy Construction from Keywords. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, New York, 2012, pp. 1433–1441.
10. *Loukachevitch N. V.* (2011), Tezaurusy v zadachah informatsionnogo poiska [Thesauri in information retrieval tasks], MSU, Moscow.
11. *Pampapathi R., Mirkin B., Levene M.* (2006), A suffix tree approach to anti-spam email filtering, Machine Learning, Vol. 65(1), pp. 309–338.
12. *Ponzetto S. P., Strube M.* Deriving a Large Scale Taxonomy from Wikipedia. In Proceedings of AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2007, pp. 78–85.
13. *Robinson P. N., Bauer, S.* (2011), Introduction to Bio-Ontologies, Chapman & Hall/CRC, USA.
14. *Sadikov E., Madhavan J., Wang L., Halevy A. Y.* Clustering query refinements by user intent. In Proceedings of the 19th International Conference on World Wide Web, New York, USA, 2008, pp. 841–850.
15. *Taxonomy of Abstracting Journal "Mathematics"* (1999), VINITI. Available at: http://www.viniti.ru/russian/math/files/271.htm
16. *Van Hage W. R., Katrenko S., Schreiber G.* A Method to Combine Linguistic Ontology-Mapping Techniques. In Proceedings of 4th International Semantic Web Conference, 2005, Galway, Ireland, pp. 34–39.
17. *White R. W., Bennett P. N., Dumais S. T.* Predicting short-term interests using activity-based search contexts. In Proceedings of 19th ACM conference on Information and Knowledge Management, Toronto, Canada, 2010, pp. 1009–1018.
18. *Wu F., Weld D.* Automatically refining Wikipedia Infobox Ontology. In Proceedings of the 17th International World Wide Web Conference, Beijing, China, 2008, pp. 635–645.
19. *Zamir O., Etzioni O.* Web document clustering: A feasibility demonstration. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, USA, 1998, pp. 46–54.
20. *Zirn C., Nastase V., Struve M.* Distinguishing between Instances and Classes in the Wikipedia Taxonomy. In Proceedings of 5th European Semantic Web Conference, Tenerife, Spain, 2008, pp. 376–387.

РУССКИЙ ЯЗЫК В ДАГЕСТАНЕ: ПРОБЛЕМЫ ЯЗЫКОВОЙ ИНТЕРФЕРЕНЦИИ[1]

Даниэль М. А. (misha.daniel@gmail.com),
Добрушина Н. Р. (nina.dobrushina@gmail.com)

НИУ ВШЭ, Москва, Россия

Дагестан представляет собой территорию, где стремительно растет число носителей русского языка. В последние десятилетия русский язык занял здесь место, которое до того времени было вакантно — он стал лингва франка всего Дагестана. При этом, поскольку русский язык усваивается местными жителями в интенсивном контакте с местными языками и одновременно все больше и больше, в своем локальном варианте, приобретает (само)идентификационные функции, он обладает многими характерными особенностями на уровне фонетики, морфологии, синтаксиса и лексики. Цель настоящей статьи — обсудив социолингвистические и языковые особенности сельского варианта русского языка в Дагестане, предположительно разделить их на те, которые мотивированы интерференцией, и те, которые мотивированы феноменом *imperfect learning*, а также, в заключении, обосновать необходимость создания корпуса дагестанского русского как исследовательского инструмента.

Ключевые слова: языковая интерференция, русский язык, языки Дагестана, региональные языковые варианты, разработка корпусных ресурсов

A CORPUS OF RUSSIAN AS L2: THE CASE OF DAGHESTAN[2]

Daniel M. A. (misha.daniel@gmail.com),
Dobrushina N. R. (nina.dobrushina@gmail.com)

NRU HSE / MSU, Moscow, Russia

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2013 году.

[2] This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE).

In Daghestan, the number of Russian speakers has been dramatically increasing over the last few decades. Russian has assumed the functional niche previously vacant in this extremely multilingual setting, becoming the first ever lingua franca of the region as a whole. Russian is acquired in a situation of strong interaction with local languages and shows contact properties on various linguistic levels: phonetics, morphology, syntax and lexicon. Its regional variant is also visibly developing as a self-identification device. The aim of this paper to discuss some (socio)linguistic properties of this idiom, attribute them either to interference or to imperfect learning, and to argue for building a corpus of Daghestanian Russian.

Keywords: language interference, imperfect learning, Russian as L2, East Caucasian, regional variants, corpus development

Роль и усвоение русского языка в Дагестане

В Дагестане более сорока языков (точное число зависит от выделения различных языков в диалектно-языковых кластерах, в первую очередь в даргинском кластере), и дистанция между этими языками очень велика. Хотя языки семьи обладают чертами нетривиального структурного сходства, даже те из них, которые являются близкородственными, такие как лезгинский и табасаранский, практически полностью исключают возможность взаимопонимания. Глубина родства языков этой семьи сопоставима с глубиной родства индоевропейских языков.

Есть данные о том, что в 19 веке в разных регионах Дагестана в качестве лингва франка функционировали разные языки. Роль языков межэтнической коммуникации часто выполняли тюркские языки — ногайский, кумыкский, азербайджанский, носители которых обитали на нижних территориях Дагестана (эти языки, например, названы в Chirikba 2008: 74). Исследователями описана модель вертикального билингвизма, при которой языки распространяются снизу вверх: жители верхних, высокогорных сел владеют языками нижних, а нижние языками верхних не владеют (Wixman 1980, Волкова 1974, Nichols forthcoming). У этого явления есть много причин, в их числе основной вектор передвижений (жители верхних сел спускались по разным хозяйственным нуждам вниз, но не наоборот), более высокий социальный статус земледельцев по сравнению со скотоводами (Карпов, Капустина 2012; в верхнем Дагестане занимались почти исключительно овцеводством), взаимная близость этих тюркских языков, которая позволяет тому, кто знает кумыкский, понимать азербайджанский и ногайский, и наоборот.

В некоторых регионах роль лингва франка выполнял аварский язык. Кроме того, на территории Дагестана функционировал арабский язык, который занимал место письменного литературного языка. По мнению Вихмана (Wixman 1980: 115), в 19 веке в Дагестане преподавание в школах происходило только на арабском, он же был письменным языком гражданских служб. Арабский был языком интеллигенции вплоть до революции, и первые лидеры Коммунистической партии Дагестана, принадлежавшие к числу наиболее образованных местных людей, тоже владели арабским.

Как бы то ни было, до середины 20 века в Дагестане не было языка, который объединял бы всю его территорию. Это исключительное место сумел занять русский язык. Процесс русификации Дагестана, по-видимому, начинается с 19 века, когда в 1820 году в результате военных действий Дагестан был присоединен к России. Есть данные, что в 1897 году на территории Дагестана было 6,3% русских, а к 1913 году их численность увеличилась в 2,2 раза (Ибрагимов 1978, цит. по Магомедханов 2008). Каковы каналы усвоения русского языка в сегодняшнем Дагестане? Ситуация в горных селах и в городах принципиально различна.

В горах Дагестана этнических русских практически нет. В высокогорных селах, в которых нам пришлось побывать в последние годы, мы встретили русскую жительницу лишь однажды: это была очень пожилая учительница, которая в 50-е годы приехала по распределению в Дагестан работать в школе, вышла замуж и осталась жить в селе (Чародинский район). Присутствие в сельских школах русских учителей было более характерно для 50–70-х годов (см. об этом Добрушина 2008), но часто речь шла об одном учителе на селение.

По данным опроса, произведенного в селениях Арчиб, Читтаб и Шалиб в 2008–2010 годах, среди людей, родившихся до 1919 года, некоторая степень владения русским языком была характерна для 18% населения Арчиба, 32% населения Читтаба и 40% населения Шалиба. Важно отметить, что у нас нет никаких инструментов для того, чтобы определить уровень владения: вполне вероятно, что во многих случаях речь идет о способности объясниться на бытовом уровне. К тому же эти данные получены от родственников, которые помнили своих бабушек и дедушек уже в преклонном возрасте, то есть факт владения русским языком теми, кто родился после 1919 года, относится к периоду 60–70 годов. На протяжении всего 20-го века знание русского языка постоянно росло, так что в настоящее время по-русски говорят практически все, за исключением инвалидов, никогда не выезжавших из села, и очень пожилых людей. На рисунке 1 представлена динамика владения русским языком в высокогорном селении Арчиб.

В 30–50 годы основным источником знания русского языка была школа, общение с русскими учителями и выезды из села на русскоговорящие территории. Сегодня важным каналом является телевидение, которое присутствует везде с 80–90-х годов, причем популярными каналами являются общероссийские. Практика в разговорном русском возникает в селе летом, когда на каникулы приезжают дети, которые живут в городах. Полевые исследования говорят о том, что во многих высокогорных селах дети совсем не говорят по-русски до школы, а полноценное, свободное владение приходит уже после школы, когда молодежь начинает получать образование вне села, юноши попадают в армию или начинают выезжать на заработки.

Как мы видим, единственным источником русского языка в том варианте, в каком он используется в тех регионах, где русский язык является первым или единственным, оказывается телевизор. Все остальные источники русского языка связаны с местными носителями, для которых русский язык часто представляет собой L2 (второй язык).

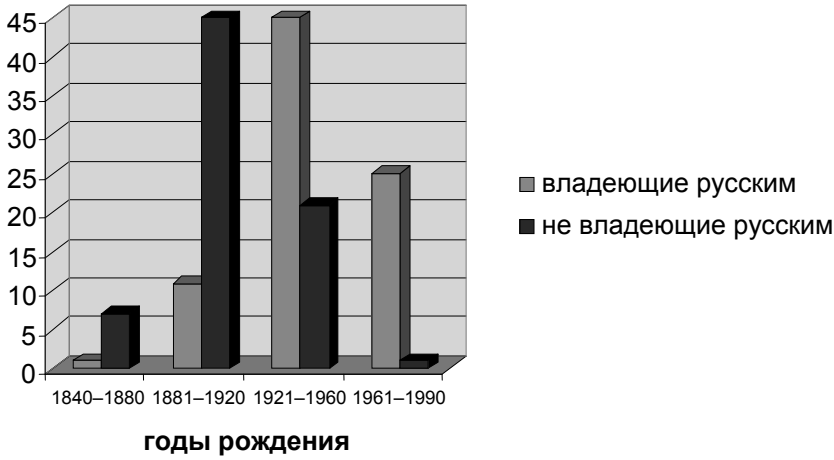


Рис. 1. Владение русским языком в дагестанском селении Арчиб по результатам опроса

В городах и, вероятно, в низинных селах (о них у нас меньше информации) ситуация иная. Русский язык получил исключительное распространение в городской жизни, поскольку города населены представителями всех дагестанских этносов, и русский функционирует как язык межэтнического общения. Дети, которые сегодня рождаются в городах, как правило, не владеют никакими этническими языками. Достаточно распространена также ситуация ограниченного владения этническим языком как L2, когда, например, ребенок общается по-аварски или по-лакски с бабушкой или по крайней мере понимает обращенную к нему речь; слышит его от односельчан, когда приезжает в село на каникулы, но сам говорить на языке не может.

Каковы каналы усвоения русского языка в дагестанских городах? Несмотря на его распространенность, главная особенность сохраняется: в окружении городского ребенка практически нет источников русского языка в его исходном варианте. Городской ребенок усваивает русский язык от родителей, которые зачастую переехали в города из сел и являются носителями русского как L2 (процесс миграции сверху вниз сегодня исключительно активен). Наблюдения показывают, что такие родители предпочитают говорить с ребенком на русском, считая этот язык более полезным для будущей жизни. Родители могут быть и городскими жителями по рождению, и тогда русский язык является для них L1. Русский язык является языком общения в детском саду и в школе и предметом школьного обучения. При этом этнических русских в городах Дагестана немного, по официальной статистике — 4,7% в 2002, 3,6% в 2010.

Таким образом, в городах единственным источником русского языка в том варианте, в каком он используется на русскоговорящих территориях России, вновь оказывается телевизор. Разница с селами состоит в том, что в окружении ребенка значительно больше таких людей, для которых русский является родным языком, и в том, что вокруг себя он слышит преимущественно русскую речь.

Итак, специфика сегодняшнего Дагестана состоит в том, что он является территорией стремительно растущего русскоязычного населения при почти полном отсутствии этнических русских. Неудивительно, что в этой ситуации формируется специфический вариант русского языка, обладающий определенными отличительными чертами. Можно было бы обозначить этот вариант термином *этнолект*. Согласно одному из наиболее популярных определений этнолекта, это "varieties of a language that mark speakers as members of ethnic groups who originally used another language or distinctive variety" (Clyne 2000:291). Применение данного определения осложнено тем фактом, что русский язык Дагестана формируется на почве многих различных этнических групп. Но поскольку механизмы, вероятно, отчасти схожи с теми, о которых пишет Клайн, мы условно примем этот термин для нашей ситуации.

Важным компонентом определения этнолекта является то, что он служит инструментом для создания групповой идентичности. Мы не можем однозначно ответить на вопрос о том, служит ли дагестанский русский для целей идентификации, поскольку не проводили никаких систематических исследований такого рода, хотя отдельные указания на это имеются (юмористические передачи дагестанцев и о дагестанцах, обсуждение слов «дагестанского языка» на форумах, составляемые дагестанцами «словарики» локального варианта и проч.). Выскажем некоторые предположения о механизмах этого процесса.

Стремительная урбанизация Дагестана и утрата знания местных языков создает препятствия для конструирования этнической идентификации у молодых городских дагестанцев: они лишены этнического языка как основной опоры этничности. В этой ситуации русский вариант Дагестана оказывается полезным субститутом. Ему соответствует этничность, которая на данный момент служит маркером «внешней» идентификации: вне Дагестана его жители называют себя дагестанцами (в то время как внутри Дагестана имеют хождение такие маркеры, как аварцы, даргинцы, лезгины и так далее). Такая этничность не имеет и не может иметь собственного языка, который отличал бы ее от других групп. Представляется, что это место может занять региональный вариант русского.

Как уже было сказано, трудность с определением дагестанского русского языка как этнолекта связана с тем, что он существует во многих вариантах. В самом деле, русский вариант Дагестана развивается в тесном контакте с местными языками. Этих языков около сорока. Значит, русская речь носителей каждого из этих языков, особенно в селах, может иметь какие-то особые черты, обусловленные именно этим языком. Это предположение необязательно верно: языки нахско-дагестанской семьи обладают значительным структурным сходством, так что не исключено и то, что больших расхождений между разными вариантами русского нет. К тому же тот русский язык, который функционирует в городе, вполне вероятно, представляет собой наддиалектный вариант — результат выравнивания сельских вариантов. Тем не менее исключить такую внутреннюю неоднородность регионального русского невозможно.

Для того чтобы перевести эти предположения в разряд утверждений, необходимы обширные исследования дагестанского русского языка в его сельских и городских вариантах. В этой статье, также как и в нескольких предшествующих

публикациях (Даниэль, Добрушина 2010; Daniel, Dobrushina, Kniazev 2011), мы сосредоточились на первой задаче. В отличие от предыдущих публикаций, здесь мы не только представляем типы нестандартных грамматических явлений в собранных нами текстах, но и предпринимаем попытку их классификации как вызванных одной из двух возможных причин: интерференции с первым языком или так называемого «неполного выучивания» (imperfect learning).

Грамматические особенности русского языка сельчан

Языковой материал, обсуждаемый в настоящей статье, опирается на доступные нам данные дагестанского русского языка в трех селениях: Арчиб, Читтаб и Шалиб. Эти три селения находятся на расстоянии пешего хода (примерно полтора часа от одного до другого), а их жители активно контактируют друг с другом. В каждом из селений свой родной язык (арчинский, аварский и лакский соответственно; все три языка принадлежат к разным ветвям нахско-дагестанской семьи). Традиционно для этих сел был характерен некоторый уровень знания языка соседей. В настоящее время есть тенденция к тому, что языком коммуникации с соседями становится русский, во всяком случае для молодежи.

Лексические заимствования, во многих отношениях представляющие собой интерес, ниже не рассматриваются; разбор нескольких ярких случаев (см. в Daniel, Dobrushina, Kniazev 2011). Там же дается более подробный обзор типов замеченных нами грамматических и синтаксических особенностей. В настоящей краткой заметке мы хотели бы акцентировать внимание на том, что наблюдаемые особенности могут проистекать из двух причин, или, точнее, иметь две составляющие: интерференция русского языка с первым языком говорящего либо эффект «неполного выучивания» (imperfect learning) второго языка. Эти составляющие могут комбинироваться, но часто они относительно четко различимы. В случае интерференции речь идет о проекции грамматических особенностей родного языка говорящего на структуры русского языка как L2. В случае «неполного выучивания» языковая система L2 оказывается «упрощена» вне зависимости от грамматических особенностей первого языка говорящего. Наконец, третья ситуация в каком-то смысле промежуточна — imperfect learning поддерживается отсутствием в родном языке того или иного грамматического явления, характерного для русского языка.

Идея неполного выучивания отнюдь не предполагает, что говорящий испытывает трудности при использовании второго языка. Еще раз подчеркнем, что подавляющее большинство живущих в селе (не говоря о городе) дагестанцев, с которыми мы встречаемся, говорят по-русски так же свободно, как и любой носитель русского языка как первого (и единственного). «Ошибки» у говорящих на таком варианте языка обычно интерпретируются как указание на то, что литературный язык просто не является для них целевой системой (target language) в процессе овладения языком. Некоторые исследователи полагают, что феномен неполного выучивания занимает центральное место в истории многих языков.

Сразу же можно отметить, что грамматические и иные особенности, наблюдаемые в наших текстах, конечно, никогда не являются абсолютно последовательными. Нестабилизированные языковые варианты (подобно диалектам, существующим в контакте с литературным языком, или креольским языкам, погруженным в посткреольский континуум) характеризуются более или менее заметными колебаниями между двумя языковыми системами — стандартным вариантом носителей языка как L1 и структурными элементами первого языка носителей этого же языка как L2. Примеры таких колебаний иногда приводятся ниже вместе с примерами нестандартных структур.

Итак, мы предлагаем различать интерференцию как «положительную» проекцию элементов языковой системы L1 на L1 (interference); неполное выучивание как элиминирование элементов языковой системы L2, отсутствующих в L1 (imperfect learning ~ interference); и неполное выучивание как результат выравнивания языковых особенностей L2 вне зависимости от структуры L1 (imperfect learning). Приведем наиболее яркие примеры для каждого из типов.

Interference: эвиденциальное *оказывается*

Ярким примером интерференции является чрезвычайно характерное для многих носителей употребление формы *оказывается*:

(1) Да. Этот самый — он умер покойный — участник войны был Мусаев Магомед. Хороший, грамотный. Хорошо он знал русский язык. Он **оказывается** в детстве ходил туда в лакское селение, дома строить, помощником. Каменщика помощником. А оттудова он убежал, попал среди русских, и он говорил, что у него два сына в Волгограде. И связи не было. Такие тоже случались. Дальше… наши люди знали и русский язык. Вот наша соседка рассказывала, она младше матери была, **оказывается**, во время гражданской войны, восемнадцатом-двадцатых годах пришли солдаты до Кубатль. И там они начали сыпать зерно лошадям. А эта женщина, **оказывается,** знала русский язык. И начала она стыдить. Люди здесь голодают, а вы зерно лошадям даете. Она поругала их, и они говорят перестали зерно брать. Это уже женщина, не мужчина. Я думаю, их наверное послали в Сибирь или что-то наверно такое было. Тоже, значит, русский язык знали люди. А азербайджанский почти второй человек знал у нас.

(2) Кочубей это черный рынок называли раньше. **Оказывается** на Кавказе был рынок в этом месте. Там продавали девушек. и поэтому говорят называли черным рынком его. такое место **оказывается** было.

(3) [А сюда кто-нибудь приходил учиться?] Приходили. Вот к этому Мамма дибиру **оказывается** приходили. Даже он предчувствовал, что умрет. У него детей не было. Он сказал хозяйке — я завтра утром умру. А похоронить

приедут люди из Советского района. Шамилевский район. В самом деле. Утром **оказывается** он умер. Вышла эта хозяйка сказать — а оттудова пришли вот эти аварцы.

Очевидна своеобразность такого употребления. В литературном русском языке это вводное слово используется для введения информации, которую говорящий представляет как неожиданную для себя (так называемый *миратив*). В приведенном же отрывке рассказа это значение очевидно отсутствует, и для носителя литературного русского языка такие употребления оказываются немотивированными.

Такое употребление является результатом интерференции с L1. В нахско-дагестанских языках вообще и в арчинском, аварском и лакском языках, в частности, присутствует категория эвиденциальности (иначе косвенной засвидетельствованности или заглазности). Эвиденциальный показатель (или конструкция) указывают, что говорящий сам не является свидетелем передаваемых им событий. По крайней мере в арчинском языке эта категория является строго обязательной (Кибрик 1977). Говоря по-русски, наши собеседники испытывают давление L1 и подбирают средство выражения в качестве функционального заместителя эвиденциальности. Интересно отметить, что выбор их, по-видимому, не случаен: категория миратива, на первый взгляд логически независимая от эвиденциальности, обнаруживает типологическую связь с последней (Плунгян 2011).

Interference: *опущение предлогов*

Несколько менее очевидный случай, но не менее яркая черта речи наших собеседников — это отсутствие предлогов.

(4) Курбамагомед он же живёт/ он **Качалибе** живёт
(Качалиб — название населенного пункта)

(5) Там много топономии вот/ топономики/ эт самое/ названий мест/ упомянуто вот **этом письме завещании**.

(6) И// до этого конечно/ до этого надо идти **нашему устару**// Наш устар/ вот шейх Мамадибир/ ээ у него надо как бы разрешение брать// Ну// Наставник/ учитель как бы религиозный/ **вот/ я ему тоже пошёл**// в этом году/ двадцать четвёртого// марта/ этот день я помню. [устар — мастер, наставник]

Мы предполагаем, что эта черта является проявлением также языковой интерференции. Для языков Дагестана предлоги вообще не характерны (они используют послелоги); кроме того, очень большой инвентарь пространственных значений выражается собственно именным склонением. В зависимости от конкретного языка, более или менее многочисленные (иногда до нескольких

десятков) комбинации пространственных значений, такие как ‘на верхнюю поверхность X-а’ или ‘домой к X-у’, могут выражаться морфологическими средствами. Говоря по-русски, под давлением структуры L1 наши собеседники иногда опускают чуждое для нее аналитическое препозитивное служебное слово. Такое опущение не является последовательным; так, в том же интервью наш собеседник употребляет предлог в очень похожем контексте.

(7) … брать разрешение/ что я вот так хочу/ с вашего разрешения/ вот так хочу/ пойти **к другому устару**/ который щас живёт/ жив/ вот// Вот **к нему** надо идти/ сделать товбу

Конечно, и в данном, и в других подобных случаях нельзя исключать наличие систематических собственно лингвистических факторов, влияющих на использование или опущение предлога. Судить об этом определенно однако нельзя по причине недостаточности доступного языкового материала.

Imperfect learning ← Interference: *колебания возвратности*

В текстах встречается большое число примеров с употреблением возвратной формы глагола в невозвратном контексте и наоборот; некоторые глагольные формы в литературном языке просто отсутствуют.

(8) Необразованные наши вот арчинки вот **попадаются** в больницу. (А)

(9) Они тоже **гостились** у нас Цурибе. (Ч)

(10) День рождения **справились**, там документы были на окошке. (Ч)

(11) Если я здесь **родила**, отец-мать арчинцы, представляешь, я арчибском языке первее говорила — как же я аварка, скажи? (А)

Здесь мы имеем дело с неполным выучиванием, мотивированным несоответствием между структурами L1 и L2. В русском языке единственный высокопродуктивный морфологический процесс, связанный с изменением аргументной структуры глагола — это понижение переходности, т. е. образование медиопассивных (возвратных) форм. Нахско-дагестанские языки, напротив, принадлежат к транзитивизирующим языкам, то есть языкам, явно предпочитающих повышение переходности — каузативизацию. Понижающие актантные деривации в этих языках либо отсутствуют, либо занимают сугубо периферийную позицию. (О делении языков на повышающие и понижающие переходности см. Nichols et al. 2004.) Таким образом, L1 не представляет никаких функциональных аналогов для данного явления L2, на которые говорящий мог бы опереться при порождении формы, и в использовании возвратных форм наблюдаются сильные колебания.

Imperfect learning: *именительный* вместо *родительного*

Для русского языка характерно употребление родительного падежа вместо именительного и винительного падежей в отрицательных контекстах. В первом из приведенных примерах такая замена в норме обязательна, но отсутствует в записанном примере. Во втором случае она в норме, наоборот, невозможна, но говорящий, по-видимому, интерпретирует отрицательную семантику контекста ('плохо знал' аналогично 'не знал') как триггер такой замены:

(12) Азербайджан ехали люди/ которые// было тяжёлое положение// Здесь у них **поля' не было**

(13) Вот// и он вот так начал// он плохо **знал аварского языка**// Ну/ русского вообще не знал//

Очевидно, что в этом случае говорить об интерференции с L1 не приходится: в дагестанских языках просто нет структур, которые могли бы следовать или не следовать этим правилам. Речь идет, так сказать, о неприменении (или применении в другом классе контекстов) правила, специфичного для грамматической системы L2.

Imperfect learning: *видовые колебания*

В некоторых случаях наблюдаются колебания в приписывании граммемы вида. Для следующего контекста вполне вероятно, что мы имеем дело просто с видовой ошибкой:

(14) А после восемьдесят пятого года/ когда уже религию как бы// не **запретили**/ вот// Щас все/ я же говорю/ все как бы начали/ и молиться/ и соблюдать пост// Особенно молодёжь

Однако в нескольких других примерах, обсуждаемых в (Daniel, Dobrushina, Kniazev 2011), речь идет об использовании совершенного вида для описания предельных ситуаций в узуальных контекстах, где в русском языке обычно происходит замена характерной для предельных ситуаций граммемы совершенного вида на граммему несовершенного вида в значении хабитуалиса; возможно, что и в предыдущем примере использование совершенного вида связано с предельностью глагола. Снова отметим характерную для L2 непоследовательность: глагол *снимать* является, очевидно, предельным, но употреблен в форме узуального несовершенного вида.

(15) Я в начальных классах училась, чуть старшие, кто седьмой, шестой класс учится, вот они вот так вот чухту, каз ходили, в школе не разрешают, они на дороге снимали, **спрятали** там.

Несмотря на существенные отличия видовременной системы нахско-дагестанских языков от славянского вида, эти языки никак нельзя считать «невидовыми» и интерпретировать соответствующие примеры аналогично колебаниям возвратности. Если принять гипотезу о связи таких контекстов с предельностью, приходится признать, что видовые системы L1 и L2 существуют тут вполне независимо; расширение контекстов употребления совершенного вида мотивировано семантическими факторами (предельность глагола).

Эти и другие (см. Даниэль, Добрушина 2010; Daniel, Dobrushina, Kniazev 2011) свойства исследуемых нами вариантов русского языка можно предварительно охарактеризовать с точки зрения их отнесения к интерференции или неполному выучиванию:

	Интерференция	**Неполное выучивание**
оказывается	+	–
нативные модели полисемии	+	–
левое ветвление	+	–
опущение предлогов	+	?
нестандартное управление	+	?
Acc (1 скл.) → Nom	+?	–?
колебания возвратности	+	+
Gen → Nom	–	+
рассогласование по роду	–	+
было как несогласуемая форма	–	+
Ipfv → Pfv	–	+

Перспективы исследования: корпус дагестанского русского языка

Как уже несколько раз говорилось, окончательное решение многих центральных вопросов невозможно до тех пор, пока не будет собрано достаточно большое количество данных. Расшифрованных на сегодняшний день текстов сугубо недостаточно. Их формат не позволяет осуществлять поиск таким образом, чтобы можно было интерпретировать результаты в количественном отношении, для чего необходимы данные о числе и типах не только контекстов, где эти явления встретились, но и контекстов, где они могли бы встретиться, но отсутствуют. Как нам кажется, нам уже известны основные типы явлений, характерные для русского как L2 в сельских районах Дагестана; но у нас твердого представления о степени их распространенности. В некоторых случаях они могут коррелировать с определенными лингвистическими параметрами контекста; и только доступность исчерпывающих подборок примеров позволит подтвердить или опровергнуть предположительное отнесение этих явлений к интерференции и/или к неполному выучиванию.

Решить часть этих проблем призван корпус русского языка в Дагестане. В настоящий момент наши тексты состоят почти исключительно из социолингвистических интервью, которые брались в ряде горных сел. Значительная часть интервью была расшифрована А. Е. Дьячковой в 2011 году в рамках Программы фундаментальных исследований РАН. В настоящий момент расшифровано 17 текстов, общим объемом 7 часов звучания (около 50 тысяч словоупотреблений). Как и приведенные выше примеры, расшифровка записей осуществляется в формате устного подкорпуса Национального корпуса русского языка (www.ruscorpora.ru); снятые в этой публикации ударения в текстах корпуса присутствуют. В течение 2013 года планируется расшифровка еще 5 часов, разметка и вывеска текстов в открытом доступе.

Разметка корпуса должна включать как социолингвистический (пол, возраст, место жительства и первый язык носителя, его лингвистическая биография), так и собственно лингвистический компонент. К последнему относится класс «ошибки» и указание дополнительных сведений, необходимых для удобного и информативного поиска по корпусу. Например, для ошибок согласования и примеров левого ветвления должны указываться характеристики вершины; при опущении предлога должен указываться этот предлог и проч.

Поскольку тексты были записаны в селах, все они представляют русский язык как L2. В таблице приведен список селений, в которых производилась запись (с указанием L1 для каждого селения).

Таблица 1. Список селений

село	**L1**
Арчиб	арчинский
Читтаб	аварский
Шалиб	лакский
Ицари	ицаринский
Мегеб	мегебский
Хив	табасаранский
Маллакент	даргинский
Янгикент	кумыкский

Увеличение объема расшифрованного корпуса позволит выяснить, каков статус описанных явлений в речи отдельных носителей; какие из них являются относительно стабильными, а какие носят скорее характер окказионализмов. Выше мы видели, что один и тот же носитель чаще всего использует и нормативную, и альтернативную конструкции; вероятно, ни один носитель не бывает вполне последовательным в следовании локальному варианту, но приверженность норме разные носители демонстрируют в разной степени.

Не менее интересным является вопрос о социальных параметрах варьирования тех или иных явлений: зависит ли их присутствие в речи и частотность от возраста, образования, пола говорящего? Какие именно явления характерны для всех категорий говорящих, а какие имеют социальную обусловленность?

Важно выяснить, есть ли такие явления в русской речи дагестанцев, которые коррелируют с их родным языком, например, встречаются в речи аварцев, но нехарактерны для носителей даргинского.

Наконец, помимо текстов сельского русского, необходимы записи городской русской речи тех, для кого русский язык является L1. Сопоставление особенностей дагестанского русского как L2 с дагестанским русским как L1 сможет дать представление о механизмах формирования нового варианта русского языка.

Литература

1. *Волкова, Н. Г.* (1974). Этнический состав населения Северного Кавказа в XVII-начало XX века. Москва: Наука.
2. *Даниэль М. А., Добрушина Н. Р.* (2010). Новые русские. *// Вопросы русского языкознания*. Вып. XIII. Фонетика и грамматика: настоящее, прошедшее, будущее. Стр. 141–158.
3. *Карпов, Ю. Ю. & Капустина, Е. Л.* (2011). Горцы после гор. Миграционные процессы в Дагестане в XX — начале XXI века: их социальные и этнокультурные последствия и перспективы. Санкт-Петербург.
4. *Кибрик А. Е.* (1977). Опыт структурного описания арчинского языка. Т. 1–3. Москва: МГУ.
5. *Магомедханов М. М.* (2007). Дагестанцы: вехи этносоциальной истории 19–20 вв. Издательство ДНЦ РАН, Махачкала.
6. *Плунгян В. А.* (2011). Введение в грамматическую семантику: Грамматические значения и грамматические системы языков мира. М.: РГГУ.
7. *Chirikba, Viacheslav A.* (2008). The problem of the Caucasian Sprachbund. In Pieter Muysken (ed.) *From linguistic areas to areal linguistics*. John Benjamins Publishing Company. pp. 25–94.
8. *Clyne, Michael.* (2000). Lingua Franca and ethnolects in Europe and beyond. Sociolinguistica, 14, pp. 83–89
9. *Nichols Johanna, David A. Peterson, Jonathan Barnes.* (2004). Transitivizing and detransitivizing languages. *Linguistic Typology* , vol. 8, no. 2, pp. 149–211.
10. *Daniel, Michael, Nina Dobrushina and Sergei Kniazev.* (2011). Highlander's Russian: Case Study in Bilingualism and Language Interference in Central Daghestan. In: *Instrumentarium of Linguistics: Sociolinguistic Approach to Non-Standard Russian*. Slavica Helsingiensia, 40, pp. 65–93.
11. *Nichols, Johanna,* forthcoming. The vertical archipelago: altitude, typology, and sociolinguistics in mountain languages. In Peter Auer, Martin Hilpert, Anja Stukenbrock and Benedikt Szmrecsanyi, eds., Space in Language and Linguistics. Berlin: Mouton de Gruyter.
12. *Thomason, Sarah, Terrence Kaufman.* (1988). Language contact, creolization, and genetic linguistics. Berkeley, University of California Press.
13. *Wixman, Ronald.* (1980). Language Aspects of Ethnic Patterns and Processes in the North Caucasus. Chicago: University of Chicago.

References

1. *Chirikba, Viacheslav A.* (2008). The problem of the Caucasian Sprachbund. In Pieter Muysken (ed.) *From linguistic areas to areal linguistics*. John Benjamins Publishing Company. pp. 25–94.
2. *Clyne, Michael.* (2000). Lingua Franca and ethnolects in Europe and beyond. Sociolinguistica, 14, pp. 83–89
3. *Nichols, Johanna, David A. Peterson, Jonathan Barnes.* (2004). Transitivizing and detransitivizing languages. *Linguistic Typology* , vol. 8, no. 2, pp. 149–211.
4. *Daniel, Michael, Nina Dobrushina and Sergei Kniazev. (2011). Highlander's Russian: Case Study in Bilingualism and Language Interference in Central Daghestan. In: Instrumentarium of Linguistics: Sociolinguistic Approach to Non-Standard Russian. Slavica Helsingiensia, 40, pp. 65–93.*
5. *Nichols, Johanna, forthcoming. The vertical archipelago: altitude, typology, and sociolinguistics in mountain languages. In Peter Auer, Martin Hilpert, Anja Stukenbrock and Benedikt Szmrecsanyi, eds., Space in Language and Linguistics. Berlin: Mouton de Gruyter.*
6. *Thomason, Sarah, Terrence Kaufman. (1988). Language contact, creolization, and genetic linguistics. Berkeley, University of California Press.*
7. *Wixman, Ronald. (1980). Language Aspects of Ethnic Patterns and Processes in the North Caucasus. Chicago: University of Chicago.*
8. *Volkova, Natalja G.* (1974). *Etnicheskij sostav naselenija Severnogo Kavkaza v XVIII — nachale XX veka*. Moskva: Nauka.
9. *Daniel', Mikhail, Nina Dobrushina.* (2010). Novye russkie. // Voprosy russkogo jazykoznanija. Vyp. XIII. Fonetika i grammatika: nastojashhee, proshedshee, budushhee. Str. 141–158.
10. *Karpov, Juriy Ju., Ekaterina L. Kapustina.* (2011). *Gorcy posle gor. Migracionnye processy v Dagestane v XX—nachale XXI veka: ix sotsial'nye i etnokul'turnye posledstvija i perspektivy*. Sankt-Peterburg: Peterburgskoe Vostokovedenie.
11. *Kibrik, Aleksandr E.* (1977). Opyt strukturnogo opisanija archinskogo jazyka. T. I. Leksika. Fonetika. T. II. Taksonomicheskaja grammatika. Izdatel'stvo Moskovskogo universiteta.
12. *Magomedkhanov, Magomed.* (2007). *Dagestancy: vekhi etnosocial'noj istorii 19–20 vv.* Izdatel'stvo DNC RAN, Makhachkala.
13. *Plungjan V. A.* (2011). Vvedenie v grammaticheskuju semantiku: Grammaticheskie znachenija i grammaticheskie sistemy jazykov mira. M.: RGGU.

СТРУКТУРА ЭМОЦИОНАЛЬНО-ЭКСПРЕССИВНОГО КОМПОНЕНТА В ТЕЗАУРУСЕ РУССКОГО ЯЗЫКА RUSSNET

Дёгтева А. В. (degteva.anna@gmail.com),
Азарова И. В. (ivazarova@gmail.com)

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

В статье рассматривается структура экспрессивно-эмоционального компонента тезауруса RussNet применительно к оценкам, выражаемым частотными и низкочастотными прилагательными в сочетании с семантическими классами существительных. Классы оценочных значений (прагматические, эстетические и морально-нравственные) связаны с объективными свойствами характеризуемого объекта/ситуации или являются выразителем субъективной оценки говорящего.

Ключевые слова: анализ тональности текста, корпус текстов, компьютерный тезаурус, эмоционально-экспрессивная лексика, структурирование значений

EXPRESSIVE COMPONENT STRUCTURING OF THE RUSSIAN THESAURUS RUSSNET

Degteva A. V. (degteva.anna@gmail.com),
Azarova I. V. (ivazarova@gmail.com)

Saint-Petersburg State University, Saint-Petersburg, Russia

The paper deals with the structure of expressive attributive word meanings implemented in the wordnet-type thesaurus for Russian (RussNet). The adjectives involved express the appraisal of objects and situations denoted by nouns, the assessment depending on the intrinsic qualities of objects or rendering the subjective attitude of the speaker.

The research was based on a 21 million word corpus of modern texts.

The sentiment meaning in RussNet is structured according to three parameters: Polarity, Domain, and Objectivity. “Polarity”, the intrinsic parameter of the class, describes a positive or negative sentiment value and its measure. “Domain” represents one of the three most commonly expressed standpoints: pragmatic, moral, and aesthetic, as well as actualization of lexical functions Ver/AntiVer, Pos/AntiPos, and Bon/AntiBon defining semantic interaction with hierarchical groups of nominal meanings (semantic trees and subtrees of the RussNet thesaurus). “Objectivity” describes the assessment source as either being personal or custom, usual or individual for the object described. The parameters listed above are organized into a rather intricate scheme but in practical work its structure can be simplified, Yet, detailed analysis can help structuring fuzzy sentiment expressions and detecting versatile evaluative content.

Keywords: sentiment analyses, text corpus, computer thesaurus, expressive lexicon, word meaning structuring

1. Представление атрибутивных значений в wordnet-словаре

В традиции wordnet-словарей, включая Принстонский WordNet [Miller, 1998], атрибутивные значения, в частности значения прилагательных, представлены парой поляризованных значений (bipolar clusters), к которым присоединяется пучок экспрессивно окрашенных способов обозначения качества, например:

(1) Sense 1
hot (vs. *cold*) “used of physical heat; having a high temperature or causing a sensation of heat”
=> *baking, baking hot*
=> *blistering, blistery* “intensely hot”
=> *boiling, scalding* ...

В проекте кафедры математической лингвистики СПбГУ RussNet [Azarova, 2008] подход к представлению атрибутивный значений для всех основных частей речи несколько иной [Azarova & Sinopalnikova, 2004]: предполагается, что значения прилагательных могут соотносится по гипонимическому типу (подобно существительным), например {*большой1*} ‘значительный по величине, размерам’ и {*высокий1*} ‘имеющий большое протяжение снизу вверх’, или тропонимическому типу (подобно глаголам), например {*высокий1*} и {*рослый1*} ‘высокого роста, крупный’. Во втором случае разница в значении определяется тем, что «родовой» термин может соотносится с разными классами видимых объектов: артефактами и объектами естественной природы, людьми и животными, а «видовое» прилагательное соотносится лишь с подгруппой объектов: людьми и животными.

1.1. Амбивалентность атрибутивных значений

Амбивалентность атрибутивных значений применительно к их значениям [Wetzer, 1996] выражается в том, что они занимают позицию на шкале-континууме «существительное-глагол», что можно связать с когнитивным доминированием [Capell, 1965] «объектов» или «событий» в конкретном языке. Косвенными свидетельствами доминирования являются количество существительных и глаголов, а также сложность системы грамматических категорий. С этой точки зрения, в русском языке представлены два центра доминирования: существительные как «доминирование объектов» (исходя из количественного преобладания) и глаголы как «доминирование событий» (исходя из категориальной сложности). Атрибутивные значения прилагательных, тяготеют к одному и/ или другому типу: прилагательные-существительные, прилагательные-глаголы и прилагательные-(глаголы|существительные).

Опираясь на синтаксическое свойство глаголов (предикацию), можно выделить прилагательные-глаголы, например краткие прилагательные рад, должен и проч. Ср. в тезаурусе WordNet также указываются атрибутивные и предикативные употребления прилагательных: {*burning* (prenominal), *fervent, fervid, fiery*} ‘extremely hot; like fire’; {*afraid* (predicate)} ‘filled with fear or apprehension’, отсутствие пометы указывает на смешанный характер синтаксического поведения.

Типичными прилагательными-существительными являются относительные, они указывают на признак через соотнесение с объектом, и в общем случае синонимичны присубстантивному генитиву: Adj ← N; N → Ngen. Синтаксическим свойством атрибутивных относительных прилагательных является контактность позиции к определяемому существительному, что приводит к образованию устойчивых терминоподобных сочетаний (*большой палец*), которые обладают непроницаемостью (ср. *ударил большой палец* референтно не тождественно *ударил большой кривой палец*).

Помимо приведенных «чистых» типов атрибутивных значений встречаются соотношения свойств обеих групп, поэтому речь идет о континууме. Ближе к прилагательным-глаголам расположены определения качеств, независимых от свойств объекта. Например, словосочетание *синее блюдце* представляет пересечение свойств объекта *блюдце* с качеством цвета *синий*. В словосочетании *большая капля* не представлено независимое свойство *большой*, оно определено относительно исходных параметров объекта *капля*. В словосочетании *бывший учитель* временная характеристика является неотделимой от характеристик объекта, она участвует в идентификации свойств объекта (*некто не является учителем в настоящее время*). В последнем случае, несмотря на невозможность предикативного использования прилагательного оно выполняет функцию сходную с предикатом, указывая на переходное состояние.

Амбивалентность прилагательных выражается семантико-синтаксически благодаря активным валентным свойствам, предопределяющим параметры локального контекста, или пассивным, когда значения предсказываются другими словами. Первое характерно для предикативных слов: например, прилагательное *знакомый* в разных значениях задает рамку: «кто-то_с чем-то» /

«что-то_кому-то», «кто-то_с кем-то»/ «кто-то_кому-то», при этом чаще (2/3 контекстов) употребляется в краткой предикативной форме. К пассивной валентности прилагательных отсылает дефиниция *лицо* (одуш. сущ.) ‘*с определением*. Отдельный человек в обществе; индивидуум (*частное лицо; исторические лица*)’. Как ясно из этого примера, необходима схема субкатегоризации прилагательных, поскольку словосочетания *бледное лицо, бессмысленное лицо* указывает на другое значение ‘Передняя часть головы человека’ и употребление как неодушевленного.

1.2. Компоненты оценочного значения vs. лексические функции

Для изучения особенностей в распределении положительных и отрицательных компонентов оценки мы взяли значения частотных (*хороший, красивый, правильный, удобный, приятный, прекрасный, плохой* и *человеческий и др.)* и менее частотных оценочных прилагательных (*милый, уродливый и некрасивый, неудобный, неправильный, неприятный, дурной* и др.)

Значения этих слов, представленные в Словаре русского языка в четырех томах под редакцией А. П. Евгеньевой (М., 1981) — далее МАС, используются для первоначальной разметки значений контекстов в корпусе современных текстов кафедры математической лингвистики объемом в 21 млн словоупотреблений [Azarova et al., 2003]. Это стандартная методика определения структуры значений для некоторой лексической группы в RussNet [Azarova, 2008], которая позволяет экстраполировать результаты разметки случайной выборочной совокупности контекстов (100–150) на корпус в целом, выявляя таким образом реальную структуру распределения значений в лексико-семантической группе.

В [Fomchenko & Azarova, 2008] была рассмотрена базовая структура значений «оценочных» прилагательных. При их корпусном исследовании было выделено 100 оценочных значений (с общий объемом контекстов 850,31 ipm[1]), которые были объединены в три группы: прагматические, эстетические и морально-нравственные. В ядре группы прагматических значения *хороший*$_1$[2]*, прекрасный*$_1$*, приятный*$_3$*, человеческий*$_3$*, плохой*$_1$, вместе с низкочастотными они встречаются примерно в 60% контекстов корпуса. Эстетическая оценка выражается ядерными значениями *красивый*$_1$*, красивый*$_2$*, прекрасный*$_2$, которые вместе с менее частотными представлены в 30% контекстов. В ядре морально-нравственной оценки значения *добрый*$_1$*, человеческий*$_2$*, хороший*$_2$*, они регулярно выражаются* прилагательными, которые употребляются в других оценочных значениях, что, возможно, свидетельствует о развивающемся характере данной оценки.

---

[1] ipm — instanses per million, имеется в виду число употреблений на миллион словоупотреблений в корпусе.

[2] Номера значений в словаре соответствуют частотности употребления WM в нашем корпусе.

Приведенные оценочные значения прилагательных пересекаются с понятием лексических функций, введенных в работах [Apresjan, 1974; Mel'chuk, 1974]. Лексические функции показывают связанное выражение в поверхностной структуре текста для передачи глубинного смысла. Провести границу между связанным и несвязанным выражением не просто, например, для выражение интенсивности (Magn) состояния удивление в корпусе можно найти определения большое, великое, глубокое, крайнее, неслыханное, искреннее, неподдельное, явное. Все ли определения в равной степени выражают значение этой функции, или только первые пять, а последние 3 выражают значение функции Ver (такое, как следует). Ясно, что разграничение истинных лексических функций и регулярных способов выражения глубинного смысла может происходить на другом уровне, при сопоставлении с сочетаемостью слов данного классов. Например, для отрицательного состояния неприязнь в исследуемом корпусе представлены атрибуты нескрываемая и устойчивая, а для положительного состояния радость: большая, великая, глубокая, глубочайшая, невероятная, неподдельная, огромная, особая, откровенная, подлинная, удивительная. Есть ли необходимость выделять из этих списков только те атрибуты, которые встречаются с определенными словами, или же стоит объединять эти списки применительно к определяемому слову, понимая, что реализация набора определений зависит от частоты упоминания соответствующего состояния. У слова радость с частотой (90,3 ipm) самый большой набор квалификаций, у слова неприязнь (7,1 ipm) — самый маленький, а у слова удивление (33,4 ipm) — промежуточный.

Среди лексических функций к оценочному компоненту имеют непосредственное отношение три: упомянутая **Ver** и **AntiVer**[3] (такой, как надо, как выражение прагматического компонента оценки), **Bon** и **AntiBon** (для выражения общей оценки) и **Pos** и **AntiPos** (положительной оценки стороннего наблюдателя). Как видно из этого перечня список лексических функций и выделенные нами компоненты оценки не конгруэнтны.

В сочетании с оценочными функциями необходимо также учитывать функцию **Magn AniMagn** как выражение интенсивности/неинтенсивности (оценки).

2. Трансформация оценки в синтаксической конструкции

В структуре оценочного компонента центральное место занимает характеристика слов, регулярно передающих субъективно-модальное или оценочное значение. Обычно в группу наиболее важных для данного аспекта значений включают глаголы и прилагательные [Kim & Hovy, 2006] или прилагательные и существительные [Ermakov & Kisilev, 2005]. Указывается на то, что оценочный статус прилагательных взаимодействует со статусом существительного: более типичным является «оценочное согласование», например *замечательный бессребреник, отвратительный маньяк, отвратительный*

[3] Префикс **Anti** передает значение противоположности.

скандал, отвратительное чавканье, чем «рассогласование» — *замечательный рассадник, отвратительное благоразумие*. Однако характер взаимодействия требует более внимательного рассмотрения, поэтому мы рассмотрим классы существительных, которые регулярно сочетаются с частотными оценочными прилагательными и иллюстрируют «сдвиги» оценочного значения. В качестве семантических категорий существительных мы используем компоненты тезауруса RussNet [Azarova et al., 2003]: — семантические деревья, связанные родовидовыми отношениями группы значений (например, *человек, артефакт, группа, пища, животное, растения, абстрактное* и проч.), подструктуры таких деревьев (например, обозначения человека посредством (а) качественной, (б) профессиональной, (в) социальной характеристики человека, или (г) родственных и (д) приязненных отношений). Подструктуры семантических деревьев могут объединяться любым способом, полученные конфигурации будут различаться как по сложности структуры, так и по объему группы.

В первом приближении классы существительных в отношении типов оценочных значений распределяются довольно ясно: прагматическая характеристика применима к обозначению артефактов (тому, что создается человеком для удовлетворения его потребностей), эстетическая — к характеристике внешности людей и вида объектов окружающего мира (и артефактов в том числе?), морально-нравственная — к характеристике социального поведения людей. Очевидно, что характеристика человека может относиться вариативно ко второму или третьему типу.

Частотное прилагательное *хороший* дает эстетическую характеристику человека в предикативной форме: *женщина была хороша*, В отдельных случаях значение дополнительно подтверждено контекстными маркерами: *хороша собой, была внешне хороша*. Синоним эстетической оценки *красивый* ‘Приятный на вид, отличающийся правильностью очертаний, гармонией красок, тонов, линий и т. п.’ в равной степени применим к обозначению людей и других объектов реального мира: *~ женщина/ девушка/ мужчина/ лицо/ глаза/ ноги/ птица/ животное/ город/ машина/ облако…* Эстетическая оценка, выраженная краткой формой хороший может относится к другим живым существам:

(2) *Щенков и правда много… Один особенно хорош…*

Однако, если живое существо может быть использовано в качестве пищи, то сразу же возникает прагматическая интерпретация ‘обладающий положительными качествами, свойствами, вполне отвечающий своему назначению’, в частности, оценка качества пищи подразумевает *вкусный*.

(3) *Нынче утром арапа ихнего в речке поймали. Ну так хорош, так хорош: весь филейный.*

Морально-нравственное оценка задается прилагательным *хороший* в атрибутивной форма. Интересным является «раскрытие» содержания оценки, которое дается в тексте в качестве пояснения.

(4) *Жена у меня, надо сказать, очень хорошая женщина. Она умеет решать за меня разные вопросы.*

(5) *Вот Татьяна Егорова, всем хорошая женщина — чувствительная, бескорыстная, с поэтическими порывами, жаждой прекрасного, острая, взбалмошная…*

Если название человека содержит характеристику родства или статуса, то имеется в виду ‹Примерно, образцово выполняющий свои обязанности, обязательства по отношению к кому-, чему-л.’

(6) *Чувства вины — вот чего, оказывается, не хватало Сабурову, чтобы сделаться хорошим отцом.*

Слова-синонимы для этого значения дают несколько разные ряды сочетаний: *заботливый муж/отец/сын/опекун/руководитель/администратор/хозяин, заботливая дочь/мать/тетя/хозяйка* или *любящий* ***друг****/муж/отец/сын/опекун*, но не *администратор/хозяин*. Выделенное отдельно в WN 2.1 значение (10) *dear, good, near* ‘with or in a close or intimate relationship’: *a good friend* применимо только к обозначению родственников и друзей: *близкий друг/родственник/человек, близкая подруга/родственница*. В этом случае мы сталкиваемся с проблемой специфичности атрибута, когда он используется, во-первых, для фиксации набора объектов/событий, к которым применим данный атрибут; во-вторых, для выделения и толкования словарного значения; в-третьих, для выявления слов-синонимов, замещающих друг друга при обозначении атрибута без явного смещения объема качеств.

Многие артефакты полифункциональны, и сочетание с оценкой *хороший* получает разные интерпретации:

(7) *…воображение поразила не только обстановка этой квартиры с хорошим письменным столом…*

(8) *…единственную проблему составляло — вырубить звезду, размером с хороший письменный стол…*

(9) *…Мы договорились о встрече. Встретились. Был хороший стол.*

Если в (9) изменение значения существительного стол приводит к трансформации простого прагматического значения ‘удобный письменный стол’ в более сложное ‘разнообразное и качественное угощение’ , то в (8) интерпретация ‘большой письменный стол’ связана больше с синтаксической конструкцией.

3. Оценочные значения прилагательных в RussNet

Проведенный анализ сочетаемостных характеристик позволяет заключить, что регулярным является наличие нескольких крупных классов имен, по отношению к которым выражаются те или иные значения, прочие варианты бывают распределены по более мелким фасетам, иногда частично смещенным в сторону фразеологизации (*хороший человек, хороший парень, красивый поступок*).

Значения, представленные в самом прилагательном, взаимодействуют с остальными словами, входящими в группу, например, прилагательное *хороший*$_1$ — пригодный, соответствующий свому назначению, зачастую определяет атрибутивную конструкцию: *хорошая типографская краска, хороший рыбий клей, хорошая карта автомобильных дорог,* из чего можно сделать вывод, что данный компонент оценки «соответствие предполагаемым качествам» требует указания на то, какие именно качества имеются в виду. Также можно отметить, что в данном случае в оценке минимально выражен субъективно-эмоциональный компонент и оценка относится скорее к тем характеристикам, которые можно проверить объективно. То же самое можно сказать про соотношение морально-нравственного оценочного компонента (*хороший человек*) и оценки профессиональных показателей (*хороший хирург*). Эта оценка зависит от характеризуемых свойств объекта, она объективна и соотносима с лексической функцией **Ver (AntiVer)**.

Более узкое по смыслу прилагательное *красивый* определяет в массе довольно узкий круг предметов: строения, архитектурные конструкции и тому подобное (*красивый вид, красивая церковь* — 12% контекстов), объекты живой природы (*красивые птицы, семена*), люди (22%) части тела человека (*красивый нос, красивое лицо* — 12% контекстов), произведения человеческого интеллекта — красивая мелодия, красивое слово. Можно предположить, что «красивым» могут являться только те объекты, от которых ожидается эстетическая форма. Это субъективная оценка говорящего, она соотносится с лексической функцией **Pos** (**AntiPos**).

Красивая физиономия — пример того, как противоречащие друг другу эмоциональные компоненты выражены прилагательным и существительным. При этом прилагательное выражает эстетический компонент, более «объективный» в данном контексте, а существительное отображает эмоциональное отношение говорящего к объекту оценки.

Оценка *красивый* чаще встречается для определения лиц женского пола, при атрибуции к обозначению мужчин часто отделяется другими словами:

(10) *Чем тебе Андрей плох? Красивый, умный, Если уж я собираюсь рожать и воспитывать без мужа, то самец должен быть эталонным. И красивым, и физически мощным, и умным, а может быть, даже и талантливым.*

При этом часто, как в предыдущем примере, противопоставляются душевные качества (или подчеркивается их неактуальность в контексте) и внешность:

красивая самка, красивый партнер. То есть *красивая женщина* — устойчивое непроницаемое сочетание, *красивый мужчина* встречается гораздо реже и, как правило, в «разобранном» виде.

(11) *Герой книги показался мне красивым, видным мужчиной, но, пожалуй, слишком уверенным в себе.*

Можно сказать, что оценочные значения прилагательные выражают, в очень большой степени опираясь на семантику существительных, к которым они относятся и к внелингвистическим сведениям о том, что эти существительные называют. Сочетание прилагательного и существительного при выражении оценочной характеристики говорит либо о том, что определяемый объект соответствует ожидаемым от него свойствам, либо речь идет о некотором несоответствии, даже если сама выражаемая оценка положительная.

4. Структура эмоционально-экспрессивного компонента в тезаурусе русского языка RussNet

В структуре wordnet-словарей нет типового представления оценочных и экспрессивных значений за исключением стандартной стилистической словарной характеристики типа *разг.* или *офиц.* и т. п. Необходимость внедрения такой информации очевидна, поэтому появляются специализированные версии типа SentiWordNet [Esuli, Sebastiani, 2006], в которых существительные и прилагательные WordNet 3.0 снабжаются числовыми индексами положительности Pos(s) и отрицательности Neg(s), на основании которых вычисляется объективная характеристика Obj(s) значения компонента по формуле Obj(s) = 1 − (Pos(s) + Neg(s)). Вычисление индексов, возможно, неоднозначно, в целях корректировки используется «обратная связь», при которой пользователь выставляет собственную оценку компонентов.

В более консервативном варианте предлагаются списки положительных и отрицательных словоформ (positive-words и negative-words), которые содержат слова разных частей речи (включая глаголы), при этом отрицательный перечень примерно в два раза превосходит положительный список и они практически не пересекаются. В разметке НКРЯ используется поля положительной и отрицательной оценки: εϖ — оценка (*толковый, мешковатый*), εϖ:ποσιτ — положительная (*везучий, ладный*), εϖ:νεγ — отрицательная (*продажный, сварливый*) для существительных, прилагательных и наречий.

В тезаурусе RussNet для представления оценочных данных используется поле, отведенное для описания валентностей, которые трактуются гибридно [Azarova et al. 2006]: совмещаются синтаксические, семантические и статистические характеристики реализаций валентности в текстах корпуса. В этой структуре интерпретация оценочных значений задается как пассивная валентность атрибутивного или предикативного типа. В xml-формате представления словаря:

```
<VALENCY _ FRAME active="no" main _ segment="attribute">
```

или

```
<VALENCY _ FRAME active="no" main _ segment="predicate">
```

Семантический класс существительных задается в поле «семантический тип»:

```
<sem _ data TYPE="tree" ID="RUS-nHuman"/>
```

где параметр "type" указывает на тип указателя вершины, значение "tree" обозначает семантическое дерево RussNet; идентификатор ID является ссылкой на вершину дерева существительных «человек».

Собственно оценка представлена тремя независимыми параметрами: полярность, сфера, объективность.

(1) Параметр «Полярность» (*Polarity*) задается двумя атрибутами:
(а) оценкой *eval* со значением *pos* (положительный) и *neg* (отрицательный) и степенью *extent*, которые принимает значение *max* (усилением) или *min* (ослаблением). Атрибут «оценка» является обязательным для оценочных значений. Например, *отличный*$_1$ будет описываться

```
<POLARITY eval="pos" extent="max"/>,
```

а *хороший*$_1$

```
<POLARITY eval=”pos”/>
```

При использовании этих данных в процедурах вычисления тональности текста значению атрибутов может быть приписано числовое значение, например, для *отличный*$_1$ — 1, а для *хороший*$_1$ — 2, но это уже «метрика» пространства тональных значений.

(2) Параметр «Сфера» (*Domain*) предполагает следующие атрибуты:
(а) тип оценочного значения *type* описанных выше видов *pragma* (прагматический), *estetic* (эстетический), *etic* (морально-этический); (б) лексическая функция *ver* с бинарными значениями *yes* и *no*; (в) лексическая функция *pos*; (г) лексическая функция *bon* с аналогичными значениями.

Например, близкие по значению прилагательные *правильный*$_1$ и *удобный*$_1$ “такой, какой нужен” и “такой, которым удобно и легко пользоваться” имеют следующее описание

```
<LITERAL>правильный<SENSE>1</SENSE>...<DOMAIN type="pragma"
  ver="yes"/>
```

```
<LITERAL>удобный<SENSE>1</SENSE>...<DOMAIN type="pragma"
  pos="yes"/>,
```

(3) Параметр «Объективность» (*Objectivity*) задается двумя атрибутами: (а) субъектом-источником оценки *bearer* со значениями *person* (личность) и *custom* (обычай); (б) индивидуальным восприятием *individual* со значениями *peculiar* (индивидуальное свойство) и *usual* (общепринятый).

Перечисленные параметры представляют довольно сложную структуру, однако в практических целях структура может упрощаться посредством выбора одного первого параметра, первого и второго и т. д. Однако при анализе значений они позволяют хотя бы частично структурировать такую сложную материю, как языковая оценка.

Некоторые значения атрибутов могут быть слишком «тонкими» для различения синсетов, в частности приведенный выше пример с прилагательными *правильный*$_1$ и *удобный*$_1$ будут «несинсетообразующими», т. е. могут различаться у компонентов синсета.

Литература

1. *Apresjan Ju. D.* (1974), Lexical semantics: Synonymic linguistic resources [Leksicheskaja semantika: Sinonimicheskie sredstva jazyka], Moskow.
2. *Azarova I.* (2008), RussNet as a Computer Lexicon for Russian // Intelligent Information Systems 2008. ISBN 978-83-60434-44-4, pp. 447–456.
3. *Azarova I., Sinopalnikova A.* (2004), Adjectives in RussNet. Proc. of the 2nd Global WordNet Conference. Brno: Masaryk University, pp. 251–258.
4. *Azarova I. V., Ivanov V. L., Ovchinnikova E. L.* (2006), RussNet Valency Frame Inheritance in Automatic Text Processing [Ispol'zovanie shemy nasledovanija ramok valentnostej v tezauruse RussNet dlja avtomaticheskogo analiza teksta] // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2006" [Komp'juternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2006"], Moscow, pp. 18–25.
5. *Azarova I. V., Mitrofanova O. A., Sinopal'nikova A. A.* (2003), Wordnet-type Computer Thesaurus for Russian [Komp'juternyj tezaurus russkogo jazyka tipa WordNet] // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2003" [Komp'juternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2003"], Moscow, pp. 43–50.
6. *Capell A.* (1965), A Typology of Concept Domination. // Lingua, v. 15, pp. 451–462.
7. *Ermakov A. E., Kisilev S. L.* (2005), The Linguistic Model for Computer Sentiment Analysis of Media Texts [Lingvisticheskaja model' dlja komp'juternogo analiza tonal'nosti publikatsij SMI] // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2005" [Komp'juternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2005"], Moscow.

8. *Esuli A. and Sebastiani F.* (2006) SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining // 5th edition of the International Conference on Language Resources and Evaluation. Genoa, Italy, pp. 417—422.
9. *Fomchenko A. V., Azarova I. V.* (2008), Interaction of pragmatic, aesthetic and moral features in the semantic structure of Russian judgment adjectives [Vzaimodejstvie èsteticheskih, moral'nyh i pragmaticheskih aspektov v semanticheskoj strukture otsenochnyh prilagatel'nyh russkogo jazyka] Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2008" [Komp'juternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2008"], Moscow, pp. 545–550.
10. *Kim S.-M., Hovy E.* (2006), Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text // Proceedings of the ACL/ COLING Workshop on Sentiment and Subjectivity in Text, Sydney, Australia, pp. 1—8.
11. *Kustova G. I.* (2010), Adjectives and Personal Nouns [Prilagatel'nye v sostave nominatsii litsa] // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2010" [Komp'juternaja Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoj Konferentsii «Dialog 2010»], Moscow, pp. 265–272.
12. *Mel'chuk I. A.* (1974), On the Theory of Linguistic Models «Sense <=> Text [Opyt teorii lingvisticheskih modelej "Smysl <=> Tekst"], Moscow.
13. *Miller K. J.* (1998), Modifiers in WordNet // WordNet: An Electronic Lexical Database / Ch. Fellbaum (ed.), MIT Press, pp. 47–67.
14. *Vol'f E. M.* (2002), Functional semantics of appraisal [Funktsional'naja semantika otsenki], Moskow.
15. *Wetzer H.* (1996), The typology of adjectival predication. Berlin: Mouton de Gruyter.
16. *Yavorskaya M. and Azarova I.* (2010), Hierarchy of perceptional adjectives in RussNet // Proceedings of the 5th Global WordNet Conference, Mumbai, pp. 216–223.

DEVELOPMENT OF LEXICAL BASIS FOR THE UNIVERSAL DICTIONARY OF UNL CONCEPTS

Dikonov V. G. (dikonov@iitp.ru)

IITP RAS, Moscow, Russia

The paper describes the current state of development of the lexical basis of an open and free lexical-semantic resource — the Universal Dictionary of UNL Concepts (UNLDC). The resource serves as a lexicon of an artificial intermediary language UNL (Universal Networking Language). It links the elementary units of UNL — concepts with lexicons of natural languages and various external lexical and semantic resources, including Wordnet and SUMO ontology. The dictionary's main goal is to support automated semantic analysis, encoding the meaning of the text as UNL semantic graphs and subsequent generation of text in different natural languages.

Keywords: lexical resources, semantics, interlingua, UNL

1. Introduction

The dictionary of the artificial interlingua UNL (Universal Networking Language), also called Universal Dictionary of UNL Concepts (UNLDC) is a part of an international project to develop UNL [Boguslavsky, et. Al, 2005], [UNL Specification 2005]. The development of this resource is supported by the members of the "U++ Consortium", which unites researchers from Russia, France, Spain and India. Although UNL is the main application of the resource, it also has certain value of its own and can be used for scientific and practical tasks not directly related to UNL.

The basic units of the dictionary are so called UNL concepts, which correspond to word senses described by traditional explanatory dictionaries. They are also similar to semanthemes in the Meaning↔Text theory by I. A. Melchuk, which includes deep syntax and semantic representations with many features in common with semantic graphs of UNL. The notion of UNL concept is well aligned with lexicographic tradition. This allows to reuse much of the natural language data already gathered by explanatory dictionaries and thesauri.

UNLDC defines the inventory of concepts in U++ UNL. Each concept receives a unique identifier called Universal Word (UW) conforming to the U++ standard. Each new UW should be added to the dictionary before being used in any UNL encoded documents in order to maintain lexical compatibility between different software tools supporting generation of natural language text from UNL. UNLDC links UWs with words and expressions of natural languages that can be used to express corresponding concepts.

The dictionary has three main parts:

1. List of the U++ Universal Words (UWs),
2. Semantic network that links the UWs together,
3. Set of local dictionaries linking UWs with words and expressions of natural languages.

Entries have links to external lexical and semantic resources, including internal dictionaries of several MT systems. The UWs and their links to natural languages are annotated to keep track of their source, status and expected quality. All data is split into several complementing each other "volumes" to simplify maintenance. Each volume is stored in a separate file. Different volumes may be used to represent languages and alternative orthographies/dialects of the same language. Large groups of entries, such as domain terms or named entities, are split away into separate volumes too.

The general overview and full introduction to UNLDC has been given in [Dikonov V., Boguslavsky I., 2009]. In this paper we describe the results of our work done in 2012 to extend the lexical coverage of UNLDC in Russian and integrate data for other natural languages provided by our partners.

2. Current state of the project

At the time of writing the total number of UWs in the dictionary has reached 2,880,661. There are seven local dictionaries of Russian, English, French, Hindi, Spanish, Vietnamese and Malay. A considerable part of the semantic network is completed. The core of the semantic network — its ontological structure — is modeled on the basis of the SUMO ontology [Pease, 2011].

The main priority is given to the core lexical part of UNLDC, which covers most frequently used words of English and Russian, and some special semantic units corresponding to lexical functions, modal words and some closed class words. In addition to this part we develop separate extension volumes containing basic terminology and named entities. The extensions are still at an early stage and contain only automatically gathered data. Table 1 shows, how many UWs are there in each part.

Table 1. Number of UWs by dictionary part in early 2013

Part	UWnumber	File	Status
General lexics	82,804	CommonUNLdict-XML-0.04.1-alpha.tar.bz2	Downloadable
Terminology	688,617	CommonUNLdict-CSV-Terminology-0.02.tar.bz2	Under development
Named Entities	2,109,240	CommonUNLdict-CSV-NamedEntities-0.01.tar.bz2	Soon to be released

The principle approach to the development of the resource is accumulation and integration of data available in the Internet with subsequent proofreading and gap filling. The initial versions of the local dictionaries are built using automated methods of import and cross-linking of different sources. The preferred sources are more reliable ones, such as translation dictionaries, or semantically annotated resources (Wordnets, ontologies, other UNL dictionaries). Since every resource contains some errors and automatic integration tends to multiply them, we have to to put a lot of effort into finding and fixing the resulting "noise". It is planned that all automatically gathered data will eventually be proofread. However, the manual verification process is far too labor and time consuming to be applied to all the data. Therefore, proofreading is done only for the most important parts of the resource. Some errors are detected by applying formal criteria. For example, if an UW is linked to words of several languages, but no translation dictionary confirms that these words are good translation equivalents of each other, we can suspect that some of the word↔UW links are wrong and lower their reliability score. Another way to sift through the data automatically is to check if the semantic argument frame manually ascribed to the UW matches its taxonomic class in the associated ontology and if both are compatible with any external semantic annotations linked to the corresponding words of natural languages.

2.1. Local Dictionaries

The existing local dictionaries are not equal in size and quality. They have been built using different approaches and from different data. The most interesting part is the general lexics part, because it is being proofread by hand. Table 2 quotes the number of UWs with translations by language and dictionary part with rough quality estimation.

Table 2. Number of UWs linked to words and expressions of natural languages in early 2013

Language	General lexics	Terminology	Names	Total	Quality estimation
English	**82,804** (40,894 words)	688,617	2,109,240	2,880,661	*****
Russian	**48,555** (30,818 words)	688,613	226,595	963,763	**** Manual proofreading in progress
French	**36,324** (25,068 words)	103,060	367,888	507,272	*** Automatic verification
Hindi	**27,815** (30,220 words)	0	10,823	38,638	*** Automatic verification
Spanish	**11,758** (6,983 words)	21,990	298,674	332,422	** Experimental
Malay	**21,861** (17,457 words)	0	46,044	67,905	** Experimental
Vietnamese	**5,927** (6,456 words)	0	171,367	177,294	*** Experimental

2.2. Volumes of general lexics

The English general lexics volume was built from the Princeton Wordnet 2.1 as a result of the work done by the Spanish UNL center. English is used to form the majority of UWs, so almost all UWs links to some English words or phrases. The Wordnet data were supplemented with new UWs standing for semantically non-void prepositions and conjunctions and some phrasal verbs. 13,811 UWs representing lexical senses of 6,723 English words were rewritten or created by hand. Those changes concerned the most frequent English words. Further, manual changes were made in about 5,000 predicate UWs to fix bad argument frame descriptions. The general English local dictionary does not contain all of the Wordnet. We rejected all multiword expressions from Wordnet and chose about 40,000 most frequent words to facilitate linking to the internal dictionary of the ETAP-3 MT system.

The Russian general volume 0.05-alpha registers over 48,000 senses of 30,800 Russian words. It is being gradually improved by the author through proofreading and adding new data. It still includes only those Russian words and word senses (with very few exceptions) that serve as translations of already existing English Wordnet senses. This happens due to the nature of the process used to construct the initial version of the Russian dictionary and the necessity to a) clean up the unavoidable errors before adding more specifically Russian lexical data and b) to maximize the percentage of UWs that have translations into all supported natural languages. The next version 0.06 is going to include Russian words that lack direct single word translations into English. The examples include such common words as *старик (old man), касса (cash register), телевизор (TV set), молчать (keep silence), белеть (*be seen as white), приходить (come on foot),* historic and cultural phenomena, e.g. *самовар (samovar), щи (cabbage soup)*, *лапоть (peasants' shoe)*, *хохлома (khokhloma)*, *по-черному (house without a chimney), etc*.

The French volume has been built automatically on the basis of the free French Wordnet WOLF 0.1.5 [Sagot, 2008]. It includes only those French words that were linked to the Princeton Wordnet synsets, the members of which had matching UWs in the subset forming the core of UNLDC. The WOLF data were supplemented by the lexical data provided by the French UNL center and further ranked through automatic comparison of the resulting possible French-Russian translations with regular French-Russian dictionaries.

The Hindi volume is made from the UNL dictionary supplied by the Center For Indian Language Technology at the Indian Institute of Technology in Mumbai. It can be expanded using the existing open Hindi Wordnet.

The Spanish volume is based on the small public Spanish Wordnet.

The Malay and Vietnamese volumes are the results of experiments in assimilating lexical data from regular translation dictionaries. We developed tools that can automatically pair the already registered UNL concepts with words/symbols or expressions of arbitrary natural language taken from dictionaries that translate them into the already supported natural languages, e.g. English, Russian and French.

2.3. Links to external resources

UNLDC is being built using data from other open resources and can in turn become a source of data for other projects. The links with external resources are important to enable easy exchange of linguistic data. The UWs in UNLDC are connected with Princeton Wordnet 2.1 and 3.0, Suggested Upper Merged Ontology (SUMO) [Pease, 2011], DbPedia Ontology. Table 3 shows the number of existing links. More external resources may be added to this list in future.

Table 3. Statistics of the links to external resources

Part	UW number	Connected with
General	77,671 77,293	Princeton Wordnet 2.1 и 3.0 SUMO ontology
Terminology	all part	Upper SUMO classes (not reliable) Domain ontologies
Named entities	all	DbPedia ontology Upper SUMO classes

UWs and Russian, English, French and Hindi words also have pointers to the entries of internal dictionaries of linguistic processors supporting those languages. Such connections are needed to convert UNL semantic graphs to text in different languages. The Russian local dictionary links 40,175 UWs with 26,188 entries of the Russian combinatorial dictionary used by the ETAP-3 system. Figure 1 shows the diagram of links between parts of UNLDC and external resources.

The ontology pictured in Fig. 1 is our custom OWL rendering of SUMO. It does not include complex axioms from SUMO due to the differences in expressivity between KIF and OWL languages. The method of producing the OWL version permits periodic synchronization with SUMO and custom changes in the resulting ontology. The ontology is used to form the ontological part of the semantic network connecting all UNL concepts. This part is under active development.

2.4. Data files

The data files are public and can be downloaded under GPLv3+ and Creative Commons CC-BY-SA. Currently the data is provided in two formats: simple tab separated tables CSV and XML for uploading into the lexical database Jibiki/Pivax [Boitet et al., 2007], which can be used as an online search tool. It is also planned to add the RDF/Turtle format conforming to the Semantic Web Linked Data principle. The released data files are available at http://atoum.imag.fr/geta/User/services/pivax/data/. The currently available files include two out of three main parts of the dictionary: the list of concepts (unlvolume) and loacal dictionaries (rusvolume, fravolume, engvolume etc.).

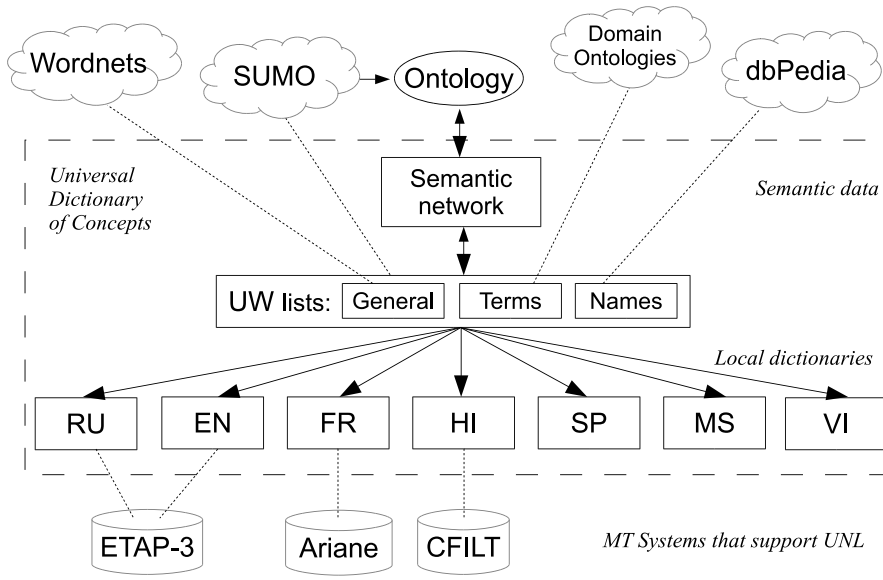


Fig. 1. Links between parts of UNLDC and external resources

3. Some features

Each UW in UNLDC has a special mark showing which language motivated its creation. If it is discovered that translations of the same UW into different languages display subtle differences in meaning or connotations and the conflict is not resolved by ontology, or the ontological classification itself is doubtful, the specified language becomes the final reference.

3.1. Levels of semantic affinity

The dictionary structure allows to distinguish full and quasi synonyms. A single UW may have several translations into the same natural language that are supposed to be full synonyms. For example, the UW *hydroplane(icl>airplane>thing,equ>seaplane)* is linked to two Russian words *гидроплан* and *гидросамолет* with exactly the same meaning. If two words are not interchangeable in most contexts, they should be connected to different UWs. The UWs themselves should than be linked by a synonymy link "equ". Most of the existing synonymy links were imported from Princeton Wordnet. Unlike the Wordnet, groups of synonyms are not treated as units representing one common meaning. Synset members are considered quasisynonyms, unless the opposite is proven. For example, the UWs *hydroplane(icl>airplane>thing,equ>seaplane)* and *seaplane(icl>airplane>thing)* can be merged together, if it turns out that in all languages but English they receive identical translations.

3.2. Special UWs

Usually in order to translate a concept into a natural language one only needs to check, what words are linked to the UW in the local dictionary. However, there are special categories of UWs, which are hard or impossible to translate in the regular way. First of all, it concerns the UWs standing for abstract concepts known as Lexical Functions (LF). The dictionary contains some UWs equivalent to "collocate" type LFs. They may be used to avoid faulty literal interpretation of certain idiomatically used words. For example, the current version of our UNL semantic analyzer renders words *take* and *have* in *I took a short walk* and *I had a short rest* as a special UW *perform_an_action(icl>do,agt>thing,obj>process)*, which corresponds to the abstract meaning of the lexical function OPER1. Its translation into other languages depends on the LF's argument and may be empty. Compare Russian translations *Я немного прогулялся (I walked a little)* and *Я слегка отдохнул (I rested a little)*.

Another group of UWs that are hard to correlate with individual words of natural languages is modal predicates. These UWs are parallel to UNL modal attributes used to encode modality and constitute a well organized system described in [Dikonov, 2009]. In many languages, including English and Russian, words used to express modality are polysemic and represent different modal meanings in different contexts. In UNL the same modal attribute or UW is used to encode a given modality regardless of what modal word was used in the source sentence. E.g. the prohibition in *You may not carry a weapon here* and *You can not smoke onboard* is expressed by two different modal verbs but in UNL the same symbol is used to represent both. The dictionary will link the special modal UW *grant-not(icl>modal>be,obj>uw,aoj>thing)* to both *can not* and *may not* leaving the choice between them to the processor or the user.

3.3. Multiword expressions

Concepts may be translated into some languages by multiword phrases and/or grammar constructions with a variable part. The target language expression may be written not only in the form of a simple n-gram, but also include syntactic relations between the words following the notation of the MT system that is supposed to support the target language. For example, the UW *bathroom(icl>room>thing)* is translated into French as the phrase *salle de bain*, which has the structure recorded in the format of the French MT system Ariane as: *@@_1:'salle' (2:'bain')::1 ": CAT(CATN),GNR(FEM),N(NC). 2 ":CAT(CATN),GNR(MAS), N(NC), NUM(PLU), ART(ABS), RSUNL(MOD)*.

In addition to recording the phrase structure in the formats understood by specific MT systems, UNLDC may use a more general format similar to the text form of UNL graphs. In such case the UWs are replaced with corresponding words of the target natural language. The words are linked with UNL semantic relations and carry UNL attributes. This option is already used in the terminological part of the dictionary. A UNL-capable system should be able to convert such phrase into a fragment of its own internal representation, if applicable. For example, the Russian verb *белеть* in *Белеет парус одинокий* must be translated into English by grammatic constructions

with a variable: *A white* (lonely sail) *can be seen/There is a white* (lonely sail)*/A white* (lonely sail) *appears* (at …). Here the brackets contain a variable NP which cannot be omitted. The English local dictionary might contain the following record:

*mod(*concrete_thing(icl>thing).*@indef.@topic,white),*

obj(see.@entry.@ability, concrete_thing(icl>thing).*@indef.@topic),*

which corresponds to *a white … can be seen*. The UWs in bold instead of words describe the general class uniting the words that could be inserted into the specified slots to build a good translation.

4. Comparable resources

The most important resources similar to UNLDC are the Wordnet family with "Inter Language Index" as a whole and parallel projects building dictionaries for other flavors of UNL. There are two major alternative UNL resources: the dictionary of UNL Development Center developed under the lead of Hiroshi Uchida and UNLarium dictionaries. All of them collect multilingual lexical data and provide some semantic annotation.

4.1. Other UNL dictionaries

The UNL Development Center (UNDC) dictionary, http://www.undl.org/unlexp/ was built by automatic corpus mining using methods common among the developers of statistical and example-based MT systems. This is confirmed by the existence of a large number of characteristic errors. The method does not take into account morphological features and in the case of languages with rich morphology, such as Russian, that dictionary resorts to crude approximation techniques trying to discover lemmas. As a result, the UNDC dictionary contains a lot of wrong lemmas and false translations. There are some differences in the UW formation standards too. UNDC permits UWs consisting of bare English headwords without any constraints that would disambiguate their meaning. The U++ UWs always include some ontological constraints and list arguments of predicates.

The second UNL resource — UNLArium UNLdic (http://www.unlweb.net/unlarium/) is based on the Wordnet and we would criticize it for using numerical synset codes form Wordnet directly as UWs to identify concepts. The numbers merely refer to the size of the offset in bytes from the beginning of a Wordnet data file to the start of the synset record. They change between versions of Princeton Wordnet. Sticking to the numbers would severely limit the range of possible concepts in the resource, so UNLarium supplements them with proper UWs for concepts not directly linkable to the Wordnet. UNLdic is license compatible and it is relatively easy to exchange data between our projects.

4.2. Lexical networks

UNLDC contains a lot of data imported from Princeton Wordnet and reuses its synonymy and antonymy links. Even parts that were left out may be imported later. However even the English section of UNLDC is not an exact copy of the Wordnet and includes some new data, in particular the description of prepositions and conjunctions. More differences between UNLDC and Wordnet were described in [Dikonov V., Boguslavsky I., 2009].

Local dictionaries of other languages, e.g. French or Spanish, can be created from open Wordnets of those languages. The data contained in UNLDC and other linked resources together makes up a superset of a typical Wordnet. It is possible to generate new Wordnets from UNL data. For example, there is still no open and free Russian Wordnet that would be larger than 30,000 words. The already existing semantic and ontology relations make it possible to join registered Russian words into synsets and form a new open Russian Wordnet. The French local dictionary might fill in some gaps in WOLF, etc.

There are other projects of multilingual lexical networks consisting of words and translation links between them. Such projects usually lack semantic annotation, but their data can be used to extend the coverage of the semantic dictionary. One example is the PanLex project [Baldwin et al, 2010]. There are several Internet sites offering croudsourced multilingual dictionaries, e. g. Wiktionary, freedict, etc. All of them collect potentially useful data that can be used in future.

The existence of other projects aiming to integrate lexical data available in the Internet, e. g. BabelNet proves a widespread interest and importance of resource integration for the development of applied linguistics.

References

1. *Baldwin T., Pool J., Colowick S.* PanLex and LEXTRACT: Translating all Words of all Languages of the World, 2010
2. *Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L.* The UNL Initiative: An Overview. Computational Linguistics and Intelligent Text Processing, 2005
3. *Boguslavsky I., Dikonov V.* Universal Dictionary of Concepts [Universal'nyj slovar' konceptov] Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009"]. Bekasovo, 2009 M.: RGGU, 2009. Issue. 8(15). pp. 91–96. ISBN 978-5-7281-1102-3.
4. *Boguslavsky I.* Guidelines for UW construction, manuscript
5. *Dikonov V. G.* Modal Attributes in UNL [Atributy modal'nosti v UNL]. Sbornik trudov 32-oj Konferencii molodyh uchenyh i specialistov IPPI RAN "Informacionnye tehnologii i sistemy (ITiS'09)" [Proceedings of the 32-nd Conference "Information technologies and systems (ITiS'09)"], Bekasovo, 2009. pp. 230–237. ISBN 978-5-901158-11-1.

6. *Dikonov V.* Semantic Network of the UNL Dictionary of Concepts. Proceedings of the SENSE Workshop on conceptual Structures for Extracting Natural language Semantics
7. *Iraola L.* Using WordNet for linking UWs to the UNL UW. International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, Alexandria, EGYPT, 2003
8. *Nguyen Hong-Thai, Boitet C., Sérasset G.* PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot, SNLP-2007, Bangkok, 2007
9. *Pease, A.* (2011). Ontology: A Practical Guide. Articulate Software Press, Angwin, CA. ISBN 978-1-889455-10-5.
10. *Sagot B., Fišer D.* Building a free French wordnet from multilingual resources. Ontolex 2008, Marrakech, Maroc, 2008
11. *UNL* Specification 2005, available at: http://www.undl.org/unlsys/unl/unl2005/

GERMAN-RUSSIAN IDIOMS ONLINE: ON A NEW CORPUS-BASED DICTIONARY[1]

Dobrovol'skij D. O.
(dm-dbrv@yandex.ru, dobrovolskij@gmail.com)

V. V. Vinogradov Russian Language Institute RAS,
Moscow, Russia

The paper focuses on the structure and principles for constructing a new German-Russian phraseological dictionary based on corpus data. Fragments of this dictionary are available on the website of the German Language Institute in Mannheim: "Deutsch-russische Idiome online" http://wvonline.ids-mannheim.de/idiome_russ/index.htm. Relevant information is also made available via the Europhras homepage at http://www.europhras.org. In section 1, I formulate certain general principles of modern bilingual phraseology. Section 2 discusses the state of the art of German-Russian phraseography and explains the need for a new German-Russian phraseological dictionary. In Section 3, key features of the new corpus-based dictionary are considered. The basic difference between the present dictionary and traditional ones is that all examples of idiom usage are taken from text corpora DeReKo and DWDS, and in individual cases from the German-language Internet. Parallel texts from the Russian National Corpus (RNC) are also used. The use of authentic examples based on text corpora is a new approach in bilingual lexicography. Traditional dictionaries were based on a limited body of randomly selected examples, and the use of the idioms was often not even exemplified. The advantages of using corpora consist not only in more detailed and well thought-out illustrations of the expressions being described, but also in additional possibilities that the corpus provides for compiling the idiom-list and structuring entries.

Key words: dictionary, corpus, bilingual lexicography, phraseology, idiom, German, Russian

1. Corpus-based bilingual phraseography and cross-linguistic equivalence

The lexicographic treatment of the notion of equivalent in dictionaries based on corpus data encounters certain problems. Not infrequently, the generally accepted equivalent of an idiom cannot always be used to translate authentic texts.

Let us take an example. The German idiom *jmdn. an der Nase herumführen* has a "standard" equivalent in Russian, namely the idiom *водить за нос кого-л.* It would

[1] This paper is based on work supported by the RGNF under Grants 11-04-00105a, 12-04-12041, 12-34-10413 and by the Basic Research Program "Corpus Linguistics" of the Presidium of the Russian Academy of Sciences.

be somewhat odd to doubt that these idioms are basically equivalent, since they are identical with respect to both their lexicalized meaning and image component. Nevertheless, it turns out that it is far from always possible to translate the expression *jmdn. an der Nase herumführen* with the Russian idiom *водить за нос кого-л.* Numerous contexts with the idiom *jmdn. an der Nase herumführen* can be found in text corpora in which this idiom has to be translated into Russian either by the verbs *надуть* and *одурачить* or by the idiom *обвести вокруг пальца*.

(1) Die Aktionäre fühlen sich vom größten deutschen Industriekonzern *an der Nase herumgeführt.* (Mannheimer Morgen, 08.08.1995)
У акционеров такое чувство, что самый большой промышленный концерн Германии *обвел* их *вокруг пальца*.
In Wahrheit hatte er [Wolfgang Schäuble] aber 100.000 Mark <…> bekommen <…>. Und das hat er im Deutschen Bundestag <…> verschwiegen und hat das erst später, vier Wochen später in einem Fernsehinterview aufgedeckt und da haben viele gesagt, <…> der hat den Deutschen Bundestag *an der Nase herumgeführt*. [www.stroebele-online.de/themen/spendenaffaere/29273.html]
На самом деле он [Вольфганг Шойбле] получил 100 000 марок. Причем он скрыл это от бундестага и только позднее, спустя четыре недели, признался в этом во время телеинтервью. И многие сказали тогда: он просто *одурачил* немецкий парламент.

Consequently, despite the intuitively felt equivalence of the expressions *jmdn. an der Nase herumführen* and *водить за нос кого-л.*, this equivalence cannot be considered complete. For the lexicographer interested in a maximally precise description of the material, such instances are problematical. Either we acknowledge that *jmdn. an der Nase herumführen* and *водить за нос кого-л.* are equivalent, in which case it is necessary to explain why the "standard" equivalent is unacceptable in a number of contexts, or we deny that a relationship of bilingual equivalence obtains between *jmdn. an der Nase herumführen* and *водить за нос кого-л.*, and focus exclusively on translating specific contexts. Such a solution, however, is counterintuitive.

There are at least two ways out of this cul-de-sac. Either we refrain from giving equivalents and replace them with an explanation (here permissible target-language correspondences can be given in a special field in the entry — cf. Lubensky 1995), or we provide the given equivalents with a commentary indicating relevant limitations.

In our dictionary we have followed the second path. Thus for the German idiom *jmdn. an der Nase herumführen* we give the Russian equivalent *водить за нос кого-л.* and explain divergences in the use of the idioms in the commentary, where we point to the fact that the Russian idiom *водить за нос кого-л.* is an imperfectiva tantum, i.e. it cannot normally be used in the perfective aspect. Contexts such as *а народ не дурак, за нос его так просто не проведешь* or *за нос такого провести нетрудно* are encountered quite rarely. The use of this idiom in the perfective aspect is licensed only in non-veridical contexts. For more detail see (Dobrovol'skij 2013).

A question that arises from the perspective of phraseological theory (especially its comparative aspects) concerns the essence of cross-linguistic equivalence of idioms. Does it really exist?

At first glance, this question seems to be quite simple. Those not involved in idiom research would immediately give a positive answer to this question. Really, how can one doubt the existence of cross-linguistic equivalence in the field of phraseology when there are so many bilingual, and even some multilingual dictionaries of idioms? The aim of such dictionaries is, above all, to provide the user with the knowledge of cross-linguistic idiom-equivalents, therefore such equivalents must exist. And besides, there is rather a long tradition of contrastive idiom research (compare, e.g., review articles Dobrovol'skij 2002 and Korhonen 2007).

1.1. Types of phraseological equivalence

Within this tradition some well-known types of phraseological equivalence are discriminated. For the sake of simplicity, these types have been reduced here to the following four main classes:

(i) "full equivalents",
(ii) "partial equivalents",
(iii) "phraseological parallels", and
(iv) "non-equivalents".

(i) "Full equivalents" (or "absolute equivalents") are idioms of L1 and L2 which are identical with regard to meaning, syntactic and lexical structure, and imagery basis. Compare German *seine Hand ins Feuer legen für etw.* and English *to put one's hand into the fire for sth.*; English *to rest on one's laurels*, German *auf seinen Lorbeeren ausruhen* and Russian *почивать на лаврах*. Some "full equivalents" allow for morphological or certain lexical alternations, cf. the singular-plural alternation in the following idioms: German *von Kopf bis Fuß* (lit.: "from head to foot") and Russian *с головы до ног* (lit.: "from head to feet"), or German *ganz Ohr sein* (singular), English *to be all ears* (plural) and French *être tout oreilles* (plural). It seems more natural to regard equivalents of this kind as (i) rather than (ii), though in the tradition of contrastive idiom research also quite different views on this issue can be found.

(ii) "Partial equivalents" are idioms of L1 and L2 which have identical or near-identical meanings, but do not fully correspond in syntactic and lexical structure, or imagery basis. Compare English *to get out of bed on the wrong side* and Russian *встать не с той ноги* (lit.: "to get out [of bed] with the wrong foot"), German *aus einer Mücke einen Elefanten machen* (lit.: "to make an elephant out of a mosquito") and Russian *делать из мухи слона* (lit.: "to make an elephant out of a fly"), or German *die Hände über dem Kopf zusammenschlagen* (lit.: "to strike the hands over the head") and Russian *схватиться за голову* (lit.: "to grip one's head").

(iii) "Phraseological parallels" are different idioms of L1 and L2 which correspond to each other in the core meaning, but not with regard to the image component. Cf. English *hot potato* and German *heißes Eisen*, English *to be like a cat on hot bricks* and German *wie auf glühenden Kohlen sitzen*, English *to buy a pig in a poke* and

German *die Katze im Sack kaufen*, German *jmd. hat nicht alle Tassen im Schrank* and Russian *у кого-л. не все дома* (lit.: "sb. does not have them all at home"), English *to take a sledgehammer to crack a nut* and Russian *стрелять изз пушек по воробьям* (lit.: "to use cannons to shoot at sparrows"), or English *to play a dirty trick on sb.* and Russian *подложить свинью кому-л.* (lit.: "to put a pig on sb."), English *spic and span* and Russian *одетый с иголочки* (lit.: "dressed from the needle"). "Phraseological parallels" are semantically similar, but their planes of expression mostly do not have much in common.

(iv) "Non-equivalents", i.e. a given L1-idiom has no idiomatic correspondences in L2; compare the Russian idiom *объяснить на пальцах что-л.* (lit.: "to explain sth. on fingers") meaning 'to explain sth. as simply as possible' which has no idiomatic counterpart in English, or German *etw. nicht übers Herz bringen können* (lit.: "not to be able to bring sth. over the heart") which can only be translated into Russian using free word combinations such as *не мочь себя заставить (сделать что-л.)* 'not to be able to force oneself (to do sth.)'.

So, at least in category (i) we seemingly are dealing with real cross-linguistic equivalents, also most members of category (ii) and even some members of category (iii) can be considered good candidates for equivalence. Also from the point of view of text translation there seems to be rather a clear case for the existence of equivalents. It is obvious that almost every literary text analyzed so far contains idioms, sometimes a lot of them. While translating these texts into other languages, translators have to tackle the issue of cross-linguistic equivalence of idioms encountered in the original.

Nevertheless, I doubt that the question as to whether cross-linguistic idiom-equivalents really exist can be answered positively without a serious linguistic discussion. It appears that there are too many stumbling blocks on the way to finding real cross-linguistic equivalents. So, it seems to be an exciting task to discover them, to discuss their nature and, above all, to investigate the reasons why idioms and other semantically corresponding lexical items which look like full equivalents do not always function as such. Cf. example (1) above.

1.2. Aspects of equivalence

It seems expedient to distinguish two different aspects of equivalence:

a) equivalence in translation; that is, the relationship between an idiom of language L1 and its translation into language L2 in a particular text, and
b) equivalence in the language system; that is, the relationship between the compared idioms of L1 and L2 on the systemic level.

One of the most important differences between translational and systemic equivalence (besides the fact that the former has to do with a concrete text and the latter with the lexical system) consists in the circumstance that equivalence in translation is a unilateral relationship, whereas equivalence in the language system is defined as bilateral. In other words, if an idiom of language L1 is equivalent to an idiom in language L2, this means that the L2 idiom is also equivalent to the corresponding L1 expression. With

respect to equivalence in translation, all that is being said is that an expression in language L2 is being used in the translation of some specific text in language L1 in such a way that between the L1 idiom from this particular text and the L2 expression there is a relationship of semantic correspondence. The fact that the translation of some L1 idiom into language L2 is its equivalent (at least with respect to this particular context) does not, of course, mean that the relationship can be reversed. That is, the L1 idiom should not be regarded as an equivalent of the expression used in the translation of this idiom into language L2 (even if this expression is an idiom, which is not at all obligatory). Obviously, the study of equivalence in translation broadens our notions about the possibilities of cross-linguistic paraphrasing and about the role of contextual conditions in the selection of adequate correspondences, and it contributes to the development of both translation theory and comparative phraseology.

As for equivalence in the language system, its study has both theoretical and practical significance for phraseology. Deserving of special attention from the theoretical point of view is the question of why one and the same concept is expressed by means of an idiom in one language but not in another. Another (no less important) problem concerns the fact that between basically similar idioms in language L1 and language L2, there are practically always certain semantic, pragmatic, and collocational differences that must be discovered and described. This is especially important in cases where a traditional description postulates a relationship of "full equivalence" but ignores the absence of functional interchangeability between the idioms. The practical aspect of systemic equivalence is what is reflected in bilingual dictionaries, where the entry consists of an idiom of language L1 (in the lemma) and its idiomatic (to the extent this is possible) correlates in L2. Can these correlates be regarded as equivalents of the L1 idiom? Yes and no. On the one hand, they must be at least "partial equivalents" or "phraseological parallels", for otherwise they could not be placed in the corresponding dictionary entry. On the other, often they cannot be used in the translation of specific texts. The reason, as a rule, is that the idioms of L1 and L2 display certain differences in their semantic, pragmatic, and collocational features. They can be considered cross-linguistic equivalents only in a rather approximate comparison of the idioms of the given languages, and are the starting point of a thorough contrastive analysis that attempts to discover the unique properties of each idiom and thereby improve the lexicological and lexicographical description of phraseology.

Obviously, aspects (a) and (b) are, as it were, two sides of the same phenomenon or two approaches to studying it. We assume that one of the principal goals of contrastive phraseology is to discover genuine equivalents — that is, those that are as close as possible with respect to their actual meanings and — ideally — with respect to the inner form of the expressions (i.e. with respect to images underlying their lexicalized meanings), and that function equally well in analogous types of situations, which does not at all imply an obligatory "idiom — idiom" relationship.

What is important for cross-linguistic correspondence, after all, is not "phraseologicalness," but **functional equivalence**. It is this type of equivalence that is most interesting from the perspective of bilingual lexicography.

Functional equivalents, how I understand them, are counterparts which can be used in the same concrete situations without any informational loss. To find them out we have

to simultaneously go two ways: from text to language system and from language system to text. As we have seen, not all systemic equivalents can function as counterparts in authentic texts, and on the other hand, not all translational equivalents can be included in the dictionary as typical parallels suitable for using in neutral contexts.

In contrast to a conception that is wide-spread within traditional phraseology, I claim that lexical units of any kind in L2 which have the identical meaning and, in the ideal case, near-identical metaphorical basis as the L1-idioms from the source text are excellent functional equivalents, so they have to be considered not only more or less appropriate translational solutions, but also real functional equivalents, i.e. parallels in the lexicons of L1 and L2, which have to be fixed lexicographically.

2. German-Russian phraseography: state of the art

The need for a new German-Russian phraseological dictionary is motivated by the fact that existing such dictionaries do not meet present requirements. Both the vocabulary and the examples in (Binovich, Grishin 1975) are out of date, and the work fails to satisfy current needs with respect to a number of other parameters as well. Although (Dobrovol'skij 1997) is on the whole more up to date, it also has certain shortcomings. Its idiom-list is rather limited, and illustrative examples are often arbitrary and unpersuasive, which may be because it was written back in the "pre-corpus era." Actually, one of the basic goals of our new lexicographical project is to eliminate all the shortcomings of this dictionary and to significantly expand its idiom-list.

Yet another dictionary of this type has appeared recently: "The New German-Russian Phraseological Dictionary" (Shekasjuk 2010). Its phraseme-list is fairly large and up to date, but the work is difficult to use, primarily because the illustrative examples are not translated into Russian, and the division of entries into meanings and selected equivalents often appears hasty and arbitrary.

Thus there is an unquestionable need for a new dictionary containing the most widely used contemporary German idioms together with carefully selected Russian equivalents, explanations facilitating the correct use of these idioms, and good, authentic examples translated into Russian. It is also important that such a dictionary exist not only in print, but also in an online version, which will not only provide easier access to the information but will also ensure continuous revision and improvement.

3. Key features of the new dictionary

The basic difference between the present dictionary and traditional ones is that all examples of idiom usage in it are taken from the text corpora DeReKo and DWDS, and in individual cases from the German-language Internet. Parallel texts from the Russian National Corpus (RNC) are also used. These examples are especially valuable because they have been translated by professional translators rather than by the

authors and editors of the dictionary. Since this part of the parallel corpus of the RNC is still rather modest in size, however, examples needed for the dictionary were rarely encountered.

The use of authentic examples based on text corpora is a new approach in bilingual lexicography. Traditional dictionaries were based on a limited body of generally randomly selected examples, and the use of the idioms was often not even exemplified. The advantages of using corpora consist not only in more detailed and well thought-out illustrations of the expressions being described, but also in the additional possibilities that the corpus materials provide for compiling the idiom-list and structuring entries. For further detail see (Dobrovol'skij 2013).

Yet another advantage of using corpora is that it increases our ability to determine the peculiarities of the formal and semantic structure of idioms, particularly in the description of the ambiguity and variation of a form. Although an analysis of examples of use clearly indicates that polysemy in phraseology is an extremely widespread phenomenon, traditional dictionaries rarely distinguish the different meanings of idioms, and seldom reflect the full diversity of variants actually represented in texts. Dictionaries often register only a single "canonized" form of an idiom that in many cases proves to be not the most frequent one.

In a number of instances text corpora allow us not only to determine the form of a lemma and a selection of its most frequent variants, but also to establish whether a given expression belongs to the sphere of phraseology. For example, Duden 11 (2002) cites four idioms with the noun *Mundwerk*: *jmds. Mundwerk steht nicht still* (ugs.) 'jmd. redet ununterbrochen'; *ein böses/lockeres/loses/freches* o.ä. *Mundwerk haben* (ugs.) 'gehässig/vorlaut/frech o.ä. reden'; *ein gutes/flinkes Mundwerk haben* (ugs.) 'sehr gewandt reden'; *ein großes Mundwerk haben* (ugs.) 'großsprecherisch reden'. Corpus analysis has shown that the noun *Mundwerk* has a much broader combinatorial profile. Compare, e.g., *flottes, vorlautes, geschliffenes Mundwerk*. This noun can also be used without any adjectives, combining with verbs of various meanings. Cf.

(2) Manchmal wäre es vielleicht sinnvoller, mein *Mundwerk* etwas zu zügeln, nach dem Motto «Reden ist Silber, Schweigen ist Gold». (St. Galler Tagblatt, 08.04.1999)

Consequently, what we have to do with here is not an idiom but a series of relatively free collocations.

References

1. *Binovich L. È., Grishin N. N.* (1975), German-Russian Phraseological Dictionary [Nemecko-russkij frazeologicheskij slovar'], Russkij jazyk, Moscow.
2. *Dobrovol'skij D. O.* (1997), German-Russian Dictionary of Current Idioms [Nemecko-russlij slovar' zhivyh idiom], Metatext, Moscow.
3. *Dobrovol'skij D. O.* (2002), Phrasemes in cross-linguistic perspective [Phraseologismen in kontrastiver Sicht], Lexikology: An International Handbook on the

Nature and Structure of Words and Vocabularies. Volume 1. de Gruyter, Berlin/New York, pp. 442–451.
4. *Dobrovol'skij D.* (2013), German-Russian phraseography: on a new dictionary of modern idiomatics, Phraseodidactic Studies on German as a Foreign Language. Verlag Dr. Kovač, Hamburg, pp. 121–138.
5. *Duden 11* (2002) = Duden — Dictionary of German Idiomatics [Duden — Redewendungen. Wörterbuch der deutschen Idiomatik (=Der Duden, Band 11)]. Dudenverlag, Mannheim etc.
6. *Korhonen J.* (2007), Issues of cross-linguistic phraseology [Probleme der kontrastiven Phraseologie], Phraseology: An International Handbook of Contemporary Research. Volume 1. de Gruyter, Berlin/New York, pp. 574–589.
7. *Lubensky S.* (1995), Random House Russian-English Dictionary of Idioms. Random House, New York.
8. *Shekasjuk B. P.* (2010), The New German-Russian Phraseological Dictionary [Novyj nemecko-russkij frazeologicheskij slovar'], Librokom, Moscow.

Digital resources

1. *DeReKo* — Das Deutsche Referenzkorpus des IDS Mannheim im Portal COSMAS II (Corpus Search, Management and Analysis System): https://cosmas2.ids-mannheim.de/cosmas2-web
2. *DWDS* — Corpora des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts: http://www.dwds.de
3. *Online* dictionary "Deutsch-russische Idiome online": http://wvonline.ids-mannheim.de/idiome_russ/index.htm
4. *RNC* — Russian National Corpus [Nacional'nyj korpus russkogo jazyka]: http://www.ruscorpora.ru

МОДЕЛИРОВАНИЕ ДИАЛОГА В ПСИХОЛИНГВИСТИКЕ: ВЗРОСЛЫЕ И ДЕТСКИЕ СТРАТЕГИИ ОПИСАНИЯ ОБЪЕКТОВ ДЕЙСТВИТЕЛЬНОСТИ[1]

Федорова О. В. (olga.fedorova@msu.ru),
Деликишкина Е. А. (skaista_diena@mail.ru),
Слабодкина Т. А. (goodword@yandex.ru),
Ципенко А. А. (eire.morrigan@gmail.com)

Московский государственный университет
имени М. В. Ломоносова, Москва, Россия

Ключевые слова: диалог, онтогенез, коммуникативные стратегии, танграммы

DIALOGUE MODELLING IN PSYCHOLINGUISTICS: LITERAL AND ANALOGICAL PERSPECTIVES AS BASES FOR ADULTS' AND CHILDREN'S REFERENCES

Fedorova O. V. (olga.fedorova@msu.ru),
Delikishkina E. A. (skaista_diena@mail.ru),
Slabodkina T. A. (goodword@yandex.ru),
Tsipenko A. A. (eire.morrigan@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

Dialogue is a fundamental part of language use. In search of systematic evidence how the dialogue mechanisms work we turn to the referential communication task originally devised by R. Krauss and specified by H. Clark. In our experiment, two students or children were seated at tables separated by an opaque screen, in front of each were 12 cards of so-called Tangram

[1] Работа выполнена при частичной финансовой поддержке РГНФ (проект № 11-04-00153) и РФФИ (проект № 12-06-00268). Авторы выражают благодарность учителю русского языка и литературы лицея № 1564 Н. Н. Ципенко за неоценимую помощь в организации эксперимента с подростками, а также Е. О. Розановой за помощь в расшифровке подростковых записей.

figures. For the Director the cards were already arranged in a target sequence, and for the Matcher the same figures lay in a random sequence. The Director's job was to get the Matcher to rearrange his or her figures to match the target ordering. They carried out the task in four trials. All conversations (36 adults' and 8 children's dialogues) were transcribed, including changes of speaker, back-channel responses, hesitations, and false starts. We consider a prediction proposed by H. Clark that people prefer analogical perspective, which focuses on the resemblances of the figures to natural objects, to literal perspective, which focuses on the literal features of the objects, i.e. their geometric parts. Our results confirm the hypothesis; we also describe some peculiarities of the child dialogue strategies.

Key words: dialogue, ontogeny, communicative strategies, tangrams

1. Экспериментальное изучение диалога в психолингвистике

Как известно, лингвистика XX века была по большей части ориентирована на изучение компетенции, а не употребления. Подобное положение дел привело к тому, что к концу XX века психолингвистическая традиция 'язык как продукт' (подробнее об этом делении см. книгу Clark 1992), направленная на исследование компетенции абстрактного носителя языка независимо от контекста, в котором были употреблены изучаемые слова или предложения, оказалась несравнимо более развитой и успешной, чем традиция 'язык как действие', направленная на исследование функционирования языка в процессе реальной коммуникации. Возможно, именно по этой причине до самого последнего времени психолингвисты не проявляли особого интереса к моделированию языкового взаимодействия в диалоге, довольствуясь традиционной математической моделью, разработанной еще в середине XX века (Shannon, Weaver 1949).

Первым крупным психолингвистом, который привлек внимание коллег к изучению не только отдельных слов и предложений, но и естественных диалогов, возникающих в процессе совместного решения испытуемыми некоторой мыслительной задачи, стал Г. Кларк. В своих работах последней четверти XX века он предложил Совместную модель взаимодействия в диалоге, основанную на формировании собеседниками общей позиции, опирающейся на моделирование чужого сознания.

С момента создания традиции 'язык как действие' ее прототипическим методом исследования является ***метод референциальной коммуникации*** (англ. Referential communication task), введенный в практику в 70-ые годы XX века социальным психологом Р. Краусом. Основная идея этого метода состоит в том, что один из собеседников, Инструктор (англ. Director) видит или знает нечто, что он должен вербально передать второму собеседнику, Раскладчику (англ. Matcher), который этого не видит или не знает. Существует два способа проведения подобных экспериментов — через специальный экран и по телефону, а также два типа задания — испытуемый должен пройти определенным путем по лабиринту (методика 'лабиринт') или разложить

в правильном порядке предметы, первоначально лежащие в беспорядке или в неправильном порядке (методика ‘беспорядок’). Отметим, что в том виде, в котором данная методика была разработана Краусcом (в частности, см. Krauss, Weinheimer 1966), в настоящее время она уже не используется. Большой вклад в усовершенствование методики ‘беспорядок’ внес Г. Кларк, опубликовав в конце XX века целую серию работ с так называемыми ***танграммами***[2] — фигурками из китайской игры-головоломки (см. рис. 1), которые с трудом поддаются вербальному описанию.

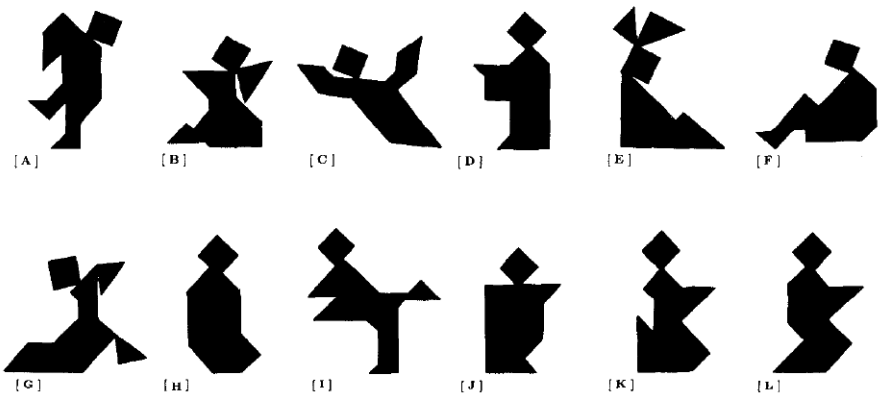


Рис. 1. Стимульный материал исследования из пионерской работы Clark, Wilkes-Gibbs 1986. Для удобства дальнейшего изложения ниже приведены условные наименования каждой танграммы: [A] — ‘медведь с топором’, [B] — ‘мышка с бантиком’, [C] — ‘привидение’, [D] — ‘монах с квадратным рукавом’, [E] — ‘зайчик у стенки’, [F] — ‘студент у стенки’, [G] — ‘чеченская танцовщица’, [H] — ‘кокон’, [I] — ‘фигурист’, [J] — ‘кавказец в бурке’, [K] — ‘горбатая японка с хвостиком’, [L] — ‘японка с ногами зигзагом’

Танграммы — это популярная игра-головоломка, которая состоит из семи частей особым образом разрезанного квадрата. Эти семь частей нужно сложить для получения другой, обычно более сложной, фигуры, которая изображает человека, животное, предмет домашнего обихода и под. Фигуру, которую необходимо получить, традиционно задают в виде внешнего контура, скрывая границы между частями. Своей популярностью в XX веке танграммы во многом обязаны известной книге Сэма Лойда «Восьмая книга Тан», вышедшей в 1903 году. Кроме 700 задач, некоторые из которых оказались неразрешимы, книга содержит

[2] Насколько нам известно, существует два параллельно употребляющихся русских перевода английской лексемы *tangram* — как в мужском роде (*танграм*), так и в женском (*танграмма*); мы будем использовать второй из них.

вымышленную историю головоломки, согласно которой эта игра была изобретена 4 тысячи лет назад божеством по имени Тан. На самом деле первые печатные упоминания о танграммах встречаются в китайской книге, изданной только в 1813 году, и у современных исследователей нет оснований предполагать ее более раннее происхождение. Легенда Лойда, однако, считается одной из самых известных мистификаций XX века, так что идея о древнем происхождении танграмм до сих пор периодически появляется в литературе.

Насколько нам известно, в научных целях танграммы были впервые использованы в в исследовании, описанном в работе Clark, Wilkes-Gibbs 1986. Процедура эксперимента состояла в следующем: перед обоими участниками эксперимента, которые сидели за соседними столами, но при этом не видели друг друга благодаря специальному экрану-перегородке, был одинаковый набор из 12 танграмм (см. рис. 1), однако перед Инструктором фигурки были расположены в одном порядке, а перед Раскладчиком — в другом. Целью Инструктора было объяснить Раскладчику свой порядок, а целью Раскладчика — воспроизвести на своем столе порядок расположения фигурок на столе Инструктора. У каждой из восьми пар испытуемых было шесть попыток, в каждой из которых набор фигурок оставался таким же, менялось только их взаимное расположение; в ходе эксперимента собеседники не менялись ролями — один из них всегда выступал в роли Инструктора, а второй — в роли Раскладчика. Все диалоги записывались на аудиоаппаратуру, а затем расшифровывались. В результате проведенных исследований авторами был собран корпус диалогов, состоящий из описаний 576 танграмм (=12 фигур по 6 попыток в каждой из 8 пар испытуемых). Транскрипты включали не только текстовую расшифровку, но также паузы хезитации, фальстарты и маркеры обратной связи. Корпус состоял из 9792 слов.

Совместная модель референции, предложенная в работе Clark, Wilkes-Gibbs 1986, позволила авторам сделать несколько важных предсказаний, которые подтвердились в ходе анализа результатов исследования. В частности, авторы предположили, что с каждой последующей попыткой количество необходимых слов (как и количество необходимых реплик) на описание одной танграммы будет уменьшаться, а сами описания будут модифицироваться таким образом, что количество нестандартных номинаций будет уменьшаться, а количество так называемых ‘ярлыков’ (англ. label), придуманных совместными вербальными усилиями Инструктора и Раскладчика, будет, наоборот, увеличиваться.

В нашем исследовании с русскоязычными испытуемыми, проведенном осенью 2009 — зимой 2010 годов, у каждой пары испытуемых было четыре попытки, в ходе выполнения заданий они также 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. Кроме того, весной 2010 года мы дополнили этот корпус данными, собранными в ходе исследования шести пар пациентов Центра патологии речи и нейрореабилитации[3],

[3] Данная работа была проделана совместно с сотрудниками Центра патологии речи и нейрореабилитации под руководством О. В. Драгой.

а в январе 2013 года провели новое исследование с 8 парами подростков возраста 11–12 лет[4]. Полученный корпус представляет собой, на наш взгляд, практически неисчерпаемый источник данных о самых разных особенностях диалогового поведения в норме, при патологии и в онтогенезе. В настоящей работе мы рассмотрим один такой конкретный вопрос, связанный со стратегиями описания танграмм взрослыми и подростками.

2. Стратегии описания: перспектива по аналогии vs. буквальная перспектива

В работе Кларка и Уилкес-Гиббс (1986) авторы рассматривают две возможные перспективы описания окружающих нас объектов действительности — холистическую (англ. holistic), при которой мы воспринимаем объект как единое целое, и сегментированную (англ. segmental), когда объект представляется нам состоящим из нескольких соположенных друг с другом сегментов. Согласно мнению авторов, в общем случае холистическая перспектива описания объектов действительности является более предпочтительной и чаще заканчивается успешной коммуникацией.

Говоря о стратегиях описания танграмм, Кларк и Уилкес-Гиббс используют термины ***перспектива по аналогии*** (англ. analogical perspective), который в общих чертах совпадает с холистической перспективой, и ***буквальная перспектива*** (англ. literal perspective), совпадающая с сегментированной перспективой (Clark, Wilkes-Gibbs 1986: 30). Приводя примеры конкретных описаний (см. пример 1 для перспективы по аналогии и пример 2 для буквальной перспективы), авторы отмечают, что в 84% номинаций, вводимых перспективой по аналогии, используются лексемы *looks like* или *resembles*, а в 89% номинаций, вводимых буквальной перспективой, используются вспомогательные глаголы *is* или *has*.

(1) танграмма [G] ‘чеченская танцовщица’
Number 5 looks like a girl dancing sort of.

(2) танграмма [J] ‘кавказец в бурке’
Number 1 has a diamond on the top and a square. The left side is, urn, like a rectangle shape, and the right side is cut off.

Согласно данным Кларка и Уилкес-Гиббс (1986), в первой попытке испытуемые обычно начинали свое описание с перспективы по аналогии, а потом

[4] Строго говоря, биологическое, психологическое и когнитивное (в том числе языковое) развитие ребенка принято делить на три больших этапа: (1) раннее детство (от 2 до 6 лет, или дошкольный период), (2) среднее детство (от 6 до 12 лет, или предподростковый период) и (3) отрочество (12–16 лет, или подростковый период), каждый из которых характеризуется своими особенностями. Таким образом, дети рассматриваемого нами возраста переходят из предподросткового периода в подростковый.

в случае необходимости добавляли конкретные геометрические детали — таким образом устроено большинство описаний танграмм в исследовании Clark, Wilkes-Gibbs 1986. Что касается соотношения перспектив, то если в первой попытке 42% всех танграмм описывались с точки зрения обеих перспектив, то к шестой попытке процент таких случаев уменьшился до 19. Таким образом, от первой к последующим попыткам собеседники устанавливают общую языковую почву, используя все больше 'ярлыков', в которых обычно или совсем не задействуются элементы буквальной перспективы или они задействуются минимально.

Данные русскоязычного исследования оказались в целом похожи на англоязычные результаты: количество использования обеих перспектив (см. пример 3) уменьшилось с 52% в первой попытке до 10% в четвертой попытке.

(3) танграмма [G] 'чеченская танцовщица', попытка 1[5]
И: *Ммм, дальше, следующая картинка это ну что-то такое тоже как будто это какая-то женщина в платье, ну то= то есть если образно так говорить. У неё эээ сз= сл= с правой стороны два таких маленьких треугольничка.*
Р: *Угу.*
И: *Вот. Значит, ещ= у неё гол= голова так скажем, да, она в левую сторону наклонена, а в правую сторону у неё два таких маленьких треугольничка и основное эээ как бы корпус её он в пра= в левой части находится картинки.*
Р: *Угу.*
И: *Ну эээ похоже ещё на башмак какой-то, не знаю, если так смотреть.*
Р: *Угу, поняла.*

Итак, большинство взрослых испытуемых в результате совместной когнитивной деятельности в первой попытке придумывали некоторый 'ярлык', который они использовали в последующих попытках для почти безошибочной идентификации танграмм. Важно отметить также, что и в англоязычном, и в русскоязычном экспериментах процент описаний, при которых взрослые испытуемые ограничивались одной буквальной перспективой, см. пример 2, а также пример 9 ниже, совсем низок (меньше 3% в исследовании Кларка и менее 1% в русскоязычном исследовании).

Обратимся теперь к результатам эксперимента с подростками, см. табл. 1 ниже. Как видно из представленных данных, с точки зрения использования 'ярлыков' подростковая кривая оказывается очень похожей на аналогичную взрослую кривую. Однако важной отличительной особенностью подростковых диалогов является устойчивое, не зависящее от номера попытки, наличие около 10% описаний танграмм с использованием только буквальной перспективы, см. пример 4.

---

[5] В транскриптах отмечены заполненные паузы хезитации (*ммм* или эээ); знак равенства обозначает обрыв слова.

Табл. 1. Стратегии описания танграмм в эксперименте с подростками.

Перспектива (в %) / номер попытки	**1**	**2**	**3**	**4**
Обе перспективы	45	36	25	16
Только перспектива по аналогии	45	50	64	74
Только буквальная перспектива	10	14	11	10

(4) танграмма [E] ‘зайчик у стенки’, попытка 1
Значит, там дальше эээ квадрат наклонен вправо, от него вверх два треугольника в разные стороны, потом идет большой треугольник снизу, и от него еще такой выступ как бы.

Характерно, что буквальную перспективу подростки обычно используют при описании самых сложных танграмм. Данная закономерность характерна и для взрослых диалогов, однако сравним коммуникативное поведение подростков и взрослых испытуемых в подобных случаях. Сначала рассмотрим все четыре попытки описания танграммы [G] ‘чеченская танцовщица’ в одном из диалогов подростков:

(5) танграмма [G] ‘чеченская танцовщица’, попытка 1
И: *Квадрат, это я не знаю ни на что не похоже, ммм квадрат дальше идет, немного похоже на треугольник, дальше идет вниз там прямоугольник наклоненный вправо, и он вместе с трапецией которая прямо стоит, ммм нашла?*
Р: *Неа.*
И: *Это как-то это похоже еще сзади там как бы один треугольник, и еще пониже второй треугольник.*
Р: *Угу, есть.*

(6) танграмма [G] ‘чеченская танцовщица’, попытка 2
Эээ дальше, там квадрат эээ от него идет справа треугольник, от него еще один треугольник маленький.

(7) танграмма [G] ‘чеченская танцовщица’, попытка 3
И: *Дальше, это уже второй ряд, эээ квадрат, ммм справа от него треугольник.*
Р: *Угу.*
И: *Справа и чуть внизу треугольник от этого треугольника справа еще треугольник, потом ниже идет эээ основание и из прямоугольника наклоненного вправо, а потом у него трапеция как бы. Нашла?*
Р: *Угу.*

В перерыве перед четвертой попыткой Раскладчик сказал Инструктору, показывая на данную танграмму: «*Эту фигуру трудно различать. Давай это будет башмак с каблуком*». Реплика Инструктора в последней попытке звучит, соответственно, таким образом:

(8) танграмма [G] ‘чеченская танцовщица’, попытка 4
Третья это башмак с каблуком.

Теперь обратимся к взрослым описаниям. Сначала рассмотрим две первые попытки описания танграммы [B] ‘мышка с бантиком’, а затем первую попытку описания танграммы [G] ‘чеченская танцовщица’.

(9) танграмма [B] ‘мышка с бантиком’, попытка 1
И: *Самая непонятная картинка. Там сверху ааа как бы в одной точке, то есть в одной вершине пересекаются квадрат, ну ромб, треугольник и кусок большого треугольника.*
Р: *Но там один треугольник или…*
И: *Ну формально, целый треугольник, отдельно стоящий, там один. Второй треугольник как бы вливается в общую фигуру. И еще у него, еще у него, еще у него…*
Р: *То есть большой квадрат и треугольник?*
И: *Еще у него, еще у него кривое основание. Если ты обратишь внимание, у него…*
Р: *Окей, понятно.*

(10) танграмма [B] ‘мышка с бантиком’, попытка 2
И: *Ааа вторая картинка, первый ряд. Ну вот сейчас мне кажется, что это похоже на мышь, на самом деле, но это вот такой один из наименее похожих на человека… один из наименее похожих на человека изображений. В одной точке ммм ромб, треугольник и часть большого треугольника вокруг одной точки.*
Р: *Угу.*

(11) танграмма [G] ‘чеченская танцовщица’, попытка 1
И: *Значит, второй ряд. Первая картинка, что-то непонятное совершенно. Вот из пят= из шести картинок, которые остались, выбери наименее похожее на человека. Это будет первая картинка во втором ряду.*
Р: *Он похож на туфлю без задней части, у кот=*
И: *Да-да-да!*
Р: *сверху…*
И: *Да-да-да.*

Таким образом, мы видим, что подростки и взрослые испытывают однотипные трудности с выбором номинации в случае с самыми сложными танграммами — описания подобных случаев при помощи буквальной перспективы занимают больше времени и сопровождаются ошибками (в частности, танграмма из примера 7 была идентифицирована Раскладчиком неправильно). Однако существенные различия между взрослыми и подростковыми стратегиями заключаются, на наш взгляд, в том, что (1) во взрослых описаниях такие случаи являются единичными; (2) взрослые так или иначе справляются

с этими трудностями уже ко второй попытке, в то время как подростковые записи демонстрируют недостаточные коммуникативно-дискурсивные навыки ведения диалога в наиболее сложных коммуникативных ситуациях.

Данный вывод хорошо согласуется с гипотезой позднего дискурсивного развития детей, которая была выдвинута К. Ф. Седовым: 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). К. Ф. Седов пишет, что становление языковой личности продолжается и в школьном возрасте, когда индивид постепенно повышает уровень своей дискурсивной компетенции, а полное овладение коммуникативно-дискурсивными механизмами происходит только в старшем подростковом возрасте (к 15–16 годам).

Литература

1. *Clark H. H.* (1992), Arenas of Language Use. Chicago, University of Chicago Press.
2. *Clark H. H., Wilkes-Gibbs D.* (1986), Referring as a collaborative process, Cognition, Vol. 22(1), pp. 1–39.
3. *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.
4. *Sedov K. F.* (2004), Discourse and personality [Diskurs i lichnost’], Moscow, Labirint.
5. *Shannon C. E., Weaver W.* (1949), The Mathematical theory of communication, Urbana, IL, University of Illinois Press.

ЧАЩА СИНТАКСИЧЕСКОГО РАЗБОРА ДЛЯ АБЗАЦА ТЕКСТА

Галицкий Б. (boris.galitsky@ebay.com),
Иворский Д. (dilv_ru@yahoo.com),
Кузнецов С. (skuznetsov@hse.ru),
Строк Ф. (fdr.strok@gmail.com)

eBay Inc.; Национальный исследовательский университет, Высшая школа экономики, Москва, Россия

Мы разрабатываем технику представления структуры предложений и абзацев текста в виде графов. Мы определяем чащу синтаксического разбора как объединение синтаксических деревьев разбора предложений. Чаща включает дуги между вершинами синтаксических деревьев для таких отношений, как кореферентность и таксономия. Эти дуги также получаются из других источников, в том числе, теории Риторических Структур и Речевых Актов. В работе предлагается алгоритм вычисления чащ разбора. Также в работе рассматриваются программные средства, предназначенные для построения чащ разбора и выполнения операции обобщения (пересечения) чащ разбора. На основе рассматриваемого подхода проводятся вычислительные эксперименты по улучшению поиска в случае, когда запрос представлен несколькими предложениями. Производится сравнение базового поиска, поиска с помощью сопоставления отдельных предложений и поиска с использованием Чащ разбора.

PARSE THICKET REPRESENTATIONS OF TEXT PARAGRAPHS

Galitsky B. (boris.galitsky@ebay.com),
Ivovsky D. (dilv_ru@yahoo.com),
Kuznetsov S. (skuznetsov@hse.ru),
Strok F. (fdr.strok@gmail.com)

eBay Inc.; National Research University Higher School of Economics, Moscow, Russia

We develop a graph representation and learning technique for parse structures for sentences and paragraphs of text. We introduce parse thicket as a set of syntactic parse trees augmented by a number of arcs for inter-sentence word-word relations such as coreference and taxonomies. These arcs are also derived from other sources, including Rhetoric Structure and Speech Act theory. We introduce respective indexing rules that identify inter-sentence relations and join phrases connected by these relations in the search index. We propose an algorithm for computing parse thickets from parse trees. We develop a framework for automatic building and generalizing of parse thickets. The proposed approach is used for evaluation in the product search where search queries include multiple sentences. We draw the comparison for search relevance improvement by pair-wise sentence generalization and thicket-level generalization.

Keywords: learning taxonomy, learning syntactic parse tree, syntactic generalization, search relevance, paragraph matching

1. Introduction

Parse trees have become a standard form of representing the linguistic structures of sentences. In this study we will attempt to represent a linguistic structure of a *text paragraph* based on parse trees for each sentence of this paragraph. We will refer to the set of parse trees plus a number of arcs for inter-sentence relations between nodes for words as Parse Thicket (PT). A PT is a graph that includes parse trees for each sentence, as well as additional arcs for inter-sentence relationship between parse tree nodes for words.

In this paper we will define the operation of *generalization of text paragraphs* to assess similarity between portions of text. The use of generalization for similarity assessment is inspired by structured approaches to machine learning versus unstructured, statistical approaches where similarity is measured by a distance in feature space. Our intention is to extend the operation of the least general generalization (antiunification of logical formula)[14] towards structural representations of paragraph of texts. Hence we will define the operation of generalization on Parse Thickets and outline an algorithm for it.

This generalization operation is a base for number of text analysis applications such as search, classification, categorization, and content generation [9]. *Generalization of text paragraphs* is based on the operation of generalization of two sentences explored in our earlier studies [8]. In addition to learning generalizations of individual sentences, in this paper we study how the links between words in sentences other than syntactic ones can be used to compute similarity between texts. To compute generalization of a pair of paragraph, we performed a pair-wise generalization for each sentence in paragraphs. This approach ignores the richness of coreference information, and in the current study we develop graph-learning means, specifically oriented to represent paragraphs of text as respective PTs with nodes interconnected by arcs for a number of relations including coreference and taxonomy relations. We also consider such discourse-related theories as Rhetoric Structure (RST) [24] and Communicative Actions (CA) [23] as a source of arcs to augment PTs. These arcs will connect nodes for words both within and between parse trees for sentences.

It is significant to note that we used "out of the box" tools for constructing parse trees for sentences. For evaluation we used OpenNLP [16] and Stanford NLP [27] frameworks, which are intended for working with constituency-based trees. Also we used ETAP-3 system [20, 25, 26], built for working with dependency-based trees. This system was applied only to construct and visualize syntactic trees for sentences from basic examples. More details about programming components of our framework will be given in evaluation section.

2. Introducing Parse Thicket

Is it possible to find more commonalities between texts treating parse trees at a higher level? For that we need to extend the syntactic relations between the nodes of the syntactic dependency parse trees towards more general text discourse relations.

What relations can we add to the sets of parse trees to extend the match? Once we have such relations as "the same entity", "sub-entity", "super-entity" and anaphora, we can extend the notion of phrase to be matched between texts. Relations between the nodes of parse trees that are other than syntactic can merge phrases from different sentences or from a single sentence which are not syntactically connected.

If we have two parse trees P_1 and P_2 of text T_1, and an arc for a relation r

r: $P_{1j} \rightarrow P_{2j}$ between the nodes P_{1j} and P_{2j}, we can match …,$P_{1,i-2}$, $P_{1,\,i-1}$, $P_{1,\,i}$, $P_{2,j}$, $P_{2,j+1}$, $P_{2,j+2}$, … of T_1 against a chunk of a single sentence of merged chunks of multiple sentences from T_2.

2.1. Finding similarity between two paragraphs of text

There are several approaches to assessing the similarity of text paragraphs:

- Baseline: bag-of-words approach, which computes the set of common keywords/n-grams and their frequencies.
- Pair-wise matching: syntactic generalization to each pair of sentences, and summing up the resultant commonalities. This technique has been developed in our previous work [9].
- Paragraph-paragraph matching.

The first approach is most typical for industrial NLP applications today, and the second is the one of our previous studies. Kernel-based approach to parse tree similarities [13, 22], as well as tree sequence kernel [21], being tuned to parse trees of individual sentences, also belongs to the second approach.

Let us consider a short example to compare the above three approaches. Fragments of this example will be used through the whole paper. The generalization operation is denoted by '^':

"Iran refuses to accept the UN proposal to end the dispute over work on nuclear weapons",

"UN nuclear watchdog passes a resolution condemning Iran for developing a second uranium enrichment site in secret",

"A recent IAEA report presented diagrams that suggested Iran was secretly working on nuclear weapons",

"Iran envoy says its nuclear development is for peaceful purpose, and the material evidence against it has been fabricated by the US",

^

"UN passes a resolution condemning the work of Iran on nuclear weapons, in spite of Iran claims that its nuclear research is for peaceful purpose",

"Envoy of Iran to IAEA proceeds with the dispute over its nuclear program and develops an enrichment site in secret",

"Iran confirms that the evidence of its nuclear weapons program is fabricated by the US and proceeds with the second uranium enrichment site"

The list of common keywords gives a hint that both documents are on nuclear program of Iran, however it is hard to get more specific details

Iran, UN, proposal, dispute, nuclear, weapons, passes, resolution, developing, enrichment, site, secret, condemning, second, uranium

Pair-wise generalization gives a more accurate account on what is common between these texts:

[NN-work IN-* IN-on JJ-nuclear NNS-weapons], [DT-the NN-dispute IN-over JJ-nuclear NNS-*], [VBZ-passes DT-a NN-resolution],
[VBG-condemning NNP-iran IN-*],
[VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]],
[DT-* JJ-second NN-uranium NN-enrichment NN-site]],
[VBZ-is IN-for JJ-peaceful NN-purpose],
[DT-the NN-evidence IN-* PRP-it], [VBN-* VBN-fabricated IN-by DT-the NNP-us]

Parse Thicket generalization gives the detailed similarity picture which looks more complete than the pair-wise sentence generalization result above:

[NN-Iran VBG-developing DT-* NN-enrichment NN-site IN-in NN-secret]
[NN-generalization-<UN/nuclear watchdog> * VB-pass NN-resolution VBG condemning NN- Iran]
[NN-generalization-<Iran/envoy of Iran> Communicative_action DT-the NN-dispute IN-over JJ-nuclear NNS-*
[Communicative_action — NN-work IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]
[NN-generalization <Iran/envoy to UN> Communicative_action NN-Iran NN-nuclear NN-* VBZ-is IN-for JJ-peaceful NN-purpose],
Communicative_action — NN-generalize <work/develop> IN-of NN-Iran IN-on JJ-nuclear NNS-weapons]*
[NN-generalization <Iran/envoy to UN> Communicative_action NN-evidence IN-against NN Iran NN-nuclear VBN-fabricated IN-by DT-the NNP-us]
condemn^proceed [*enrichment site*] <leads to> *suggest^condemn* [*work Iran nuclear weapon*]

One can feel that PT-based generalization closely approaches human performance in terms of finding similarities between texts. To obtain these results, we need to be capable of maintaining coreferences, apply the relationships between entities to our analysis *(subject vs relation-to-this subject),* including relationships between verbs *(develop* is a partial case of *work).* We also need to be able to identify communicative actions and generalize them together with their subjects according to the specific patterns of speech act theory. Moreover, we need to maintain rhetoric structure relationships between sentences to generalize at a higher level above sentences.

The focus of this paper will be to introduce parse thicket and their generalization as paragraph-level structured representation. It will be done with the help of the above example. Fig. 1 and Fig. 2 show the dependency-based parse trees for the above texts T_1 and T_2. Each tree node has labels as part-of-speech and its form (such as SG for 'single'); also, tree edges are labeled with the syntactic connection type (such as 'composite'). Source trees were built and visualized using ETAP-3 system [20, 25, 26]. Then we added specific "red" arcs to them in order to illustrate the idea of parse thicket.

Generalization of parse thickets, being the set of maximal common sub-graph (sub-parse thicket) can be computed at the level of phrases as well, as the set of maximal common sub-phrases. However, the notion of phrases is extended now: *thicket phrases* can contain regular phrases from different sentences. The way these phrases are extracted and formed depends on the source of non-syntactic link between words in different sentences: thicket phrases are formed in a different way for communicative actions and RST relations.

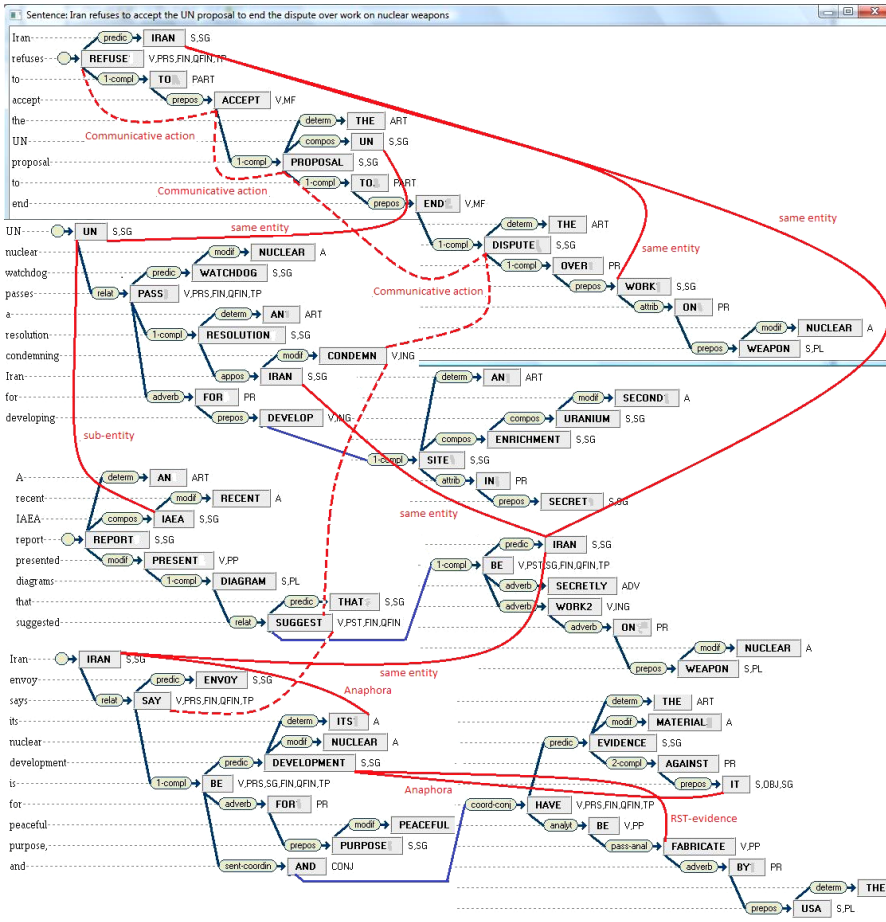


Fig. 1: Parse thicket for text T_1

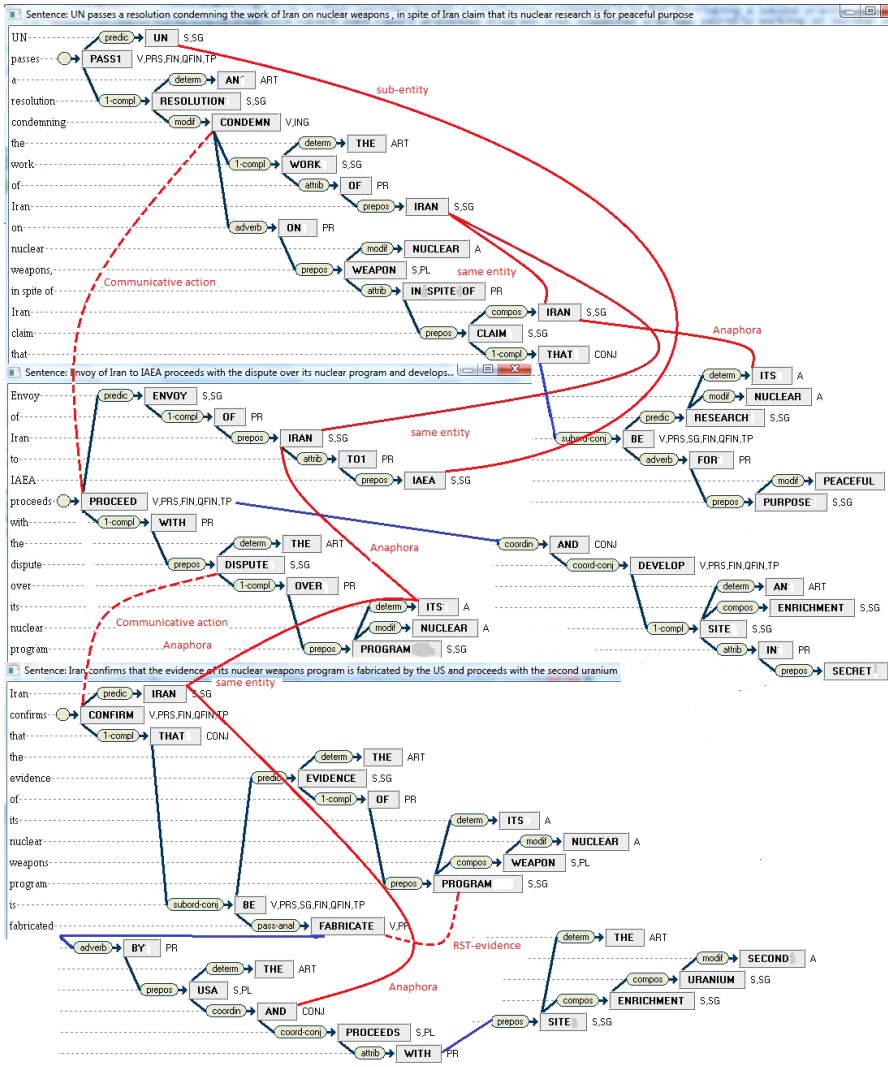


Fig. 2: Parse thicket for text T_2

3. Arcs of parse thicket based on theories of discourse

Using the unified framework we develop two approaches to textual discourse based on

- Rhetoric structure theory (RST) [24],
- Communicative Actions (CA) [23].

We used a vocabulary of Communicative actions to

1. find their subjects [23],
2. add respective arcs to the parse thicket,
3. index combination of phrases as subjects of communicative actions

For RST, we introduce explicit indexing rules which will be applied to each paragraph and

1. attempt to extract an RST relation,
2. build corresponding fragment of the parse thicket, and
3. index respective combination of formed phrases (noun, verb, prepositional), including words from different sentences.

3.1. Generalization based on Rhetoric structure arcs

The theory of Rhetoric structures (RST)[24] was developed to explain the coherence of texts, seen as a kind of function, linking parts of a text to each other.

Two connected clouds represented on the right of Fig.3 show the generalization instance based on RST relation "RCT-evidence". This relation occurs between the phrases

evidence-for-what [Iran's nuclear weapon program] and *what-happens-with-evidence [Fabricated by USA]* on the right-bottom, and

evidence-for-what [against Iran's nuclear development] and *what-happens-with-evidence [Fabricated by the USA]* on the right-top.

Notice that in the latter case we need to merge (perform anaphora substitution) the phrases *'its nuclear development'* and *'evidence against it'* to obtain *'evidence against its nuclear development'*. Notice the arc *it — development,* according to which this anaphora substitution occurred. *Evidence* is removed from the phrase because it is the indicator of RST relation, and we form the subject of this relation to match. Furthermore, we need another anaphora substitution *its — Iran* to obtain the final phrase.

As a result of generalizations of two RST relations of the same sort (evidence) we obtain

Iran nuclear NNP — RST-evidence — fabricate by USA.

Notice that we could not obtain this similarity expression by using sentence-level generalization.

Green clouds indicate the sub-PTs of T_1 and T_2, which are matched. We show three instances of PT generalization.

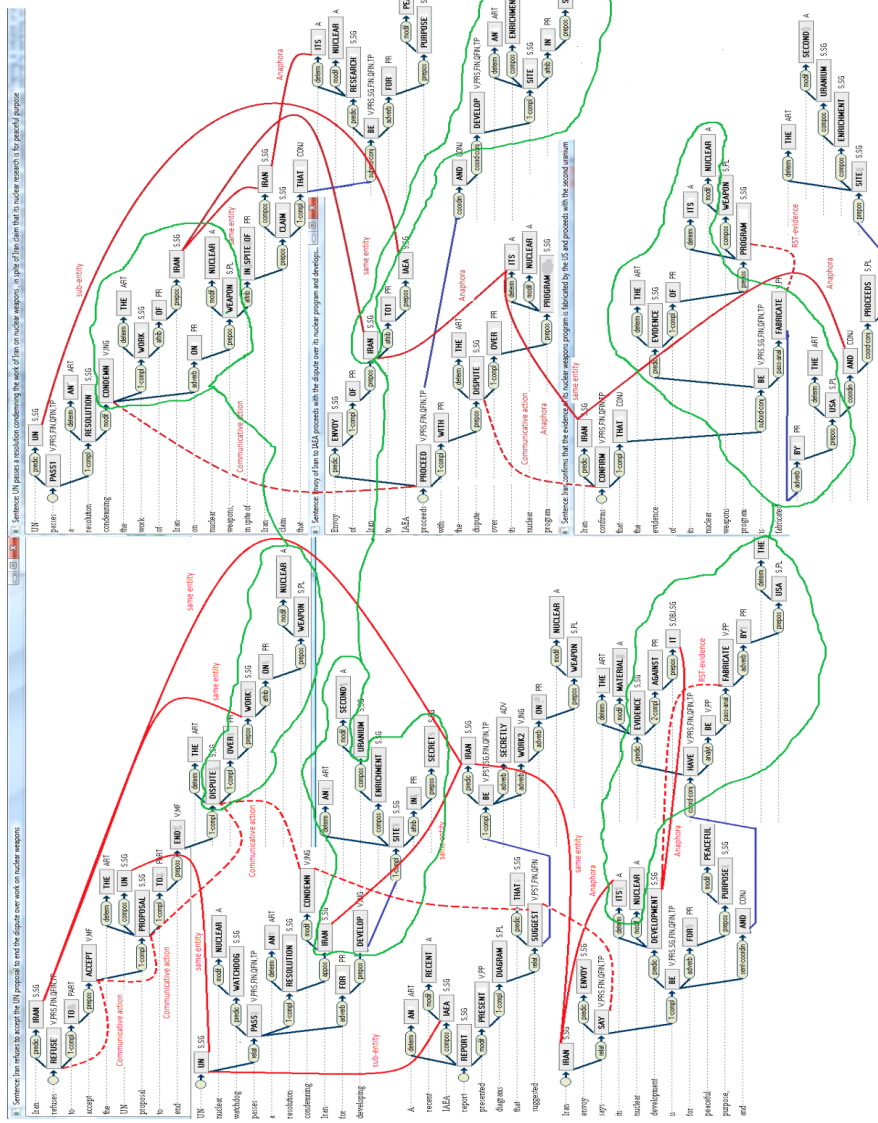


Fig. 3: Three instances of matching between sub-PTs shown as connected clouds

3.2. Generalization based on Communicative action arcs

Communicative actions (CA) are used by text authors to indicate a structure of a dialogue or a conflict [23]. Hence, analyzing the arcs of communicative actions of PT, one can find implicit similarities between texts. We can generalize

1. one communicative action with its subject from T_1 against another communicative action with its subject from T_2 (communicative action arc is not used) ;
2. a pair of communicative actions with their subjects from T_1 against another pair of communicative actions from T_2 (communicative action arcs are used).

In our example, we have the same communicative actions with subjects with low similarity:

> *condemn* [‘*Iran for developing second enrichment site in secret’*] vs *condemn* [‘*the work of Iran on nuclear weapon*’] or different communicative actions with similar subjects.

Two communicative actions can always be generalized, which is not the case for their subjects: if their generalization result is empty, the generalization result of communicative actions with these subjects is empty too. The generalization result here for the case 1 above is:

> *condemn^dispute [work-Iran-on-nuclear-weapon].*

Generalizing two different communicative actions is based on their attributes and is presented elsewhere [7].

T_1		T_2
condemn [second uranium enrichment site]	↔	*proceed [develop an enrichment site in secret]*
↓ communicative action arcs		↓
suggest [Iran is secretly working on nuclear weapon]	↔	*condemn [the work of Iran on nuclear weapon]*

which results in

condemn^proceed [*enrichment site*] <leads to> *suggest^condemn* [*work Iran nuclear weapon*]

Notice that generalization

condemn [second uranium enrichment site]	↔	*condemn [the work of Iran on nuclear weapon]*
↓ communicative action arcs		↓
suggest [Iran is secretly working on nuclear weapon]	↔	*proceed [develop an enrichment site in secret]*

gives zero result because the arguments of *condemn* from T_1 and T_2 are not very similar. Hence we generalize the subjects of communicative actions first before we generalize communicative actions themselves.

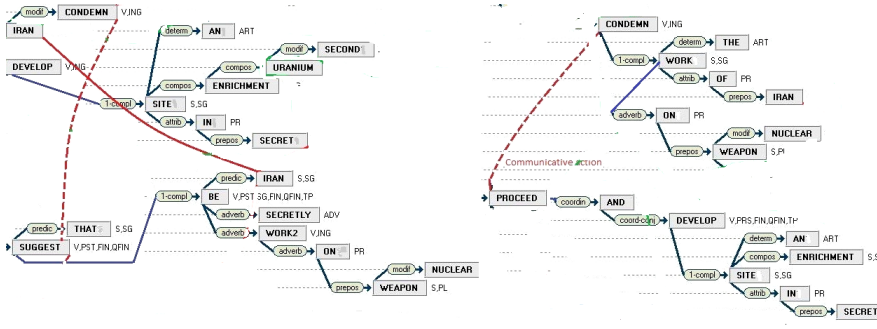


Fig. 4: A fragment of PT showing the mapping for the pairs of communicative actions

4. Generalization of thickets

4.1. Definition of generalization operation on two parse thickets

Given two parse thickets $C_x=(V_x, E_x)$ and $C_y=(V_y, E_y)$, their generalization denoted by $C_x \hat{} C_y$ is defined as the set $\{G_1, G_2,\ldots,G_k\}$ of all inclusion-maximal common subgraphs of C_x and C_y. See e.g. [28].

4.2. Algorithm for forming thicket phrases for generalization

We will now outline the algorithm of forming thicket phrases.

For each sentence S in a paragraph P
  Form a list of previous sentences in a paragraph S_{prev}
  For each word in the current sentence:
    - If this word is a *pronoun*: find all nouns or noun phrases in the S_{prev}, which are
      * The same entities (via anaphora resolution)
    - If this word is a *noun*: find all nouns or noun phrases in the S_{prev}, which are
      * The same entities (via anaphora resolution)
      * Synonymous entity
      * Super entities
      * Sub and sibling entities
    - If this word is a *verb*:
      * If it is a communicative action:
        Form the phrase for its subject $VBCA_{phrase}$, including its verb phrase VB_{phrase}
        Find a preceding communicative action $VBCA_{phrase0}$ from S_{prev} with its subject and form a thicket phrase $[VBCA_{phrase}, VBCA_{phrase0}]$
      * If it indicates RST relation
        Form the phrase for the pair of phrases, which are the subjects $[VBRST_{phrase1}, VBRST_{phrase2}]$, of this RST relation, $VBRST_{phrase1}$ belongs to S_{prev}.

4.3. Sentence-level generalization algorithm

Below we outline the algorithm on finding a maximal sub-phrase for a pair of phrases, applied to the sets of thicket phrases for T_1 and T_2.

1. Split parse trees for sentences into sub-trees which are phrases for each type: *verb, noun, prepositional* and others; these sub-trees are overlapping. The sub-trees are coded so that information about occurrence in the full tree is retained.
2. All sub-trees are grouped by phrase types.
3. Extending the list of phrases by adding equivalence transformations
4. Generalize each pair of sub-trees for both sentences for each phrase type.
5. For each pair of sub-trees yield an alignment, and then generalize each node for this alignment. For the obtained set of trees (generalization results), calculate the score.
6. For each pair of sub-trees for phrases, select the set of generalizations with highest score (least general).
7. Form the sets of generalizations for each phrase types whose elements are sets of generalizations for this type.
8. Filtering the list of generalization results: for the list of generalization for each phrase type, exclude more general elements from lists of generalization for given pair of phrases.

5. Evaluation of multi-sentence search

5.1. PT-processing framework

There are two system components, which include Parse Thicket building and phrase-level processing.

The textual input is subject to a conventional text processing flow such as sentence splitting, tokenizing, stemming, part-of-speech assignment, building of parse trees and coreferences assignment for each sentence. Unlike our previous studies [19] computing parse thickets becomes fully automatic and includes not only RST and CA arcs but also coreference (via anaphora resolution) and taxonomic (same-entity, sub-entity, super-entity) relations. This flow is implemented by either OpenNLP or Stanford NLP, and the parse thicket is built based on the algorithm presented in this paper. The coreferences and RST component strongly relies on Stanford NLP's rule-based approach to finding correlated mentions based on the multi-pass sieves.

Phrase-level processing for the phrases of individual sentences has been described in detail in our previous studies [8, 19]. In this study we collect all phrases for all sentences of one paragraph of text, augment them with thicket phrases (linguistic phrases which are merged based on the inter-sentence relation), and generalize against that of the other paragraph of text.

5.2. Evaluation results

Having described the system architecture and engineering aspects, we proceed to evaluation of how generalization of PTs can improve multi-sentence search, where one needs to compare a query as a paragraph of text against a candidate answer as a paragraph of text (snippet). We refer the reader to [8] for the details on evaluation settings.

Table 1: Evaluation results

Query type	Query complexity	Relevance of baseline Bing search, %, averaging over 100 searches	Relevance single-sentence phrase-based generalization search, %, averaging over 100 searches	Relevance of thicket-based phrase generalization search, %, averaging over 100 searches
Product recommendation search	1 compound sentence	62.30	69.10	72.40
	2 sentences	61.50	70.50	71.90
	3 sentences	59.90	66.20	72.00
	4 sentences	60.40	66.00	68.50
Travel recommendation search	1 compound sentence	64.80	68.00	72.60
	2 sentences	60.60	65.80	73.10
	3 sentences	62.30	66.10	70.90
	4 sentences	58.70	65.90	72.50
Facebook friend agent support search	1compound sentence	54.50	63.20	65.30
	2 sentences	52.30	60.90	62.10
	3 sentences	49.70	57.00	61.70
	4 sentences	50.90	58.30	62.00
Average		58.15	64.75	68.75

Evaluation results are shown in Table 1. Three domains are used in evaluation:

- Product recommendation, where an agent reads chats about products and finds relevant information on the web about a particular product.
- Travel recommendation, where an agent reads chats about travel and finds relevant information on the travel websites about a hotel or an activity.
- Facebook recommendation, where an agent reads wall postings and chats, and finds a piece of relevant information for friends on the web.

In each of these domains we selected a portion of text on the web to form a query, and then filtered search results delivered by Bing search engine API. One can observe that unfiltered precision is 58.2%, whereas improvement by pair-wise sentence generalization is 11%, thicket phrases give additional 6%. One can also see that the higher the complexity of sentence, the higher the contribution of generalization technology, from sentence level to thicket phrases.

6. Related work and conclusions

Usually, classical approaches to semantic inference rely on complex logical representations. However, practical applications usually adopt shallower lexical or lexical-syntactic representations, but lack a principled inference framework. Paper [2] proposed a generic semantic inference framework that operates directly on syntactic trees. New trees are inferred by applying entailment rules, which provide a unified representation for varying types of inferences. The current work deals with syntactic tree transformation in the graph-learning framework, treating various phrasings for the same meaning in a more unified and automated manner.

In our previous works we observed how employing a richer set of linguistic information such as syntactic relations between words assists relevance tasks [8, 9, 19]. To take advantage of semantic discourse information, we introduced parse thicket representation and proposed the way to compute similarity between texts based on generalization of parse thickets. In this work we build the framework for generalizing PTs as sets of phrases.

The operation of generalization to learn from parse trees for a pair of sentences turned out to be important for text relevance tasks. Once we extended it to learning parse thickets for two paragraphs, we observed that the relevance is further increased compared to the baseline (Bing search engine API), which relies on keyword statistics in the case of multi-sentence query. Parse thicket is intended to represent the syntactic structure of text as well as a number of semantic relations for indexing. To this end, a parse thicket contains relations between words in different sentences, such that these relations are essential to match queries with portions of texts to serve as answers.

We considered the following sources of relations between words in sentences: coreferences, taxonomic relations such as sub-entity, partial case, predicate for subject etc., rhetoric structure relation and speech acts. We demonstrated that search relevance can be improved if search results are subject to confirmation by parse thicket generalization, when answers occur in multiple sentences.

Using semantic information for query ranking has been proposed in [1]. Moreover, relying on matching of parse trees of a question and an answer has been the subject of [13] and [15]. However we believe that our study improves multi-sentence search, relying both on learning with parse tickets as connected parse trees and on linguistic theories on text coherence.

References

1. *Aleman-Meza, B., Halaschek, C., Arpinar, I. and Sheth, A.* "A Context-Aware Semantic Association Ranking," Proc. First Int'l Workshop Semantic Web and Databases (SWDB '03), pp. 33–50, 2003.
2. *Bar-Haim, R., Dagan, I., Greental, I. Shnarch, E.* Semantic Inference at the Lexical-Syntactic Level AAAI-05.
3. *Bhogal, J., Macfarlane, A., Smith, P.* A review of ontology based query expansion. Information Processing & Management. Volume 43, Issue 4, July 2007, Pages 866–886.
4. *Sadid, C. Y., Hasan, A., Joty, S. R.:* Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels. Inf. Process. Manage. 47(6): 843–855 (2011).
5. *Ercan, G., Cicekli, I.* Using lexical chains for keyword extraction. Information Processing & Management, Volume 43, Issue 6, November 2007, Pages 1705–1714.
6. *Galitsky, B.* Natural Language Question Answering System. Technique of Semantic Headers. Advanced Knowledge International, Australia (2003).
7. *Galitsky, B., González, M. P., Chesñevar, C. I.* A novel approach for classifying customer complaints through graphs similarities in argumentative dialogue. Decision Support Systems, 46–3, 717–729 (2009).
8. *Galitsky, B., de la Rosa, J. L., Dobrocsi, G.* Inferring the semantic properties of sentences by mining syntactic parse trees. Data & Knowledge Engineering. Volume 81–82, November (2012a) 21–45.
9. *Galitsky, B.* Machine Learning of Syntactic Parse Trees for Search and Classification of Text. Engineering Application of AI, http://dx.doi.org/10.1016/j.engappai.2012.09.017, (2012b).
10. *Kapoor, S., Ramesh, H.* Algorithms for Enumerating All Spanning Trees of Undirected and Weighted Graphs, SIAM J. Computing, vol. 24, pp. 247–265, 1995.
11. *Jung-Jae, K., Pezik, P.* and Rebholz-Schuhmann, D. MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline. Bioinformatics. Volume 24, Issue 11 pp. 1410–1412 (2008).
12. *Mann, W. C., Matthiessen C. and Thompson, S.* (1992). Rhetorical Structure Theory and Text Analysis. Ed. by W. C. Mann and S. A. Thompson. Amsterdam, John Benjamins: 39–78.
13. *Moschitti, A.* Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany, 2006.
14. *Plotkin, G. D.* A note on inductive generalization. In B. Meltzer and D. Michie, editors, Machine Intelligence, volume 5, pages 153–163. Elsevier North-Holland, New York, 1970.
15. *Punyakanok, V., Roth, D., & Yih, W.* (2004). Mapping dependencies trees: an application to question answering. In: Proceedings of AI & Math, Florida, USA.
16. *OpenNLP* 2013. http://incubator.apache.org/opennlp/documentation/manual/opennlp.htm

17. *Marcu, D.* From Discourse Structures to Text Summaries. In I. Mani and M. Maybury (eds) Proceedings of ACL Workshop on Intelligent Scalable Text Summarization, pp. 82–8, Madrid, Spain, 1997.
18. *Mihalcea and Tarau.* TextRank: Bringing Order into Texts. Empirial Methods in NLP 2004.
19. *Galitsky, B., Usikov, D., Kuznetsov, S. O.:* Parse Thicket Representations for Answering Multi-sentence questions. 20th International Conference on Conceptual Structures, ICCS 2013 (2013).
20. *Apresian, J., Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V., Tsinman, L.* ETAP-3 linguistic processor: A full-fledged NLP implementation of the MTT //MTT 2003, First International Conference on Meaning — Text Theory. — 2003. — С. 279–288.
21. *Sun, J., Zhang, M., Lim Tan, C.* Tree Sequence Kernel for Natural Language. AAAI-25, 2011.
22. *Zhang, M., Che, W.; Zhou, G. Aw, A., Tan, C., Liu, T. and Li, S.* 2008. Semantic role labeling using a grammar-driven convolution tree kernel. IEEE transactions on audio, speech, and language processing 16(7):1315–1329.
23. *Searle, J.* Speech acts: An essay in the philosophy of language. Cambridge, England: Cambridge University. 1969.
24. *Galitsky, B., Kuznetsov S. O.* Learning communicative actions of conflicting human agents. J. Exp. Theor. Artif. Intell. 20(4): 277–317 (2008).
25. *Apresian, J., Boguslavsky, I., Iomdin, L., Tsinman, L.* Lexical Functions as a Tool of ETAP-3. First International Conference on Meaning-Text Theory (MTT'2003). June 16–18, 2003. Paris: École Normale Supérieure, 2003.
26. *Иомдин Л., Петроченков В., Сизов В., Цинман Л.* Синтаксический анализатор системы Этап и его современное состояние. Papers from the national annual conference "Dialogue", 2012.
27. *Stanford NLP.* http://nlp.stanford.edu/
28. *Kuznetsov, S. O. and Samokhin, M. V.* Learning Closed Sets of Labeled Graphs for Chemical Applications. In: Proc. 15th Conference on Inductive Logic Programming (ILP 2005), Lecture Notes in Artificial Intelligence (Springer), Vol. 3625, pp. 190–208., 2005.

СТАТЬЯ ТАКАЯ СТАТЬЯ. ОБ ОДНОМ ТИПЕ РЕДУПЛИКАЦИИ В СОВРЕМЕННОМ РУССКОМ ЯЗЫКЕ

Гилярова К. А. (hilaris@gmail.com)

Институт лингвистики РГГУ, Москва, Россия

Ключевые слова: редупликация, коннотация, прототип, предикат, тавтология, семантика, дискурс, язык Интернета, блоги, социальные сети

PAPER SUCH A PAPER. ON A REDUPLICATION PATTERN IN MODERN RUSSIAN

Gilyarova K. A. (hilaris@gmail.com)

Russian State University for the Humanities, Moscow, Russia

The paper presents a semantic and pragmatic analysis of noun reduplication in colloquial Russian and the Internet language. We consider the repetition of a noun within the same prosodic unit separated by a particle “takoj” (‘such’) as in “statja takaja statja” (‘paper such a paper’). Drawing on a corpus of examples gathered from Internet texts we categorize the semantics of this reduplication pattern into six types: (1) prototype and connotation, (2) non-fitting a stereotype, (3) condescension and irony, (4) expression of emotions, (5) discourse topic and scene-setting topic (6) object nomination and ellipsis. Compared to the model “such X-X”, the model “X such X” more often points to the negative attitude. We also consider the syntactic structure of the given reduplication pattern.

Keywords: reduplication, connotation, prototype, predicate, tautologies, semantics, discourse, Internet language, blogs, social networking services

К сожалению, мне на кафедре сообщили,
что статья такая статья не прокатит.
(С аспирантского форума)

Введение

Настоящее исследование посвящено конструкции «Х такой Х», быстро завоевывающей позиции в языке Интернета и русской разговорной речи.

(1) *Чуть позже напьются и глаза друг другу повыцарапают… **Леди такие Леди**. **А свадьба такая свадьба**!*[1]

(2) *а я не **девочка такая девочка**, а вовсе даже женщина-вывеска-не-влезай-убьет, которая в обычной жизни ходит в джинсах, дурацких майках да кедах, а с маникюром-педикюром справляется за просмотром мультиков с подрощенной деткой или бизнес-новостями по работе.*

Сфера применения подобных конструкций с удвоением и актуализирующей коммуникативной частицей *такой* — блоги, социальные сети, форумы, чаты, в меньшей степени — устная речь. Они или употребляются непосредственно в текстах, или используются как заглавие поста, как подпись или комментарий к фотографии, видеоролику, комиксу, как метка (тэг), маркирующая серию постов или фотографий на одну тему.

Впервые внимание на описываемую конструкцию обратила Е. Л. Вилинбахова [Вилинбахова 2011]. На первый взгляд, конструкция «Х такой Х» напоминает синтаксическую редупликацию «такой Х-Х» [Гилярова 2010], [Ghomeshi 2004], [Horn 1993], но при ближайшем рассмотрении оказывается, что в синтаксисе, семантике и прагматике указанных конструкций немало отличий.

Прежде чем перейти к непосредственному анализу материала, отметим, какого рода примеры оказались вне поля нашего исследования:

(3) ***Больной*** (сущ.) ***такой больной*** (прилагат.)

(4) *Ах **лето, такое лето!***

(5) ***Дети — такие дети***.

В (4) мы имеем дело с повтором со значением усиления, запятая указывает на наличие паузы. Пример (5) — это биноминативное псевдотавтологическое высказывание, в котором дети в первой позиции — субъект, а во второй — предикат, ср. *Дети есть дети* [Апресян 1995]. В речи первая и вторая части

[1] Все примеры взяты из Интернета, орфография и пунктуация сохранена.

высказывания разделяются паузой, а на письме — тире или запятой. Иногда вставляется местоимение 3-го лица: ***Бабушка,*** *она* ***такая бабушка***. Между тем, в примерах (1) и (2) конструкция «X такой X» неделима и представляет собой целостную просодическую единицу, поэтому мы определяем её как синтаксическую редупликацию.

1. Синтаксические особенности

1.1. Исследуемая конструкция «X такой X»

Исследуемая конструкция встречается в предикативной и непредикативной роли. Предикативное употребление первично и более частотно. Выражение «X такой X» или целиком является односоставным предложением (простым или в составе сложного, ср. (6), (7)), или исполняет роль предиката в двусоставном предложении: (8), (9).

(6) ***Лингвисты такие лингвисты!***

(7) *Каким бы пропащим не был наш Ричард, мы все равно его любим и ждем, ведь* ***семья такая семья****.*

(8) *Личная жизнь* ***мужик такой мужик*** *весь из себя, никогда бы не подумала* (девушка узнала, что ее парень зарегистрирован на гей-форумах)

(9) *Мы* ***москвичи такие москвичи****, а вы там лапти все.*

1.2. Примеры использования конструкции «X такой X»

Е. Л. Вилинбахова рассматривает и такие примеры [Вилинбахова 2011]:

(10) *Моя* ***собака такая собака****, что любит все фрукты овощи-ягоды красного цвета…*

(11) *<…> Здесь можно сделать отступление и поделиться тем, какую «свинью» подложили ей журналисты <…> Хорошо по этому поводу выразилась Аня «****Журналисты такие… журналисты****».*

На наш взгляд, в (10) и (11) речь идёт о других явлениях: в (10) представлена конструкция «такой X, что», в (11) вторую позицию занимает, как верно замечает автор, эвфемизм, то есть подразумевается «X такой Y». И в обоих случаях предложение членится на субъект и предикат. А объектом нашего исследования является неделимая конструкция, в которой выражение «X такой X» как целое выступает в роли предиката или, реже, в непредикативной позиции.

Безусловно, в разграничении членимых и нечленимых псевдотавтологических конструкций есть некоторая сложность. Они отличаются интонационно: в первых присутствует пауза. На письме пауза должна бы отражаться постановкой тире, запятой или многоточия, однако пользователи Интернета, как известно, зачастую пренебрегают знаками препинания. Так что про большинство примеров вида «X такой X» мы не можем сказать с полной уверенностью, что автор высказывания имел в виду именно редупликацию «X такой X», а не конструкцию «X — такой X», аналогичную конструкциям «X есть X», «X он и есть X», описанным в [Апресян 1995], [Gibbs 1990].

Тем не менее, в пользу нечленимости исследуемой нами конструкции есть немало аргументов. Прежде всего, это её способность употребляться непредикативно, склоняться, присоединять к себе определения (см. (12), (13), (14)), а также выступать в виде целостного предиката в двусоставных предложениях (см. (2), (8), (9)). Во-вторых, по количеству употреблений в Интернете вариант с тире заметно уступает варианту без тире. Более того, самые устойчивые выражения вида «X такой X» пишутся через дефис: *девочка-такая-девочка, мужик-такой-мужик*.

Сомнения вызывают случаи, когда конструкция «X такой X» представляет собой отдельную клаузу и перед ней стоит определение: *Современные учителя такие учителя. Наша бабушка такая бабушка*. Иркутское (балтийское, питерское) лето такое лето. Остается открытым вопрос, относится ли определение только к первой позиции или ко всей конструкции. Неискушенному в сетевом жаргоне носителю языка и исследователю более вероятным кажется первый вариант, однако в пользу второго говорят примеры, в которых конструкция «X такой X», предваряемая определением, выступает в роли именной группы:

(12) *Серое дождливое **утро такое утро** надо мной всегда, и всё тихо, спокойно и идёт своим чередом.*

Таким образом, само по себе наличие определения при конструкции «X такой X» не может служить тестом для выявления её синтаксической структуры: и членимая, и нечленимая конструкции могут содержать в себе определения или сочетаться с ними.

1.3. Непредикативное употребление конструкции «X такой X»

Непредикативное употребление конструкции «X такой X» вторично по отношению к предикативному [Вилинбахова 2011] и встречается реже. Редуплицированное существительное может выступать в роли подлежащего или дополнения, склоняться и присоединять к себе именную группу.

(13) *Вокалист — пример умного, талантливого, красивого, харизматичного **мужика такого мужика**.*

(14) (Комментарий к статье про мужские обязанности) *Часть обязанностей во мне зародила мысль о том, что мужчина слепоглухонемощный инвалид, за которым нужен уход, еще часть заронила вопрос — а чем тогда* ***семья такая семья*** *отличается от свободных встреч с любимым время от времени?*

2. Семантика

2.1. Редупликация существительных «Х-Х»

В работе [Гилярова 2010] показано, что редупликация существительных «Х-Х» в современном русском языке Интернета может выражать 1) прототип и коннотацию Х-а, 2) высокую степень Х-а, 3) положительную оценку Х-а, 4) точное, буквальное значение Х-а и уточнение значения в случае его многозначности, 5) определенный референциальный статус Х-а. У конструкций «такой Х-Х» и «Х такой Х» схожие сферы употребления; на первый взгляд, они почти идентичны, так что естественно ожидать от них и семантического и прагматического сходства. Посмотрим, подтвердится ли эта гипотеза при более подробном анализе материала.

2.2. Значения конструкции «Х такой Х»

У конструкции «Х такой Х» выделяются следующие значения (2.2.1–2.2.6).

2.2.1. Прототип и коннотация

Х такой Х = ‘прототипический, обладающий характерными признаками или коннотациями Х-а Х’.

(15) *А по поводу вождения — вчера видела как из двора в довольно-таки сложной ситуации выежает* ***бабушка такая бабушка*** *по виду — я хорошо разглядела, т. к. она долго не могла выехать…*

Значение прототипа — основное и для редупликации «Х-Х». Однако при ближайшем рассмотрении оказывается, что конструкция «Х такой Х» чаще употребляется, чтобы подчеркнуть негативные прототипические черты, в то время как простое удвоение «Х-Х» более нейтрально. Так, в (16) имеется в виду свадебное торжество со всеми традиционными атрибутами, а (1) и (17) отсылают к таким неотъемлемым, с точки зрения русской языковой картины мира, характеристикам свадьбы, как пьянство и драка:

(16) *У вас прямо* ***свадьба-свадьба*** *была, или вы ограничились регистрацией и узким семейным кругом?*

(17) *А у нас как-то из-за свадьбы соседей спиртным из смежной с их квартирой розетки пахло.* ***Свадьба такая свадьба****.*

Как и биноминативные псевдотавтологические конструкции типа «X есть X» (*Война есть война*, *Boys will be boys*), конструкция «X такой X» может выявлять коннотации лексемы, то есть «несущественные, но устойчивые признаки выражаемого лексемой понятия, которые воплощают принятую в данном языковом коллективе оценку соответствующего предмета или факта действительности» [Апресян 1995, с.159]. Как и в случае с прототипом, описываемая нами конструкция вскрывает почти исключительно отрицательные коннотации:

(18) ***понедельник такой понедельник***. *прорвало кран, залило всю квартиру и соседей. в этом году новоселья не будет* (понедельник = 'тяжёлый день')

(19) ***Москва такая Москва***. *<…>Вместо 15-ти минут, запланированных на дорогу, у нас ушло около полутора часов.* (в Москве всегда затруднено уличное движение)

Если у существительного несколько коннотативных значений, актуализовано может быть любое из них, поэтому выражение «X такой X» может быть многозначно при одном и том же X-е. Так, Санкт-Петербург ассоциируется, прежде всего, с холодной дождливой погодой, но также и со своим городским диалектом, отличным от московского:

(20) *По поводу погоды можно сказать одно:* ***Питер такой Питер***.

(21) *Самый прикол был когда в Сапсане предлагали на выбор рыбу или куру :-)* ***Питер такой Питер*** *:-)))*

Также и выражение *Москва такая Москва* показывает столицу многогранно: как город перенаселенный и многонациональный, опасный и бандитский, как город больших денег и завышенных цен, как город транспортных пробок (ср. (19), (24)). А пятница в следующих примерах предстаёт перед нами как день, когда можно пораньше уйти с работы (22), когда можно вечером выпить (23) и как день, когда в городе затруднено уличное движение (24). Хотя все эти отличительные черты пятницы суть следствия одного и того же: пятница — последний день рабочей недели.

(22) *Все-таки* ***пятница такая пятница****! В отделе* (номинально) *— десять человек, из них <…> на рабочем месте присутствуют двое: я и мой начальник.*

(23) ***пятница такая пятница*** *и машина при себе))) В связи с чем вопрос, если кто пользовался подобной услугой, киньте телефончик эвакуаторов, которые за разумные деньги доставят меня вместе с машинкой домой!*

(24) ***Пятница такая пятница***. *И* ***Москва такая Москва***.
(заглавие поста с видеороликом про пробки)

В [Апресян 1995] справедливо отмечается, что понятие коннотации примиряет два принципиально различных подхода к интерпретации биноминативных конструкций: «радикально прагматический» подход Грайса и «радикально семантический» подход Вежбицкой [Grice 1975], [Wierzbicka 1987]. По Грайсу, столкнувшись с такой конструкцией, адресат «пытается реконструировать то субъективное содержание, которое вложил в свои слова его конкретный собеседник в конкретной ситуации общения» [Апресян 1995, с. 166]. Вежбицка же считает, что каждая псевдотавтологическая конструкция имеет устойчивую семантическую интерпретацию для данного языка. Для конструкции «X такой X» подход Вежбицкой работает лишь частично. Правильнее было бы говорить об устойчивой коннотации для данной группы носителей, порой весьма ограниченной. В этом смысле показательны примеры, где в роли X-а выступает имя или фамилия.

(25) ***Маша такая Маша!*** (снова опоздала, забыла, потеряла)

(26) *да-да,* ***Шиханович такой Шиханович****. причем я как раз получила у него пятерку, чем до сих пор страшно горда :)*

2.2.2. Несоответствие стереотипу

X такой X = ‘противоположный прототипическому, не обладающий характерными признаками X-а X’. Любовь такая любовь (название поста с видеороликом про измену). Бабушка такая бабушка (про молодую бабушку в джинсах на мотоцикле).

(27) *Теплая одежда. Обязательно.* ***Лето такое лето.*** *Ветровка и свитер.*
(Статья о сборах ребенка в лагерь)

«Неправильная» любовь, «неправильная» бабушка, «неправильное» лето. У конструкции «X-X» это значение полностью отсутствует.

2.2.3. Снисходительность

X такой X = ‘X проявил свои характерные отрицательные признаки, но ничего другого мы от него и не ждали и не сердимся’.

Иногда конструкция «X такой X» не просто показывает коннотации X-а, но и добавляет элемент иронии, чаще незлобной: ‘ну что с него взять, ничего другого я и не ждал’. Самым показательным примером может служить устойчивое выражение *мама такая мама*, которое практически стало интернет-мемом. Под заголовком или меткой *мама такая мама* собраны анекдоты, истории из жизни, серии картинок, комиксы, в которых мама предстаёт уже немолодой женщиной, отставшей от жизни, дающей глупые советы, нелогичной, чрезмерно опекающей своего ребёнка и вмешивающейся в его жизнь. Несмотря на то, что образ по такому описанию складывается отрицательный, в большинстве примеров прослеживается хорошее отношение говорящего к «маме такой маме».

(28) ***мама такая мама***
смирившись, видимо, что дочь увлеклась серфом, она, как настоящая мама, решила в этом поучаствовать. посмотрела ТВ передачу, где девачку учили вставать на серф. затем просмотрела мои португальские фотки. после чего авторитетно заявила мне по телефону: у тебя, говорит, пятка передняя неправильно на доске стоит! (на минуточку, она ваще боится заходить в воду глубже, чем по грудь)

(29) *– на спутнике собираются сделать разворотное кольцо для троллейбусов. – это надо столбы опять ставить.провода вести.может они уже какие нибудь бесконтактные троллейбусы придумали? – да мам.и они называются автобусы.* (с) ***мама такая мама***

(30) *шла домой, зашла в магаз, просто так купила маме коробку конфет, подарила. мама 100% подумала, что я где-то накосячила.* ***мама такая мама***

Интересно, что выражение *такая мама-мама* актуализует в сознании адресата совсем другие представления о прототипической маме: молодая, с маленькими пока ещё детьми, заботливая, уделяющая детям много внимания и времени [Гилярова 2010].Чаще всего это выражение встречается как комментарий к фотографии цветущей матери с маленьким ребенком на руках. И употребляется выражение *такая мама-мама* по отношению к чужой маме, а *мама такая мама* — к своей. Таким образом, на данном примере видна разница в семантике конструкций «Х-Х» и «Х такой Х». Даже там, где области их значений пересекаются, значения не полностью идентичны: «Х такой Х» апеллирует больше к отрицательным прототипическим свойствам и коннотациям, а «Х-Х» — к положительным или нейтральным.

2.2.4. Выражение оценки и эмоций

Х такой Х=‘хороший/плохой Х’, ‘Х меня радует/раздражает’.

У редуплицирующей конструкции «Х-Х» есть значение ‘хороший, красивый, стоящий Х’. Это же значение наблюдаем и у «Х такой Х»:

(31) ***Дом такой дом****! Поздравляю! Мне безумно-безумно нравится ванная комната.*

Положительные эмоции говорящего видны и в (32):

(32) ***Москва такая Москва*** *<…>. Сколько бы проблем не было с регистрациями, проездными и прочей гадостью, всегда есть настолько приятные впечатления, что о проблемах даже и не вспоминаешь))*

Однако такие случаи выражения положительных эмоций единичны. Как правило, конструкция «Х такой Х» выражает отрицательную оценку и негативные эмоции:

(33) *Всё было, как обычно: скучно и безрадостно. Последний урок,* ***учительница такая учительница****, и ещё кто-то голодный отгрыз кусочек моей парты. Сами понимаете, не фонтан.*

(34) ***Утро такое утро****. Сначала дождь, потом это долбанное метро, теперь еще и работать надо.*

Из примера (33) видно, что понятие ‘плохой X’ бывает почти синонимично понятию ‘непрототипический, неправильный X’. Однако это верно отнюдь не для всех существительных, ср. (34): никакого соответствия или несоответствия прототипу в дождливом рабочем утре нет.

2.2.5. Дискурсивный топик

Исследуемая конструкция часто выступает в роли названия поста или вступительного предложения, зачина. Любовь такая любовь — название текста про разные виды любви у древних греков. Счастье такое счастье — заглавие научно-популярной статьи про то, почему человек чувствует себя счастливым. Среда такая среда — заметка о том, как рассказчик провёл среду. Учителя такие учителя — воспоминания о своих школьных учителях. Работа такая работа — рассказ о некотором случае на работе (и случай никак работу не характеризует). В этих примерах конструкция «X такой X» не несёт никакого дополнительного значения: ни иронии, ни отсылки к стереотипам, ни оценки или эмоции. Она используется просто как дискурсивный маркер — вводит тему.

Интересны и другие примеры, когда «X такой X» фиксирует время или место происходящего:

(35) ***вторник такой вторник****. Только что минут 15 общалась по телефону с полнейшим неадекватом. А что, даже весело оказалось!*

Пример (35) имел бы коннотативное значение, если бы на месте вторника был понедельник или пятница, но вторник никаких коннотаций в русской языковой картине мира не имеет, а значит, выступает здесь в роли дискурсивного маркера. В (36) похожая ситуация: это не рассказ о лете, лето служит лишь фоном:

(36) ***Лето такое лето***
Ни разу в жизни больше не возьмусь рисовать брусчатку вручную. Лень, лень не даёт мне нормально работать с фонами — вот почему я не профессионал. Но квота есть — уже хорошо, да. А ещё я скоро уеду — не смогу нормально рисовать. И только сегодня прибыли в нашу школу результаты экзамена по русскому. Не люблю эту систему.

В аналогичной функции могут выступать и пространственные объекты: Москва такая Москва, дача такая дача. Таким образом, помимо обычного

дискурсивного топика, конструкция «X такой X» может служить топиком, задающим место и время (scene-setting topic) [Jacobs 2001]. В этом случае набор используемых существительных ограничивается семантическим полем пространства и времени.

Редупликация «X-X» в качестве дискурсивного маркера не используется.

2.2.6. Номинация объекта

X такой X='X'.

В следующих примерах значение описываемой конструкции близко к предыдущему, хотя и не идентично ему. «X такой X» выступает в назывной функции и может быть заменено на X без потери какого-либо дополнительного значения. Изменится только стиль текста — на сетевой жаргон.

(37) *Хочется макдачки.идти лень.**дождь такой дождь**. С Лизой проснулись час назад..поспали всего лишь часа 2..и спать не хочется..*

(38) *Помнится, **фильм такой фильм** «Рысь» с Макаровым. Там, у депутатши детей похитили.*

Конструкция может заменять не только отдельное существительное, но и всё содержащее его высказывание:

(39) *Я к вас к сожалению не могу. **дача такая дача**.*
(='я уезжаю на дачу')

(40) *в целом таж фигня: «тапо» — светофор, «ко» — кошка и пр. Есть и целые слова, конечно, их даже большинство, сейчас повторяет все почти, но **каша такая каша**, ага* (= 'у ребёнка в речи каша')

(41) *о, по-моему, ты единственный здесь рисуешь историю, а не отдельные иллюстрации. Стиль? Что-то есть. Но **карандаш такой карандаш**! Тушь, маркеры? Все сливается, контраста не хватает. Тебя убьет твоя консервативность.* (='но ты рисуешь только карандашом')

Возможно, такие употребления связаны с темпом общения в интернете. У пишущего не хватает времени на печатание полных фраз, и замена на «X такой X» — своеобразный эллипсис. И снова конструкция «X-X» в подобной функции выступать не может.

2.3. Набор используемых существительных

Теперь, когда мы выделили основные значения конструкции «X такой X», обратим внимание на то, какие существительные могут быть использованы в этой конструкции. В [Вилинбахова 2001] справедливо отмечается, что

«наиболее частотны термины родства, этнонимы, наименования лиц по возрасту, полу, далее возможны имена собственные, наименования животных, обозначения ситуаций, как-то: свадьба, бизнес, работа и пр. Реже всего встречаются конкретные неодушевленные существительные. Тем не менее, по нашим данным, употребление конкретных неодушевленных существительных всё же возможно, ср. (31), (41), а также:

(42) ***Банан такой банан***. *Он вначале растет вверх и верх, а когда становится пальмой начинает бананить.*

(43) ***Шоколад такой шоколад*** **облизывается** (подпись под изображением шоколада)

(44) *эта.. **стол такой стол**.. все делают, во. типа — опенсорс проект, собирай кто хочет ;)*

Как показывает проведенное исследование, круг существительных, используемых в конструкции «Х такой Х», шире, чем в «такой Х-Х», и это связано с различием в семантике двух конструкций. Самое распространенное значение конструкции «такой Х-Х» — обращение к прототипу или коннотации, а коннотативное значение, как правило, бывает как раз у терминов родства, этнонимов, наименований животных, в то время как у конкретных неодушевленных существительных оно встречается реже. А конструкция «Х такой Х», как мы показали, может выполнять и другие функции: назывную, дискурсивную, оценочную, эмотивную — и тогда выбор используемых существительных фактически неограничен.

2.4. Устойчивые выражения

Вернёмся к вопросу о целостности конструкции «Х такой Х». В 1.4. её членимость и нечленимость обсуждалась с точки зрения синтаксиса, теперь же мы можем рассмотреть проблему и с точки зрения семантики. Нам представляется, что у выражений «Х такой Х» есть три степени устойчивости.

1. Почти идиома — устойчивое сочетание, за которым закреплено отдельное значение, не выводимое из значения Х-а и известное всем носителям, употребляющим данное выражение. Например, *девочка-такая-девочка*. Устойчивость таких сочетаний закреплена и в графике — они часто пишутся через дефис, и в синтаксических свойствах — они могут употребляться в непредикативной функции, склоняться, присоединять к себе именную группу — исключительно как целое. Набор существительных, выступающих в таких устойчивых сочетаниях, ограничен одушевленными существительными с развитой коннотативной зоной.

2. Устойчивое сочетание, допускающее несколько семантических интерпретаций в зависимости от контекста. Значение сочетания, как правило, также не выводится из значений компонентов и понятно либо всем носителям, либо отдельным группам (социальным, профессиональным, региональным). Например, *пятница такая пятница, Питер такой Питер*. Такие сочетания могут употребляться как предикативно (как правило), так и непредикативно (существенно реже). Круг существительных, выступающих в таких сочетаниях, шире, чем в первом случае (например, включает в себя имена собственные), но по-прежнему ограничен существительными, способными порождать коннотации.

3. Неустойчивые, можно даже сказать регулярные образования, не имеющие никакого отдельного значения, которое было бы понятно всем носителям или даже группе лиц. Например, *вторник такой вторник, карандаш такой карандаш*. Используются как дискурсивный маркер, в назывной функции, для выражения эмоции или оценки. Употребляются только предикативно и только в виде отдельного предложения. Набор существительных, задействованных в конструкции, неограничен, в том числе это могут быть неодушевленные конкретные существительные.

Заключение

Проведённое исследование конструкции «X такой X» позволяет сделать следующие выводы.

- Конструкция «X такой X» может иметь следующие значения: 1) прототип и коннотация, 2) несоответствие прототипу, 3) «снисходительность», 4) оценка и выражение (негативных) эмоций, 5) дискурсивный топик и пространственно-временной маркер, 6) номинация объекта и дискурсивный эллипсис.
- В отличие от другой редуплицирующей модели, «X-X», конструкция «X такой X» актуализует прежде всего отрицательные коннотации и выражает отрицательные эмоции.
- Так как конструкция «X такой X» выполняет, среди прочего, назывную и дискурсивную функции, набор существительных, используемых в этой конструкции, не ограничен. Тем не менее, наиболее частотны существительные с развитой коннотативной зоной.
- В зависимости от значения и от выбранных существительных «X такой X» может быть как устойчивым сочетанием, так и продуктивной синтаксической моделью. Это отражается и в синтаксической структуре фразы, и просодически в устной речи, и графически на письме.

Проведению дальнейших исследований редупликации в современном русском языке очень способствовал бы корпус блогов или программа, позволяющая искать в поисковых системах Интернета любые повторы.

Литература

1. *Апресян Ю. Д.* Коннотации как часть прагматики слова (лексикографический аспект) // Интегральное описание языка и системная лексикография. М.: Языки русской культуры, 1995. Т. 2. С. 156–177.
2. *Вилинбахова Е. Л.* О конструкции вида «муж такой муж» в русском языке (на материале Интернет-источников) // Тезисы конференции «Русский язык: конструкционные и лексико-семантические подходы», Санкт-Петербург, 24–26 марта 2011 г.
3. *Гилярова К. А.* Такая девочка-девочка. Семантика редупликации существительных в русской разговорной речи и языке Интернета // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.) Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 90–96.
4. *Ghomeshi J., Jackendoff R., Rosen N., Russell K.* Contrastive focus reduplication in English (the Salad-Salad paper) // Natural Language & Linguistic Theory, 2004. V. 22. P. 307–357.
5. *Gibbs R. W., McCarrel N. S.* Why boys will be boys and girls will be girls: understanding colloquial tautologies // Journal of Psycholinguistic Research, 1990. Vol. 19. P. 125–145.
6. *Grice H. P.* Logic and Conversation // Syntax and Semantics. Vol.3, Speech Acts. N. Y., etc., 1975.
7. *Horn L.* Economy and Redundancy in a Dualistic Model of Natural Language // S. Shore and M. Vilkuna (eds.) Yearbook of the Linguistic Association of Finland. SKY1993. P. 31–72.
8. *Jacobs J.* The Dimensions of Topic-Comment // Linguistics, 2001. 39(4). P. 641–81.
9. *Wierzbicka A.* Boys will be boys // Language, 1987. Vol. 63. N. 1.

References

1. *Apresjan Ju. D.* (1995), Connotations as a part of the word's pragmatics (lexicographical aspect), [Konnotatsii kak chast' pragmatiki slova (leksikograficheskij aspekt)], Integral language description and systemic lexicography, [Integral'noe opisanie jazyka i sistemnaja leksikografija], V. 2, Jazyki russkoj kultury, Moscow, pp. 156–177.
2. *Ghomeshi J., Jackendoff R., Rosen N., Russell K.* (2004), Contrastive focus reduplication in English (the Salad-Salad paper), Natural Language & Linguistic Theory, Vol. 22, pp. 307–357.
3. *Gibbs R. W., McCarrel N. S.* (1990), Why boys will be boys and girls will be girls: understanding colloquial tautologies, Journal of Psycholinguistic Research, Vol. 19, pp. 125–145.
4. *Giljarova K. G.* (2010), Such a girl-girl. Semantics of noun reduplication in colloquial Russian and the Internet language [Takaja devochka-devochka. Semantika reduplikatsii sushchestvitel'nyh v russkoj razrovornoj rechi i jazyke Interneta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii. Po materialam Mezhdunarodnoj Konferentsii "Dialog 2010" (Bekasovo, 26–30 maja 2010)], Vol. 9(16), Izdatel'stvo RGGU, Moscow, pp. 90–96.
5. *Grice H. P.* (1975), Logic and Conversation, Syntax and Semantics, Vol. 3, Speech Acts, N. Y., etc.
6. *Horn L.* (1993), Economy and Redundancy in a Dualistic Model of Natural Language, S. Shore and M. Vilkuna (eds.), Yearbook of the Linguistic Association of Finland, SKY, pp. 31–72.
7. *Jacobs J.* (2001), The Dimensions of Topic-Comment, Linguistics, Vol. 39(4), pp. 641–81.
8. *Vilinbahova E. L.* (2011), On a costruction "husband such a husband" (lit.) in Russian (based on Internet sources) [O konstruktsii vida "muzh takoj muzh" v russkom jazyke (na material Internet-istochnikov)], Proceedings of the conference "Russian language: constructional and semantic approaches" [Tezisy konferentsii "Russkij jazyk: konstruktcionnyje I semanticheskie podhody"], St. Petersburg.
9. *Wierzbicka A.* (1987), Boys will be boys, Language, Vol. 63., N. 1.

LINGUISTIC ANALYSIS OF SOCIAL MEDIA

Grefenstette G. (Gregory.Grefenstette@3ds.com)

3DS Exalead, Paris, France

One can look upon the Web as a large corpus that can teach us about language use, and also about the real world. In order to determine what is new or interesting we need to know what the norm for language use is. This involves creating a language model that corresponds to what is found on the web. Since the web is so big, it is impossible to download it all and count appearances of words and phrases, so one must use the technique of probing: generating things to be tested and submitting them to a search engine to find their frequency of occurrence. It has been shown using Google to gather statistics is perilous since Google does not provide exact counts but rather estimates the number of pages containing an expression. These counts can be very far from the reality of what is really in Google's index. Using another search engine, such as Exalead, is one solution, but then the problem of index coverage comes into play. Google has declared having seen 1 trillion unique URLs (in 2008) but estimates of the size of Google's index are about 50 billion pages, so some hidden choice has been made of what is in the index and what is not. This means that frequency based language models derived from search engines are only approximate.

Nonetheless, it is possible to make rough, relative judgments of how often one linguistic phenomenon appears with respect to another, and using probing can provide some information of the relative frequency of these phenomena. Over a long period, it is possible to generate and test a great number of possibilities, some examples of the usefulness of this technique are finding what words commonly occur with other words, what colors are often associated with nouns, what are the most common translation of multiword expressions, what are the most likely transliteration of English terminology and names into Japanese, for example.

The Web is not a uniform corpus, far from it. There are many different language registers even within one language: there are professionally edited well written articles, there are more colloquial blog posts, there are hastily written error-filled comments, all which generate different language models. One recent exploitation of user-generated content on the web has been the mining of opinions concerning some subject, or company, or product. Affect analysis is now a thriving market and a true commercial success for natural language processing. Many other areas of text mining remain to be explored. For example, the particular language used to tag photos in social media sites (such as Panoramio or Flickr) and reveal many things about the user (especially in conjunction with GPS and time data). This language is different from that found in the general web, or on Wikipedia. We can use it to find out the interesting things to visit in a city, we can predict where a tourist can go, we can even guess whether a user is a woman or a man, from their tagging behavior. Mining this information can lead to additional applications that exploit this new knowledge.

ЖЕСТИКУЛЯЦИОННЫЕ ПРОФИЛИ РУССКИХ ПРИСТАВОК[1]

Гришина Е. А. (rudi2007@yandex.ru)

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

Ключевые слова: соотношение слова и жеста; языки сателлитного типа; приставки; русская жестикуляция; Мультимедийный русский корпус (МУРКО)

GESTURAL PROFILES OF RUSSIAN PREFIXES

Grishina E. A. (rudi2007@yandex.ru)

V. V. Vinogradov Russian Language Institute RAS, Moscow, Russia

The study analyzes the main types of gestures, which accompany the Russian verbs with/without prefixes. The gestures are described from the topological point of view: any hand/head movement is placed along the Cartesian coordinates and the statistical correspondence between prefixes and topological characteristics of gestures is detected. The paper presents the gestural profiles (the set of gestural attributes) of 16 Russian prefixes. The study makes use of the data of the Multimodal Russian Corpus (MURCO).

Key words: gesture-word coordination, satellite-framed languages, prefixes, Russian gesticulation; Multimodal Russian Corpus (MURCO)

1. Введение

В работах Л. Талми [Talmy 1985, 1991] было введено типологическое различение языков глагольного типа (verb-framed) и языков сателлитного типа (satellite-framed)[2]: в первых корневая часть глаголов движения преимущественно выражает **траекторию** перемещения, а **способ** перемещения выражается до-

[1] Исследование проведено при поддержке программы Президиума РАН «Корпусная лингвистика», а также гранта РФФИ № 10-06-00151.

[2] Перевод терминов заимствован из работы [Майсак 2005].

полнительными средствами, которые располагаются во фразе вне глагола; во вторых, напротив, корневая зона глагола выражает **способ** перемещения, а сателлиты (превербы, адвербы) выражают **траекторию**. К языкам глагольного типа относят романские, семитские, тюркские, японский, к языкам сателлитного типа — славянские, германские, финно-угорские.

Примеры (заимствованы из работы [Панов 2011: 3]). Русскому

(1) *Машина з а е х а л а в гараж,*

где корень глагола *ехать* передает способ перемещения (с помощью транспортного средства), а преверб-сателлит (приставка) *за-* передает направление, траекторию движения, можно противопоставить итальянское

(2) *La macchina è e n t r a t a nel garage,*

где глагол *entrare* никак не специфицирует способ перемещения (это может быть полет, перемещение с помощью транспорта, пешком и т. д.)

В работе [Slobin 2004] эта типологическая классификация была дополнена языками эквиполентного типа, в которых и способ перемещения, и его траектория выражены в рамках сериальной глагольной конструкции (это, в частности, китайский язык).

Типологическое противопоставление языков глагольного, эквиполентного и сателлитного типов оказалось весьма провоцирующим для **жестикуляционной лингвистики**. Исследователи заинтересовались вопросом, отражаются ли эти противопоставления в сопровождающей речь жестикуляции. В основном проблема изучалась на материале экспериментов (носители разных типов языков пересказывали один и тот же мультфильм, а затем сопоставлялась жестикуляция, которую носители типологически разных языков использовали в одной и той же зоне пересказа). Одной из первых работ такого плана, по-видимому, является статья [McNeill, Duncan 2000], где были сопоставлены данные (американского) английского и китайского языков. При том что впоследствии появились сомнения в обоснованности выводов этого исследования ([Chui 2009, 2012]), сама методология была воспринята как чрезвычайно продуктивная и воспроизводилась во многих последующих работах[3].

Русский язык относится к языкам сателлитного типа — направление движения выражается в нем прежде всего приставками (*восходить*), а также предлогами (*идти в школу*) и наречиями (*ходить вокруг да около*), в то время как корень передает способ перемещения (*вскарабкаться, вползти, прилететь*). Нам показалось интересным проследить на русском материале соотношение

[3] Целая серия работ была посвящена сопоставлению глагольной жестикуляции английского, турецкого и японского языков ([Kita, Özyrek 2003], [Özyrek et al., 2005], [Kita et al., 2007]. Аналогичные исследования были проведены на материале разноязычной детской речи ([Allen et al. 2007], [Özyrek et al., 2008], [Chen 2007], [Gullberg et al., 2008]), а также на материале изучения иностранных языков ([Brown, Gullberg 2008, 2010], [Choi, Lantolf, 2008]).

жестикуляции и глаголов (приставочных и бесприставочных) и попытаться определить, отражается ли семантика приставок в жестикуляции, и если отражается, то как именно. Именно этой проблеме посвящено настоящее исследование.

Здесь следует дать специальное пояснение. Наша работа **ни в коей мере** не является работой по семантике русских приставок. Именно поэтому мы позволили себе не обращаться к огромному пласту уже имеющихся разысканий в этой области. Исследование является попыткой выяснить, как именно русский приставочный фонд отражается в жестикуляционном сопровождении устной речи, — не более того. Но и не менее.

2. Материал

В отличие от упомянутых выше работ по жестикуляционному исследованию превербов, наша работа не имеет, во-первых, сопоставительного характера — мы планируем все время оставаться в рамках русского языка, не сравнивая его с другими языками (типологически близкими или далекими). Второе отличие — исследование проводилось на довольно обширном корпусном материале[4], а не на материале психолингвистических экспериментов. В-третьих, хотя большая часть материала представляет собой контексты с глаголами движения, но мы не ограничивали себя только ими: если жестикуляционный ряд включал в себя движения рук и головы, синхронизированные с глаголами других типов (например, ментальными глаголами, глаголами речи и под.), то этот материал также включался в рассмотрение. И наконец, в отличие от многих работ по теме, мы включали в рассмотрение не только жесты рук, но и жесты головы. В результате основу нашей работы составили клипы из Мультимедийного русского корпуса (МУРКО), в которых нами были зафиксированы 1225 головных и 3086 ручных жестов, сопровождающих глаголы[5].

Такое значительное количество контекстов позволило нам использовать в работе статистические методы. Однако даже база данных такого объема оказалась недостаточной для ряда относительно редких русских приставок — в дальнейшем из рассмотрения выводятся приставки *без-, из-, об-, пред-*, примеры на которые есть, но их количество не дает нам возможности получать

---

[4] Материал был собран из Мультимедийного русского корпуса (МУРКО), функционирующего в рамках Национального корпуса русского языка.

[5] Здесь нужно сделать два уточнения: 1) под жестом, сопровождающим глагол, имеется в виду, прежде всего, жестикуляционная ситуация, когда ударная часть того или иного жеста синхронизирована с глаголом; однако поскольку русский язык в значительной степени дублирует значение приставки в предлоге, то в рассмотрение включались также жесты, сопровождающие предложные актанты соответствующего глагола, если эти актанты непосредственно примыкали к глаголу (хотя таких примеров и существенно меньше, чем контекстов первого типа); 2) МУРКО на момент проведения исследования в основном включал в себя кинематографическую речь, так что бóльшая часть контекстов относится к русскому кинематографу 1930–2010-х гг., однако некинематографический материал, который входит в МУРКО, также анализировался.

достоверные статистические распределения. Кроме того, практически не включены в статистику сочетания приставок.

3. Методология и понятийный аппарат

При обследовании материала мы исходили из следующей посылки: **жестикуляционное сопровождение глагола не должно противоречить его типологическим характеристикам.** Это означает следующее: поскольку типологически русский глагол является сателлитным, т. е. **направление** движения передается в нем приставками, а **способ** передвижения — корнем, жестикуляционное сопровождение глагола либо никак не реагирует на это свойство русского глагола, либо также является сателлитным. Иными словами, глаголы с одной и той же приставкой статистически предпочитают одни и те же пространственные характеристики. Таким образом, мы заранее отказались от потенциального противоречия: глагол является сателлитным, а сопровождающая его жестикуляция — глагольной (т. е. мы считали невозможным, что в глаголах с корнем *ехать* жестикуляция будет регулярно отражать способ перемещения — *на транспорте* — и не будет отражать пространственных характеристик приставок *заехать, приехать, выехать, отъехать, проехать, поехать, уехать* и проч.). Сразу скажем, что эта посылка полностью подтвердилась.

Здесь следует подчеркнуть, что все описанные ниже закономерности имеют исключительно статистический характер — по одной важной причине: жестикуляционная топология в русском языке, как и в большинстве языков мира, имеет не только лексический, но и грамматический характер, а именно — направление жестов предназначено в значительной степени для выражения таких собственно грамматических и квазиграмматических противопоставлений, как значения времени (актуальное — неактуальное время), типа иллокуции (вопрос, императив) и эвиденциальности (противопоставление мнения и факта, см. об этом [Гришина 2013а]). Так, в примере (3)

(3)

Речевой ряд	*Вы завтра утром на рассвете*	*уезжаете*	*в Ленинград*
Жестовый ряд		движение ладонью сверху вниз	

Ю. Махульский, А. Бородянский. Дежавю, 1989

на глаголе *уезжаете*, включающем приставку *у-*, которая, как мы покажем ниже, характеризуется движением налево, рука говорящего движется сверху вниз. Движение руки вниз здесь вызвано двумя факторами: во-первых, тем, что в этой фразе говорящий использует уверенное будущее время (противопоставляемое неуверенному будущему и ирреалису), для которого как раз и характерно движение руки или головы сверху вниз; во-вторых, движение вниз связано с семантическим компонентом ‘уничтожение’, ‘исчезновение’, которое присутствует в глаголе *уезжать*. Таким образом, в данном случае параметр времени и семантика глагола оказываются для говорящего более существенными,

чем пространственные характеристики, выражаемые приставкой, что и отражается в жестикуляции.

Следовательно, каждый жест, сопровождающий глагол, оказывается **результатом выбора** грамматического или лексического параметра, существенного для говорящего в данный момент, из чего следует, что все закономерности, которые мы надеемся наблюсти, могут иметь только статистический характер.

Поскольку жестикуляция осуществляется в пространстве, то для ее характеристики разумно использовать декартовы координаты, адаптированные к человеческому телу (таким образом, например, описывается темпоральная жестикуляция, краткий обзор литературы по этой теме см. в [Гришина 2013а]): каждый жест осуществляется в трех измерениях, соответственно, может быть разложен на три вектора по трем осям (табл. 1, рис. 1):

Таблица 1

Название оси	**Направление**	**Значения**
сагиттальная (коммуникативная)	вперед-назад	вперед, на себя, за спину (AB)
поперечная (когнитивная)	право-лево	направо, налево (CD)
вертикальная	вверх-вниз	вверх, вниз (EF)

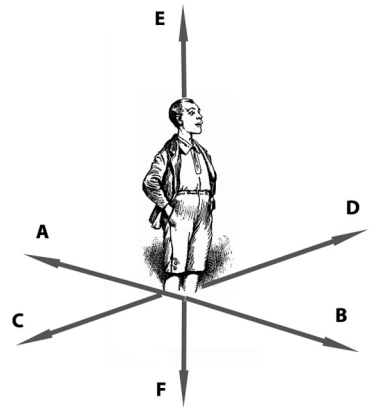


Рис. 1

Таким образом, каждый жест, который сопровождает глагол, мы раскладываем по этим трем осям и приписываем жесту соответствующее значение (так, в примере (3) движение открытой ладонью, сопровождающее глагол *уезжаете,* будет описано как имеющее характеристики *вперед* (по коммуникативной оси) и *сверху вниз* (по вертикальной оси)).

Следует, однако, отметить, что хотя значительная, если не бóльшая часть жестов является **векторной**, т. е. включает не более одного вектора вдоль одной оси, но это не единственная возможность. Для ряда движений характерно последовательное или одновременное (для двуручных жестов) использование

двух векторов. Такие жесты в дальнейшем мы будем назвать **маршрутными**. Маршрутные жесты могут быть осуществлены только руками, но не головой. Проиллюстрировать их удобнее всего рисунками (табл. 2, рис. 2–3).

Таблица 2

Траектория	**Направление движения**
выгнутая дуга	вверх-вниз (рис. 2)
вогнутая дуга	вниз-вверх (рис. 2)
изнутри наружу	направо-налево: обе руки одновременно в стороны от зоны коммуникации (рис. 3)
снаружи в центр	слева-справа: обе руки одновременно в зону коммуникации (рис. 3)

Рис. 2

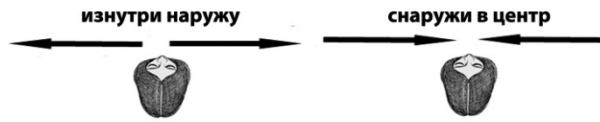


Рис. 3

Заметим, что вертикальные маршрутные жесты (выгнутая и вогнутая дуги) могут быть одновременно охарактеризованы и как векторные, потому что им ничто не мешает располагаться как вдоль коммуникативной оси (вперед), так и вдоль когнитивной оси (т. е. они могут быть направлены влево или вправо). В отличие от вертикальных, горизонтальные маршрутные жесты (*развести руками* и *сдвинуть руки*) — это жесты, симметричные относительно зоны коммуникации, и, как следствие, они могут осуществляться только вдоль поперечной оси, а единственное доступное им направление по сагиттальной оси — *вперед* — является вынужденным и, следовательно, семиотически непоказательным.

И наконец, в данной работе нам потребуется разделение на **объектно-связанные** и **объектно-свободные (когнитивные)** жесты, которое мы ввели

и активно использовали в работах [Гришина 2013а,б]. Объектно-связанными мы называем жесты, направленные либо на присутствующий в поле зрения говорящего объект, либо на отсутствующий объект, направление к которому говорящему хорошо известно, см. (4)–(5).

(4)

Речевой ряд	*Ну*	*вали*	*отсюда!*
Жестовый ряд		боковой кивок в сторону входной двери	

А. Серый, Г. Горин. Ты — мне, я — тебе, 1977

(5)

Речевой ряд	*Мэтр, я вам клянусь, кто-то меня*	*бросил*	*на пол…*
Жестовый ряд		движение двумя ладонями вниз, в сторону пола	

А. Сурикова, М. Агранович. Ищите женщину, 1982

Когнитивными (объектно-свободными) жестами мы называем жесты, направленные на отсутствующий объект, актуальное направление на который неизвестно или неважно, или на объект, не существующий (в настоящий момент) в реальности, см. (6)–(7).

(6)

Речевой ряд	*Он бросается на меня, я*	*бросаюсь*	*в седло, и вот я здесь.*
Жестовый ряд		движение двумя кулаками сверху вниз и изнутри наружу	

Л. Квинихидзе. Соломенная шляпка, 1974

(7)

Речевой объект	Право любой свободной индивидуальности перейти	на сторону сильного
Жестовый ряд		боковой кивок налево

В. Басов, В. Кожевников. Щит и меч, 1968

При указании на когнитивные объекты выбор направления указания определяется в значительной степени не реальным расположением вещей в пространстве, а представлениями говорящего о мире. Последнее влечет за собой усиление влияния лингвистических факторов на пространственную ориентацию жеста.

Вышесказанное предопределяет методику статистической обработки материала. Как мы уже писали выше, каждый жест характеризуется с точки зрения его векторного строения и с точки зрения траектории, если таковая наличествует. При этом для векторных жестов учитываются случаи только когнитивного (объектно-свободного) использования, чтобы исключить возможность влияния на ориентацию жеста в пространстве экстралингвистических факторов (реального расположения объектов в окружающем мире). Для маршрутных жестов такого ограничения нет, и мы работаем со всем материалом, т. е. и с объектно-связанными жестами тоже. Каждому жесту в соответствие

ставится некоторый набор векторов и траекторий, а потом для каждой приставки производятся соответствующие подсчеты. Полученные табличные данные проверяются на достоверность по критерию χ-квадрат. Если распределение признается достоверным, а параметры связанными, то мы получаем для каждой приставки наиболее характерный для нее набор векторов и траекторий. Кроме того, будут в общих чертах проанализированы наиболее характерные для приставок конфигурации ладони. Каждой приставке мы предполагаем поставить в соответствие некоторый набор характеристик. Именно этот набор мы называем **жестикуляционным профилем приставки**.

В заключение раздела опишем устройство таблиц. В каждой таблице перед слэшем стоит реальное число, зафиксированное на нашем материале для того или иного параметра, а после слэша — ожидаемое, среднее число, которое мы могли бы наблюдать, если бы данная приставка не была связана с данными параметром. Если мы работаем с достаточно объемным материалом, то значимым расхождением между ожидаемыми и реальными данными мы считаем случаи, когда χ-квадрат для данной ячейки равен или превосходит 2. Если же материала не слишком много, то мы оставляем за собой право ориентироваться и на менее мощные значения χ-квадрата (≥ 1). В случае, если реальные значения больше ожидаемых, ячейка выделена полужирным шрифтом, если меньше — то курсивом.

4. Данные

4.1. Приставки и основные жестикуляционные оси

В табл. 3 мы приводим данные по векторным характеристикам бесприставочных и приставочных глаголов.

Таблица 3. Векторные характеристики приставок

Ось / Приставка	Коммуникативная			Когнитивная		Вертикальная	
	вперед	**за спину**	**на себя**	**налево**	**направо**	**сверху вниз**	**снизу вверх**
бесприставочные глаголы	426/411	20/17	66/60	196/189	199/213	472/478	162/173
в-	40/40	2/2	8/6	*12/19*	17/21	**60/47**	12/17
воз-	34/38	1/2	3/6	12/18	17/20	*25/45*	**52/16**
вы-	67/68	2/3	9/10	34/32	30/35	76/80	30/29
до-	73/67	1/3	7/10	*17/31*	31/35	**102/78**	*19/28*
за-	48/55	**5/2**	**14/8**	22/25	25/29	68/65	26/23
на-	55/60	2/2	5/9	29/28	26/31	**86/70**	23/25
о-	64/66	2/3	*3/10*	*16/31*	30/34	**115/77**	*19/28*
от-	35/40	0/2	5/6	**25/18**	26/21	38/46	20/17
пере- (пре-)	58/57	2/2	*3/8*	**47/26**	**50/29**	*44/66*	*9/24*

Ось / Приставка	Коммуникативная			Когнитивная		Вертикальная	
	вперед	за спину	на себя	налево	направо	сверху вниз	снизу вверх
по-	98/87	3/4	13/13	34/40	44/45	94/101	40/37
под-	24/25	0/1	5/4	10/11	11/13	*20/29*	**23/10**
при-	72/68	5/3	**18/10**	25/31	35/35	77/79	24/29
про-	32/30	0/1	3/4	13/14	16/15	30/35	**18/13**
раз-	42/42	2/2	*2/6*	20/19	**29/22**	42/49	20/18
с-	*49/62*	2/3	**18/9**	**38/29**	31/32	68/72	27/26
у-	70/70	4/3	6/10	**42/32**	41/36	82/82	*19/30*
χ^2=305, p⩽1,28^{-23}; параметры связаны, распределения достоверны							

Табл. 3 демонстрирует, что бесприставочные глаголы не имеют каких-то предпочтений в векторных характеристиках, что выглядит вполне естественным, если мы полагаем, что ориентация жеста в пространстве определяется значением приставки. Однако и среди приставочных глаголов не все имеют векторные характеристики. Таковыми 0-векторными приставками являются приставки *вы-* и *по-*. Остальные приставки разбиваются на три группы: ориентированные по коммуникативной оси (рис. 4), по когнитивной оси (рис. 5) и по вертикальной оси (рис. 6)[6].

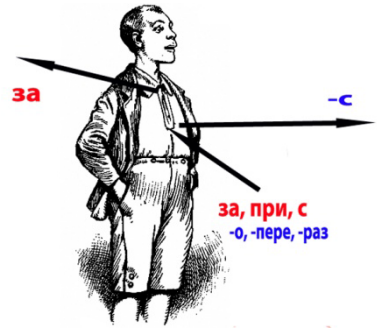


Рис. 4. Приставки вдоль коммуникативной оси

Обратим внимание, что приставка *за-* по сагиттальной оси имеет два способа реализации — *на себя*, т. е. на говорящего, и жест *за спину*. К этому вопросу мы вернемся ниже, в разделе 4.2, примечание 4.

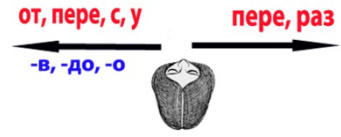


Рис. 5. Приставки вдоль когнитивной оси

[6] Минусом на рисунках снабжены приставки, которым данный вектор «противопоказан», т. е. отмеченные курсивом в табл. 3.

Примечание 1. Тяготение приставки *раз-* к правой зоне, как и симметрия приставки *пере-* (равное тяготение к правой и левой зоне) объясняются, вероятнее всего, маршрутностью этих приставок (см. 4.2). Для приставки *пере-* в высшей степени характерна траектория *выгнутая дуга* (см. рис. 2), что подразумевает, среди прочих направлений, симметричное движение справа налево или слева направо по поперечной оси. Маршрутная приставка *раз-*, в классическом случае сопровождаемая движением обеих рук *изнутри наружу* (см. рис. 3), в редуцированном варианте, по-видимому, сопровождается жестом одной правой (доминантной) руки направо, что и определяет тяготение этой приставки к правой зоне. Показательно, что на поперечной оси **ни одна** из чисто векторных приставок (*от-, с-, у-*) не обнаруживает тяготения вправо — все они ориентированы влево.

Здесь также нужно отметить, что, согласно когнитивным и семантическим исследованиям, левое и правое направления в русском языке аксиологически неравноценны, а именно, левое направление связано с отрицательными характеристиками и коннотациями (см. [Cienki 1999]).

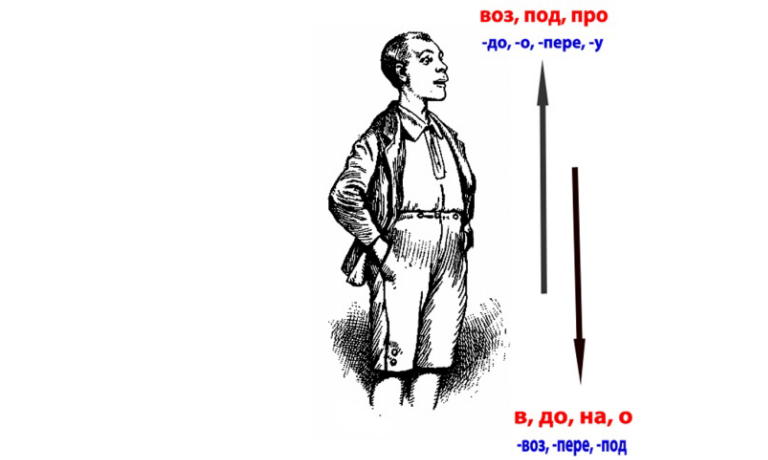


Рис. 6. Приставки вдоль вертикальной оси

Как видим, группировка приставок по осям довольно последовательна: только одна приставка, *с-*, тяготеет к двум осям одновременно — к коммуникативной оси (*на себя*) и к когнитивной оси (*налево*). И только одна приставка является симметричной, т. е. располагается на одной и той же оси и при этом имеет противоположные значения (*пере-* — одновременно лево- и правоориентированная приставка).

4.2. Приставки с точки зрения маршрута

Теперь проанализируем маршрутные характеристики этих же приставок (табл. 4).

Таблица 4. Маршрутные характеристики приставок

Траектория / Приставка	Когнитивная ось		Вертикальная ось	
	изнутри наружу	снаружи в центр	выгнутая дуга	вогнутая дуга
бесприставочные глаголы	51/49	17/18	97/103	17/13
в-	6/6	2/2	14/12	0/2
воз-	1/4	2/2	8/9	**5/1**
вы-	7/5	1/2	11/11	1/1
до-	*3/7*	3/2	15/14	4/2
за-	4/7	**6/3**	16/15	1/2
на-	*1/5*	1/2	**17/11**	1/1
о-	**15/9**	1/3	18/19	0/2
от-	5/4	1/1	8/8	1/1
пере- (пре-)	*8/18*	4/6	**53/38**	*2/5*
по-	11/8	2/3	15/18	3/2
под-	1/3	2/1	4/6	3/1
при-	7/9	2/3	26/20	0/3
про-	4/3	1/1	4/6	2/1
раз-	**19/9**	1/3	*12/19*	1/2
с-	11/10	**12/3**	*12/20*	1/3
у-	9/7	1/3	15/15	2/2
χ^2=124,7, p⩽9,75^{-9}; параметры связаны, распределения достоверны				

Как видим, маршрутными характеристиками обладает существенно меньшее количество приставок, по сравнению с векторными характеристиками. Снова, как и в случае векторов, бесприставочные глаголы и глаголы с приставками *вы-* и *по-* не имеют предпочтительных траекторий. Но 0-маршрутными являются также глаголы с приставками *в-, до-, от-, под-, при-, про-, у-*.

Зрительно данные таблицы переданы на рис. 7–9.

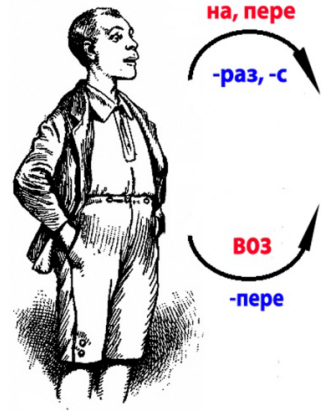


Рис. 7. Вертикальные траектории

Примечание 2. Значимые зоны выгнутой дуги существенно различаются в случае приставок *на-* и *пере-*. Для приставки *на-* выгнутая дуга — это вариант вектора *сверху вниз*, характерного для векторного варианта этой приставки (см. рис. 6). Точно так же для приставки *воз-* вогнутая дуга является вариантом вектора *снизу вверх*. То есть для приставок *на-* и *воз-* значимой частью дуги является ее конечная часть, направленная вниз в случае *на-* и вверх — в случае *воз-*. Для приставки *пере-* ситуация существенно иная: здесь значимым является именно переход от вектора, ведущего вверх, к вектору, ведущему вниз, т. е. центральная часть дуги, которая как раз символизирует собой траекторию движения объекта, преодолевающего препятствие **поверх** этого препятствия. Поэтому приставка *пере-* может считаться в первую очередь маршрутной приставкой. Расположение же ее на поперечной оси (рис. 5) является конвертацией кривой маршрута в прямую вектора.

Рис. 8. Траектория изнутри наружу

Примечание 3. Для приставки *о-* на поперечной оси мы, по-видимому, можем постулировать несколько иное значение, чем эта приставка имеет на вертикальной оси (вектор *сверху вниз*). На вертикальной оси *о-* означает движение к некоторой конечной точке (*окунуть, опустить*). На поперечной (когнитивной) оси движение обеих рук изнутри наружу в случае приставки *о-* означает формирование некоего объемного объекта между руками говорящего (*ожить, оформить*) или обозначение некоего кругового движения (*окружить, оглянуться*). Приставка *раз-* жестикуляционно обозначает растяжение объекта или линии в ширину (*растянуть*) или разведение в стороны двух или более объектов (*разбегаться*).

Рис. 9. Траектория снаружи в центр

Примечание 4. Приставка *за-* в маршрутном варианте на поперечной оси передает идею присвоения, приближения к себе, вовлечения в свою зону коммуникации (*захватить, зажать*). Это же значение трансформируется в движение руки *на себя*, к говорящему, по сагиттальной оси (см. рис. 4). Таким образом, мы можем постулировать у приставки *за-*, с точки зрения жестикуляции, два значения: 1) захват, приближение к себе и 2) помещение объекта за какое-л. препятствие (*заходить в хвост, зайти с тылу, зайти за что-л.*). См. также примечание 5 о приставе *с-*.

Маршрутными приставками являются приставки *воз-, за-, на-, о-, пере-* (*пре-*), *раз-, с-*. Интересно при этом, что в наибольшей степени траектория выражена у тех приставок, которые не имеют омонимичных предлогов: *воз-* (соответствующий χ^2=12,8), *пере-/пре-* (6,1), *раз-* (11,8).

Примечание 5. Единственное исключение — приставка *с-* (20,9), для которой в высшей степени характерна траектория *снаружи в центр*. Мы полагаем необходимым прокомментировать это исключение. Возвращаясь к табл. 3–4, мы можем констатировать, что приставка *с-*, по жестикуляционным данным, демонстрирует три разных значения.

1) При ориентации *на себя* (коммуникативная ось, векторное значение) приставка *с-*, с жестикуляционной точки зрения, имеет значение присоединения к некоторой конечной точке, притяжения к этой точке, и при этом точка присоединения или притяжения привязывается говорящим к его собственному месту расположения, с чем и связано появление автоуказания. Таким образом, мы здесь сталкиваемся с использованием тела говорящего как точки отсчета. В примере (8) предполагается, что собирать Х будет слушающий, но жестикуляция дает нам знать, что в этой фразе говорящий отождествляет себя со слушающим, соответствующим образом смещая точку зрения. В примере (9) говорящий отождествляет себя с персонажем своей речи (*демоны*).

(8)

Речевой ряд	*Собирайте*	*их капля по капле*
Жестовый ряд	движение кулаком на себя	

Ф. Эрмлер и др. Великий гражданин, 1937, 1939

(9)

Речевой ряд	*Демоны тебя*	*схватили…*
Жестовый ряд	движение кулаком на себя	

Л. Гайдай и др. Иван Васильевич меняет профессию, 1973

В этом значении приставка *с-* является векторной и имеет парный предлог *с* (*взять кого-л. с собой*). Аналогичным образом в значительном числе случаев объясняется автодейксис на приставке *при-*, см. (10).

(10)

Речевой ряд	*Кристаллы притягиваются к источнику сигнала*
Жестовый ряд	многократное указание на себя

Г. Шенгелия и др. Классик, 1998

2) При ориентации *налево* (когнитивная ось, векторное значение) приставка *с-* имеет значение отсоединения от некоторой начальной точки (11)–(12).

(11)

Речевой ряд	*и решили*	*спрятать*	*куда-нибудь драгоценности*
Жестовый ряд		движение ладонью налево и сверху вниз	

Е. Татарский, Э. Дубровский. Приключения принца Флоризеля, 1979

(12)

Речевой ряд	*Я наверно*	*сорвала*	*вас?*
Жестовый ряд		движение ладонью налево	

И. Масленников, В. Валуцкий. Зимняя вишня, 1985

В этом значении приставка *с-* также является векторной и имеет парный предлог *с* (*убрать со стола*).

3) При траектории *снаружи в центр* (когнитивная ось, маршрутное значение) приставка *с-* имеет значение собирания чего-л. из более чем одного источника, причем не в некоторой точке, а в некотором месте, имеющем объем или площадь (13).

(13)

Речевой ряд	*а внизу собирается*	*порядка трех-четырех*	*самцов*
Жестовый ряд		многократное движение обеими ладонями снаружи в центр	

Д. Дьяченко и др. День радио, 2008

В этом значении приставка *с-* является маршрутной и, по-видимому, не имеет парного предлога, потому что наиболее точным образом сходное значение передается не предлогом, а наречием *вместе*.

4.3. Приставки и движение рук на когнитивной оси

Движения руки, сопровождающие глаголы с той или иной приставкой, могут быть проанализированы с точки зрения еще одного, достаточно своеобразного критерия, который, насколько нам известно, пока не привлекался к изучению жестикуляционного сопровождения превербов. Этот критерий связан с выбором руки для осуществления жеста (правой или левой) и типом ее движения относительно зоны коммуникации по когнитивной оси[7].

Здесь мы выделяем 4 возможности.

1) **Комфортное движение** (правой или левой) руки в свой половине (правой или левой) по направлению ***от*** зоны коммуникации (т. е. правая направо и левая налево) — см. рис. 10.

---

[7] Заметим, что этот параметр играет существенную роль не только в характеристике приставок, но и во временных и эвиденциальных характеристиках русского глагола.

Рис. 10. Комфортное движение

2) **Притяжение** — движение (правой или левой) руки в своей половине (правой или левой) по направлению ***к*** зоне коммуникации (т. е. правая налево и левая направо) — см. рис. 11.

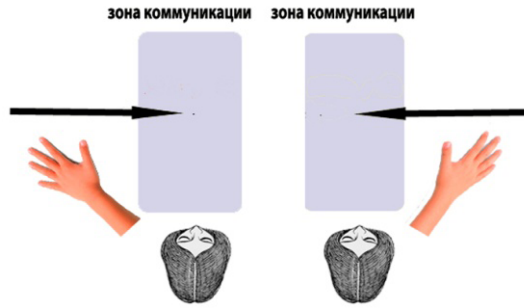


Рис. 11. Притяжение

3) **Пересечение** — движение (правой или левой) руки в том же направлении, что и при притяжении (т. е. правая налево, а левая направо), но при этом ***с пересечением*** зоны коммуникации и с попаданием руки в противоположную зону (правой — в левую, а левой — в правую), см. рис. 12.

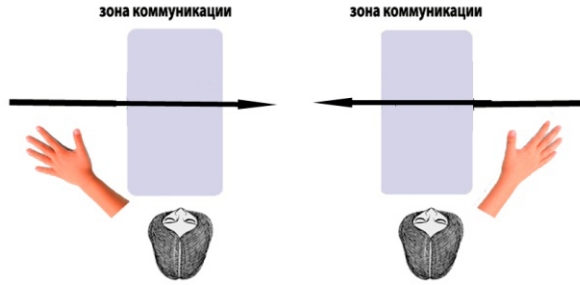


Рис. 12. Пересечение

4) **Радикальное пересечение** — перемещение (правой или левой) руки в противоположную зону (правой — в левую, левой — в правую) с последующим возвратом в исходную зону; т. е. при радикальном пересечении мы имеет дело с двойным пересечением зоны коммуникации — сначала рука занимает максимально некомфортную, далекую от исходной позицию, а затем возвращается в исходную позицию, см. рис. 13.

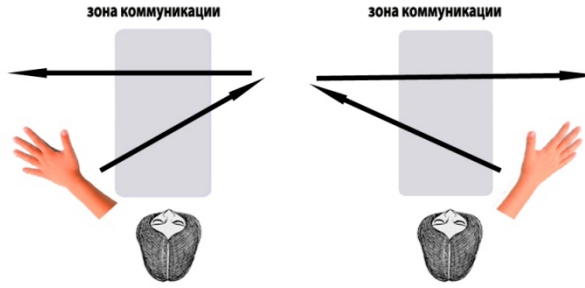


Рис. 13. Радикальное пересечение

Поскольку для каждой приставки отдельно соответствующих данных не очень иного –даже на собранном нами достаточно обширном корпусе, — мы провели некоторую группировку, пользуясь, в частности, данными предшествующего раздела. Были сформированы 4 группы приставок:

1) приставки, ориентированные вверх (*воз-, под-, про-*)
2) приставки, ориентированные вниз (*в-, до-, на-, о-*)
3) приставки ориентированные налево (*от-, у-*); сюда же вошла приставка *вы-*, которая, по предварительным прикидкам, вела себя со статистической точки зрения так же, как и левоориентированные приставки
4) приставки, ориентированные на говорящего или за него (*за-, при-*).

Примечание 6. Таким образом, из рассмотрения были выведены бесприставочные глаголы и глаголы с приставкой *по-* — как не показавшие на предыдущих этапах обследования каких бы то ни было предпочтений в отношении осей; приставка *с-* — как имеющая три разнородных значения (см. выше), которые можно было бы развести, но в результате на каждое из значений оставалось считанное количество примеров, что делало рассмотрение этой приставки статистически недостоверным, — а также приставки *пере-* и *раз-*, которые являются, судя по статистическим характеристикам и отсутствию парных предлогов, прежде всего маршрутными, а не векторными.

В табл. 5 мы приводим соответствующие результаты.

Таблица 5

Тип приставок		ориентация вниз	ориентация вверх	*вы-* + левоориентированные	ориентация наговорящего
Рука+направление \ Приставки		*в-, до-, на-, о-*	*воз-, под-, про-*	*вы-, у-, от-*	*за-, при-*
комфортное движение		64/65	**41/29**	*64/74*	45/46
притяжение		19/19	7/9	*12/23*	**25/14**
пересечение	левая рука далеко направо	7/7	1/3	11/8	5/5
	правая рука далеко налево	**31/26**	*8/12*	32/29	*13/18*
радикальное пересечение		19/23	*6/10*	**40/25**	*10/16*
χ^2=39,4, p⩽.9,1^{-5}; параметры связаны, распределения достоверны					

Мы видим, что параметры в табл. 5 связаны между собой, причем уровень связи достаточно велик, чтобы можно было строить какие-то умозаключения. Прежде всего, обратим внимание на то , что по всем параметрам, кроме пересечения, левое и правое направление движения, как и выбор левой или правой руки, дают согласованные результаты:

1) для приставок, ориентированных вверх (*воз-, под-, про-*), на когнитивной (поперечной) оси характерен выход руки из зоны коммуникации в комфортном для руки направлении, т.е. без пересечения зоны коммуникации; тем самым мы видим, что движение вверх по вертикальной оси семиотически приравнено в русской жестикуляции к выходу за пределы зоны коммуникации, осуществляемому без напряжения (без приложения каких-либо специальных усилий) любой рукой в естественном для этой руки направлении[8] (рис. 14);

Рис. 14. Приставки, ориентированные вверх, на когнитивной оси

[8] Если раздельно рассматривать данные по комфортному движению для левой и для правой руки, то можно заметить, что движение правой рукой направо в случае приставок, ориентированных вверх, более частотно, чем движение левой руки налево. Это, скорее всего, связано с доминантностью правой руки, но может быть связано также с противопоставлением «верхних» приставок — «нижним» (*в-, до-, на-, о-*), которые с жестикуляционной точки зрения тяготеют влево, см. следующий пункт.

2) для приставок, ориентированных вниз (*в-, до-, на-, о-*), напротив, на когнитивной (поперечной) оси мы имеем дело с совершенно иной картиной: во-первых, для жестикуляционного сопровождения этих приставок выбирается движение, требующее усилий (пересечение зоны коммуникации для попадания в «чужую» зону); во-вторых, расположение этих приставок на когнитивной оси асимметрично — они тяготеют к левой, а не к правой зоне; таким образом, мы видим, что движение вниз по вертикальной оси семиотически приравнено к движению влево, осуществляемому правой рукой с пересечением зоны коммуникации; таким образом, «конечный пункт» этих приставок расположен **далеко** от исходного пункта движения (см. рис. 15);

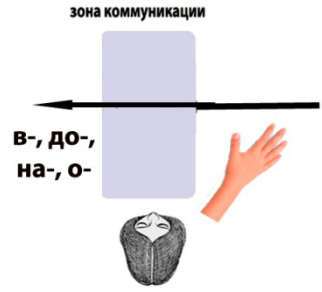


Рис. 15. Приставки, ориентированные вниз, на когнитивной оси

3) для приставок, ориентированных на говорящего (*за-, при-*), на когнитивной оси характерен вход в зону коммуникации без ее пересечения, т. е. движение на говорящего или за говорящего на горизонтальной (коммуникативной) оси на поперечной оси семиотически приравнивается к введению правой или левой руки в зону коммуникации (рис. 16);

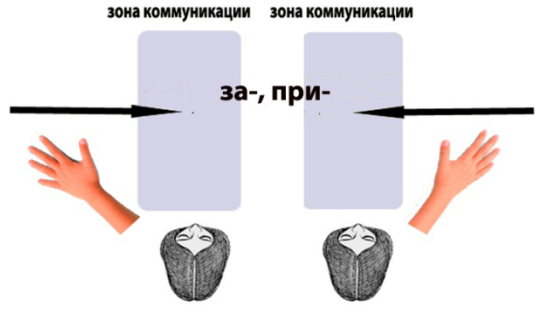


Рис. 16. Приставки, ориентированные на говорящего, на когнитивной оси

4) левоориентированные приставки (*от-, у-*), которые на когнитивной оси располагаются в основном слева от говорящего, а также приставка *вы-*,

которая с точки зрения «разложения» по осям не имеет каких-то предпочтений, при анализе движения левой и правой рук относительно зоны коммуникации демонстрируют отчетливое тяготение к радикальному пересечению[9], т. е. говорящий сначала фиксирует исходную точку движения на (семиотически) наиболее далеком расстоянии (левая рука в правой зоне, а правая рука в левой зоне), а потом пересекает зону коммуникации, чтобы вернуть руку в естественную для нее зону (см. рис. 17); неслучайно все три приставки включают в свою семантику компонент ‘исходная точка’, для отрыва от которой требуется известное напряжение, связанное с преодолением сопротивления; таким образом, мы видим, что приставка *вы-*, не обладая никакими пространственными предпочтениями, тем не менее может быть охарактеризована как приставка а) фиксирующая исходную точку движения в любом направлении и б) приставка с повышенной энергетикой; таковыми являются также приставки *от-* и *у-* (и, по-видимому, приставка *с-* во втором значении, см. выше).

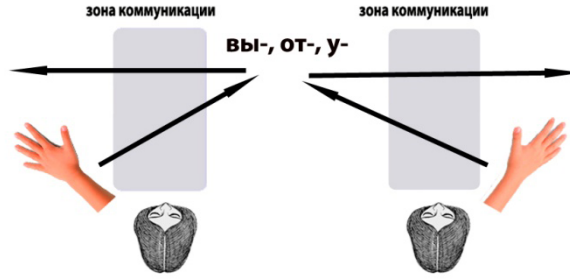


Рис. 17. Левоориентированные приставки и приставка *вы-* на когнитивной оси

В заключение этого раздела нужно отметить, что данные по русским приставкам, приведенные в этом разделе, подтверждают идею семиотической конвертации жестикуляции с оси роста (вертикальная ось) и оси прогресса (сагиттальная ось) на поперечную ось, — идеи, которая была высказана Ж. Кальбрис в работе [Calbris 2008].

5. Приставки с точки зрения конфигурации ладони

В нашей базе данных зафиксированы следующие основные конфигурации руки при осуществлении жестикуляции, сопровождающей глаголы (см. сноска 4): указание большим или указательным пальцем, движение открытой

[9] Если проанализировать подробнее предпочтения этих приставок, можно заметить, что радикальное пересечение правой рукой слева направо выражено здесь сильнее, чем радикальное пресечение левой рукой справа налево.

ладонью, а также рука, сформированная в кулак и в щепоть. Поскольку данных на указание большим пальцем и на руку, сформированную в щепоть, относительно немного, далее мы их учитывать не будем. В работе [Гришина 2012] были предложены основные прагматические и референциальные характеристики двух базовых конфигураций указующей руки в системе русских ручных указательных жестов — указательный палец и открытая ладонь, а из основных работ по жестикуляционной лингвистике (см., среди прочего, [Calbris 2011]) известно, что кулак обозначает силу или напряжение. Используя эту информацию, мы даем в табл. 6 основные значения этих трех конфигураций руки.

Таблица 6

Конфигурация руки	**Значение**
Указательный палец	наличие конечной точки движения, определенность, точечность объекта
Открытая ладонь	отсутствие конечной точки движения, неопределенность, диффузность объекта
Кулак	сила, напряженность

В табл. 7 мы приводим распределение конфигураций руки между группами приставок (каждую группу формируют приставки, имеющие один и тот же набор предпочтений).

Таблица 7

Приставки / **Конфигурации ладони**	***в-, до-, за-, пере-***	***вы-, на-, с-***	***по-***
Кулак	27/32	**41/26**	*3/14*
Открытая ладонь	*286/309*	250/249	**156/134**
Указательный палец	**117/89**	*55/72*	*27/38*
χ^2=40, p⩽.4,3^{-8}; параметры связаны, распределения достоверны			

Используя совокупность значений, предложенных нами для конфигураций руки в табл. 6, мы предлагаем следующую интерпретацию данных из табл. 7:

1) группа приставок ***в-, до-, за-, пере-***: наличие конечной определенной точки движения (*указательный палец*);
2) группа приставок ***вы-, на-, с-***: движение с напряжением, с силой, с энергией (*кулак*); для ***вы-, с-*** — наличие только исходной, начальной точки движения, для ***на-*** — неточечный, объемный или плоскостной характер объекта-цели (отрицательные данные по параметру *указательный палец*);
3) приставка ***по-***: отсутствие конечной точки движения, неопределенность и диффузность цели движения (*открытая ладонь*).

6. Заключение: жестикуляционные профили приставок

В качестве заключения подытожим все, сказанное выше о 16-ти приставках, послуживших объектом исследования. Далее при каждой приставке (или группе приставок) дается в семантических кавычках значение приставки, а в фигурных скобках — жестикуляционная характеристика, на основании которой мы предлагаем именно такое значение. Совокупность значений и жестикуляционных характеристик может рассматриваться как жестикуляционный профиль приставки или пар приставок — на сегодня максимально полный; не исключено, разумеется, что в ходе дальнейших исследований появятся новые данные, которые внесут уточнения в наши результаты.

в-, до-	векторные приставки	‘движение вниз’ {направление вниз по вертикальной оси; пересечение влево}	‘наличие дистанции’ {пересечение влево}	‘наличие конечной точки’ {указательный палец}	
воз-	векторная приставка	‘движение вверх’ {направление вверх по вертикальной оси, комфортное движение по сагиттальной оси}	‘спуск перед подъемом’ {вогнутая дуга}		
вы-		‘фиксация начальной точки движения’ {радикальное пересечение}	‘напряжение, усилие’ {радикальное пересечение, кулак}		
***за*^{1}-**	векторная приставка	‘нахождение объекта за препятствием или его перемещение за препятствие’ {за спину говорящему по сагиттальной оси}	‘наличие конечной точки’ {указательный палец}		
***за*^{2}-**	маршрутная приставка	'присвоение, собирание' {движение на говорящего по сагиттальной оси, притяжение и движение снаружи в центр по когнитивной оси}			
на-	векторная приставка	'движение вниз' {направление вниз по вертикальной оси; пересечение влево}	‘наличие дистанции’ {пересечение влево}	‘подъем перед спуском’ {выгнутая дуга}	‘напряжение, усилие’ {кулак}
***о*^{1}-**	векторная приставка	‘движение вниз’ {направление вниз по вертикальной оси; пересечение влево}	‘наличие дистанции’ {пересечение влево}		
***о*^{2}-**	маршрутная приставка	‘объем’ {изнутри наружу по когнитивной оси}			
от-, у-	векторные приставки	‘отсоединение от чего-л.’ {в сторону от говорящего по когнитивной оси}	‘фиксация начальной точки движения’ {радикальное пересечение}	‘напряжение, усилие’ {радикальное пересечение, кулак}	‘негативная оценка’ {левая зона когнитивной оси}

при-	векторная приставка	'приближение, соединение с чем-л.' {движение на говорящего по сагиттальной оси, притяжение}		
раз-	маршрутная приставка	‘расширение, удлинение’ {изнутри наружу по когнитивной оси, правая зона когнитивной оси}		
с[1]***-***	векторная приставка	‘отсоединение от чего-л.’ {в сторону от говорящего по когнитивной оси}	‘негативная оценка’ {левая зона когнитивной оси}	‘напряжение, усилие’ {кулак}
с[2]***-***	векторная приставка	'соединение с чем-л.' {движение на говорящего по сагиттальной оси}		
с[3]***-***	маршрутная приставка	'собирание' {снаружи в центр по когнитивной оси}		

Можно видеть, что для различения некоторых пар приставок предложенных нами в данной статье жестикуляционных параметров недостаточно: так, не различаются приставки *от-* и *у-*, а также *под-* и *про-*. Напротив, в некоторых приставках (*за-, о-, с-*) жестикуляция различает даже разные значения. Возможно, для различения пар приставок нам просто не хватило материала или мы пока не заметили жестикуляционный различительный признак, необходимый для их «разведения».

Интересно также отметить, что две приставки не имеют вообще никаких предпочтений в отношении их распределения в пространстве (*по-* и *вы-*), так что пока нам даже не удалось определить, являются ли эти приставки векторными или маршрутными (характеристики приставки *вы-* — фиксация начальной точки движения и напряженность осуществления, а единственная характеристика приставки *по-* — отсутствие конечной точки движения, диффузность цели). Мы не стали приводить соответствующие данные, но заметим здесь, что глаголы с приставкой *по-* ведут себя так же, как бесприставочные глаголы: у последних, как мы уже писали выше, отсутствует предпочтительное направление в пространстве, кроме того, они характеризуются отсутствием конечной точки движения и диффузностью цели, что выражается в предпочтении движений открытой ладонью, а не указательным пальцем.

Безусловно, значительные коррекции в предложенную картину может внести изучение жестикуляционного сопровождения предлогов и пространственных наречий, которые, естественно, должны быть связаны с пространственной характеристикой префиксов.

В любом случае, можно констатировать, что семантика приставок представлена в русской жестикуляции достаточно разнообразно. Расширение материала, безусловно, изменит предложенные распределения, но вряд ли отменит саму жестикуляционную асимметрию русских приставок относительно русских жестов.

Литература

1. *Гришина Е. А.* (2012). Указания рукой как система (на материале Мультимедийного русского корпуса). Вопросы языкознания, 3, с. 3–50.
2. *Гришина Е. А.* (2013а). Русская темпоральная жестикуляция. Известия РАН, ОЛЯ, 1, с. 3–31.
3. *Гришина Е. А.* (2013б). Указания головой как система. *Вопросы языкознания, 3,* с. 119–159
4. *Гришина Е. А.* (2014). Вертикальная ось в жестикуляции: лингвистические аспекты (рукопись).
5. *Майсак Т. А.* (2005). Типология грамматикализации конструкций с глаголами движения и глаголами позиции. Языки славянских культур, Москва.
6. *Панов В. А.* (2011). К типологии и диахронии глаголов движения в латинском языке. *Лингвистика и методика преподавания иностранных языков. Периодический сборник научных статей (электронное научное издание), 3,* available at: http://www.iling-ran.ru/library/sborniki/for_lang/2011_03/
7. *Allen S., Ozyürek A., Kita S., Brown A., Furman R., Ishizuka T., Fujii M.* (2007). Language-specific and universal influences in children's syntactic packaging of manner and path: A comparison of English, Japanese, and Turkish. Cognition, 102, pp. 16–48.
8. *Brown A., Gullberg M.* (2008). Bidirectional cross-linguistic influence in l1-l2 encoding of manner in speech and gesture: A study of Japanese speakers of English. Studies in Second Language Acquisition, 30(2), June 2008, pp. 225–251.
9. *Brown A., Gullberg M.* (2010). Changes in encoding of path of motion after acquisition of a second language. Cognitive Linguistics, 21(2), pp. 263–286.
10. *Calbris G.* (2008). From left to right… Coverbal gestures and their symbolic use of space. *Metaphor and gesture. A. Cienki and C. Müller (Eds).Benjamins, Amsterdam/Philadelphia,* pp. 27–53
11. *Calbris G.* (2011). Elements of meaning in gesture. Benjamins,Amsterdam/Philadelphia.
12. *Chen Liang* (2007). The acquisition and use of motion event expressions in Chinese. LINCOM publ., Munich.
13. *Choi S., Lantolf J. P.* (2008). Representation and embodiment of meaning in L2 communication: Motion events in the speech and gesture of advanced L2 Korean and L2 English speakers. Studies in Second Language Acquisition, 30(2), June 2008, pp. 191–224.
14. *Chui K.* (2009). Linguistic and imagistic representations of motion events. Journal of Pragmatics, 41 (9), pp. 1767–1777.
15. *Chui K.* (2012). Cross-linguistic comparison of representations of motion in language and gesture. Gesture, 12(1), pp. 40–61.
16. *Cienki A.* (1999). The strengths and weaknesses of the left/right polarity in Russian: Diachronic and synchronic semantic analyses. Stadler L. de, Eyrich C. (Eds.). Issues in Cognitive Linguistics: 1993 Proceedings of the International Cognitive Linguistics Conference, Berlin, pp. 299–329.
17. *Gullberg M., Hendriks H., Hickmann M.* (2008). Learning to talk and gesture about motion in French. First Language, 28(2), May 2008, pp. 200–236.

18. *Kita S., Ozyürek A., Allen S., Brown A., Furman R., Ishizuka T.* (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and Cognitive Processes, 22(8), pp. 1212–1236.
19. *Kita S., Özyürek A.* (2003). What does cross-linguistic variation of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48(1), January 2003, pp. 16–32.
20. *McNeill D., Duncan, Susan D.* (2000). Growth points in thinking-for-speaking. McNeill D. (Ed.). Language and gesture. Cambridge Univ. Publ., Cambridge, pp. 141–161.
21. *Özyürek A., Kita S., Allen S., Furman R., Brown A.* (2005). How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities. Gestural communication in nonhuman and human primates, Special issue of Gesture 5(1/2), 2005, pp. 219–240.
22. *Özyürek A., Kita S., Allen S., Brown A., Furman R., Ishizuka T.* (2008). Development of cross-linguistic variation in speech and gesture: motion events in English and Turkish. Developmental psychology, 44(4), pp. 1040–1054.
23. *Slobin D. I.* (2004). The many ways to search for a frog: Linguistic typology and the expression of motion events. *Strömqvist S., Verhoven L. (Eds.). Relating events in narrative. Vol. 2*, Mahwah, NJ, pp. 219–257
24. *Talmy L.* (1985). Semantics and syntax of motion. Kimball J. (Ed.). Syntax and Semantics, 4, NY, pp. 131–138.
25. *Talmy L.* (1991). Path to realization: a typology of event conflation. Proceedings of the 17th Annual Meeting of the Berkley Linguistic Society, Feb. 15–18, Berkeley, CA, Berkeley Linguistic Society, pp. 480–519.

References

1. *Allen S., Ozyürek A., Kita S., Brown A., Furman R., Ishizuka T., Fujii M.* (2007). Language-specific and universal influences in children's syntactic packaging of manner and path: A comparison of English, Japanese, and Turkish. Cognition, 102, pp. 16–48.
2. *Brown A., Gullberg M.* (2008). Bidirectional cross-linguistic influence in l1–l2 encoding of manner in speech and gesture: A study of Japanese speakers of English. Studies in Second Language Acquisition, 30(2), June 2008, pp. 225–251.
3. *Brown A., Gullberg M.* (2010). Changes in encoding of path of motion after acquisition of a second language. Cognitive Linguistics, 21(2), pp. 263–286.
4. *Calbris G.* (2011). Elements of meaning in gesture. Benjamins, Amsterdam/ Philadelphia.
5. *Calbris G.* (2008). From left to right… Coverbal gestures and their symbolic use of space. *Metaphor and gesture. A. Cienki and C. Müller (Eds).Benjamins, Amsterdam/Philadelphia,* pp. 27–53
6. *Chen Liang* (2007). The acquisition and use of motion event expressions in Chinese. LINCOM publ., Munich.

7. *Choi S., Lantolf J. P.* (2008). Representation and embodiment of meaning in l2 communication: Motion events in the speech and gesture of advanced L2 Korean and L2 English speakers. Studies in Second Language Acquisition, 30(2), June 2008, pp. 191–224.
8. *Chui K.* (2009). Linguistic and imagistic representations of motion events. Journal of Pragmatics, 41 (9), pp. 1767–1777.
9. *Chui K.* (2012). Cross-linguistic comparison of representations of motion in language and gesture. Gesture, 12(1), pp. 40–61.
10. *Cienki A.* (1999). The strengths and weaknesses of the left/right polarity in Russian: Diachronic and synchronic semantic analyses. Stadler L. de, Eyrich C. (Eds.). Issues in Cognitive Linguistics: 1993 Proceedings of the International Cognitive Linguistics Conference, Berlin, pp. 299–329.
11. *Grishina E. A.* (2012). Hand pointing as a system (on MURCO data) [Ukazanija rukoj kak sistema (na materiale Mul'timedijnogo russkogo korpusa]. Voprosy jazykoznanija. Вопросы языкознания, 3, pp. 3–50.
12. *Grishina E. A.* (2013a). Russian temporal gesticulation [Russkaja temporal'naja zhestikuljatsija]. Izvestija RAN, OLJA, 1, pp. 3–31.
13. *Grishina E. A.* (2013б). Head pointing as a system [Ukazanija golovoj kak sistema]. *Voprosy jazykoznanija, 3,* pp. 119–159
14. *Grishina E. A.* (2014). Vertical axis in Russian gesticulation [Vertikal'naja os' v zhestikuljatsii] (manuscript)
15. *Gullberg M., Hendriks H., Hickmann M.* (2008). Learning to talk and gesture about motion in French. First Language, 28(2), May 2008, pp. 200–236.
16. *Kita S., Özyürek A.* (2003). What does cross-linguistic variation of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48(1), January 2003, pp. 16–32.
17. *Kita S., Ozyürek A., Allen S., Brown A., Furman R., Ishizuka T.* (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and Cognitive Processes, 22(8), pp. 1212–1236.
18. *Majsak T. A.* (2005). Verbs of movement and verbs of position: Typology of grammaticalization of constructions (Tipologija grammatikalizatsii konstruktsij s glagolami dvizhenija i glagolami pozitsii]. Jazyki slavjanskih kul'tur, Moscow.
19. *McNeill D., Duncan, Susan D.* (2000). Growth points in thinking-for-speaking. McNeill D. (Ed.). Language and gesture. Cambridge Univ. Publ., Cambridge, pp. 141–161.
20. *Özyürek A., Kita S., Allen S., Brown A., Furman R., Ishizuka T.* (2008). Development of cross-linguistic variation in speech and gesture: motion events in English and Turkish. Developmental psychology, 44(4), pp. 1040–1054.
21. *Özyürek A., Kita S., Allen S., Furman R., Brown A.* (2005). How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities. Gestural communication in nonhuman and human primates, Special issue of Gesture 5(1/2), 2005, pp. 219–240.
22. *Panov V. A.* (2011). Verbs of movement in Latin:Typology and diachrony [K tipologii i diahronii glagolov dvizhenija v latinskom jazyke]. Lingvistika i metodika

prepodavanija inostrannyh jazykov. Periodicheskij sbornik nauchnyh statej, *3,* available at: http://www.iling-ran.ru/library/sborniki/for_lang/2011_03/

23. *Slobin D. I.* (2004). The many ways to search for a frog: Linguistic typology and the expression of motion events. *Strömqvist S., Verhoven L. (Eds.). Relating events in narrative. Vol. 2*, Mahwah, NJ, pp. 219–257
24. *Talmy L.* (1985). Semantics and syntax of motion. Kimball J. (Ed.). Syntax and Semantics, 4, NY, pp. 131–138.
25. *Talmy L.* (1991). Path to realization: a typology of event conflation. Proceedings of the 17th Annual Meeting of the Berkley Linguistic Society, Feb. 15–18, Berkeley, CA, Berkeley Linguistic Society, pp. 480–519.

ЧИТАТЬ НЕ ЧИТАЛ, НО…: ОБ ОДНОЙ РУССКОЙ КОНСТРУКЦИИ С ПОВТОРЯЮЩИМИСЯ СЛОВЕСНЫМИ ЭЛЕМЕНТАМИ[1]

Иомдин Л. Л. (iomdin@gmail.com)

ИППИ РАН, Москва, Россия

Ключевые слова: микросинтаксис, нестандартные синтаксические конструкции, синтаксические фраземы, синтаксис, семантика

CHITAT' NE CHITAL, NO…: ON A RUSSIAN CONSTRUCTION WITH REPEATED LEXICAL ELEMENTS

Iomdin L. L. (iomdin@gmail.com),

Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

The paper discusses several types of Russian microsyntactic units — non-standard syntactic constructions and syntactic idioms with repeated verbal elements. The primary construction discussed is of the type *chitat' ne chital (no sdelal chto-to menee sil'noe)* ≈ 'one did not really read it (but one did something less strong)'. In this construction, two copies of the same verb in different inflectional forms (one in the infinitive and the other in finite form) are present, the latter preceded by the negative particle. Since lexical instantiation of the verbal positions is virtually free, the only restriction being imposed on their lexical coincidence, the construction should be treated as lexically unbound and, hence, as a non-standard microsyntactic construction. There are two more constructions that appear to be lexically and syntactically close to the primary one: the so-called emphatic tautological infinitive construction of the type *s'jest'-to on s"est* ≈ 'he will definitely eat it' and a syntactic idiom with lexically bound repeated verbal elements of the type *Ja tebja znat' ne znaju* ≈ 'I don't know you and have no wish to do anything with you'. We focus on the semantics of these three units and ways to discriminate them in human and automatic natural language processing tasks.

Key words: microsyntax, non-standard syntactic constructions, syntactic idioms, syntax, semantics

[1] Данное исследование частично поддержано Российским фондом фундаментальных исследований (грант № 12-07-00663-а) и Российским гуманитарным научным фондом (грант № 11-04-00070а). Обоим фондам автор выражает глубокую признательность.

1. Вводные замечания

В последние годы автор активно исследовал разнообразные, по большей части маргинальные, явления русского синтаксиса, которые прежде не получали адекватного отражения ни в грамматических описаниях, ни в лексикографических штудиях и словарях (см., например, Iomdin 2003, 2005, 2006 a, b, 2007, 2008, Apresjan et al. 2010).

Общая синтаксическая система языка состоит из двух неравных и существенно различных частей. Это, во-первых, основной синтаксис языка, который включает относительно небольшое число базовых конструкций, и, во вторых, периферийный синтаксис, содержащий во много раз большее количество конструкций.

Базовые конструкции высокочастотны, неидиоматичны и строятся на основе весьма общих грамматических правил. В русском языке таковы, например, (а) элементарная предикативная конструкция типа *дети играют*, состоящая из именного подлежащего и глагольного сказуемого и характеризующаяся согласованием членов по числу и лицу/роду, или (б) определительная конструкция типа *маленький ребёнок*, состоящая из определяемого существительного и адъективного определения к нему и характеризующаяся согласованием членов по падежу, числу, роду и одушевленности.

Каждая из периферийных синтаксических конструкций встречается в текстах значительно реже, чем любая базовая, хотя в целом их частотность очень высока. Часть синтаксиса языка, образуемую периферийными конструкциями, автор предпочитает именовать «м и к р о с и н т а к с и с о м», который является составной, пусть и и специфической, частью общего синтаксиса[2].

В состав микросинтаксиса входят, по нашему мнению, объекты двух основных типов — **(1)** синтаксические фраземы и **(2)** нестандартные синтаксические конструкции. Граница между этими объектами достаточно условна; главным различающим критерием тут является степень **лексикализации** этого объекта: синтаксические фраземы образуются из конкретных слов, а нестандартные синтаксические конструкции к конкретным словам не привязаны или почти не привязаны.

Безусловно к **первому микросинтаксическому типу**, т. е. к синтаксическим фраземам, относятся единицы типа

(1) ***руки чешутся*** (сделать что-л.)*;*

ср. *У людей руки чешутся кидать камни, а тут объявляется, что это священная обязанность* (Ф. Искандер).

---

[2] Взгляд на микросинтаксис как на часть общего синтаксиса весьма близок к идеям грамматики конструкций (construction grammar; см., например, Fillmore *et al.* 1988, Goldberg 1995, 2006, Rakhilina 2010 и др.). Обсуждение сходств и различий между грамматикой конструкций и микросинтаксическим подходом, развиваемым автором независимо, выходит, однако, за рамки настоящей статьи.

От обычных фразеологических единиц, лишенных синтаксической специфики (таких, например, как смотреть сквозь розовые очки, плыть по течению и т. п., синтаксически ведущих себя как свободные словосочетания того же лексического состава), (1) отличается идиосинкратической способностью управлять инфинитивом — способностью, которой лишены оба ее лексических элемента[3].

Близкие к (1) по синтаксическому и лексическому устройству конструкции типа

(2) ***душа болит*** или

(3) ***сердце кровью обливается***

имеют и другие идиосинкратические управляющие свойства: помимо инфинитива (*душа болит смотреть на ее страдания, сердце кровью обливается слышать такое*), они управляют, в частности, (а) группой с предлогами *за* или *о*; ср. *душа болит за брата, И обо всех у тебя душа болит* (В. Шукшин), *сердце обливается кровью за Отечество* и (б) придаточным, вводимым союзами *когда* и *как*; ср. *Сердце кровью обливается, когда слышишь, что осудили невиновного человека* (Н. Гритчин), *Как вспомню о ней — сердце кровью обливается* (В. Панова); *У меня душа всегда болит, когда я вижу, что люди злые* (Л. Чарская)[4]. Разумеется, по отдельности ни у одного из составляющих эти синтаксические фраземы лексических элементов таких управляющих свойств нет.

Аналогичным образом ведут себя такие полуфразеологические сочетания, как *иметь значение, играть роль, иметь в виду, вызывать восхищение, нет сомнения, слава богу* и пр., каждое из которых способно, в отличие от составляющих элементов, управлять придаточным предложением с союзом *что: Не имеет никакого значения, что у вас нет опыта работы в торговле; Слава богу, что вы живы* (Б. Акунин) и т. п.

Синтаксической спецификой обладают такие разные фразеологические и полуфразеологические выражения, в свое время подробно исследованные автором, как ***всё равно*** (во всех значениях этой фраземной вокабулы, ср. *Я всё равно стану космонавтом, Нам* ***всё равно****, куда ехать, Писать — это* ***всё равно,*** *что чувствовать*), *Этому человеку* ***хоть бы что****;* ***Какого чёрта*** *ты здесь сидишь?* и т. д.

---

[3] Такое понимание синтаксической фраземы не является общепринятым. В частности, И. А. Мельчук (Mel'čuk 2012) считает синтаксическими фраземами именно нестандартные, т. е. создающие семантическую некомпозициональность, синтаксические конструкции.

[4] Здесь и далее большинство литературных примеров взяты из Национального корпуса русского языка (ruscorpora.ru).

С другой стороны, например, инфинитивно-модальная безличная конструкция с дательным падежом субъекта *Z-у X-овать* ‘Z-у предстоит X-овать’ (***Тебе выходить*** *на следующей; Вздохнул я и открылся: из такого я, девочки, вагона, что* ***вам жить****, а* ***мне умирать*** (А. Солженицын)) относятся ко **второму микросинтаксическому типу**, так как лексическое заполнение обоих элементов конструкции, в общем, не ограничено.[5]

К этому же типу следует, по нашему убеждению, отнести и инфинитивно-модальную конструкцию *Z-у не X-овать* ≈ ‘Отсутствует возможность, что Z сможет X-овать’ (***Одному*** [Z] *тут* ***не справиться*** [X], *Ах, не плыть по голубому морю, /* ***Не видать*** [X] ***нам*** [Z] *Золотого Рога, / Голубей и площади Сан-Марка (М. А. Кузмин); Он чувствовал, что* ***Фоме*** [Z] ***не выбраться*** [X]*, его уничтожат* (Д. Гранин); ***Мне*** *этот бой* ***не забыть*** *нипочём* (В. Высоцкий)): хотя один элемент этой конструкции (смысл которой, заметим, отнюдь не сводится к отрицанию предыдущей конструкции) лексически жёстко ограничен (он реализуется единственной лексической единицей *не*), тем не менее другие позиции конструкции лексически свободны, а ограничения затрагивают слово весьма общей семантики, по существу, слово служебное.

С другой стороны, конструкцию *Z-у не до X-а* ≈ ‘Z занят более важными делами, чем X, и считает, что X-ом можно пренебречь’ (***Мне*** *было* ***не до*** *обеда,* ***Богам не до*** *этого — их удел — творенье* (В. Голованов)), автор отнес бы к первому микросинтаксическому типу, т. е. к синтаксическим фраземам, поскольку здесь присутствует, помимо отрицательной частицы, бесспорно лексически связанный словесный элемент — предлог *до* (причем, по-видимому, в уникальном значении).[6]

К сфере микросинтаксиса относятся и многочисленные типы конструкций с лексически повторяющимися элементами.

В одних случаях эти элементы повторяются в одной и той же грамматической форме: ***Работа есть работа*** (Б. Окуджава); ***Культ не культ,*** *а чего не случается* (А. Галич); ***Умный-то*** *ты* ***умный****, а кумекать не умеешь* (Б. Васильев); ***Стучи не стучи****, тебя никто не услышит.*

Этим же свойством обладают настойчивые обращения с повтором именительного и особенно звательного падежа, ср. ***Дядя*** *Анискин,* ***а дядя*** *Анискин, ты не уснул ли?* (В. Липатов), ***Зой, а Зой,*** *ты только глянь на него…* (Б. Окуджава) — с уникальным употреблением междометия *а*, которое за пределами этой конструкции, по-видимому, не встречается.

---

[5] Иногда подобные образования именуются фразеосхемами (ср. Bulygina-Shmelev 1997). Заметим, впрочем, что инфинитивно-модальные конструкции настолько обыкновенны для русского языка, что могут быть отнесены и к основному синтаксису. Тем самым, как это часто бывает в лингвистике, и сама граница между основным синтаксисом и микросинтаксисом оказывается не столь уж неприкосновенной.

[6] Интересно отметить, что в двух последних микросинтаксических единицах частица *не*, строго говоря, не является обязательной: существуют вопросительные варианты обеих единиц, в которых *не* вообще не фигурирует, а вместо него выступают частицы *ли* и *разве*: *Справиться ли тут одному?*; *Разве одному тут справиться?*; *Его забыть ли нам когда-нибудь?* (И. Северянин); *До обеда ли мне было? Разве мне было до обеда?* Почти всегда такие варианты оформляют риторические вопросы.

Отдельный подкласс здесь образуют сочинительные конструкции с лексически повторяющимися элементами, разнообразные значения и употребления которых подробно изучал В. З. Санников (Sannikov 1989), ср. *Ну* ***упал и упал*** ≈ ‘в падении не было ничего особенного’; *Он* ***сидит и сидит*** ≈ ‘он продолжает сидеть’; *Бывают* ***аварии и аварии*** ≈ ‘аварии бывают разные, одни можно оправдать, другие нет’.

Близкие к сочинительным глагольные конструкции с повторяющимися элементами типа *дед* ***бил-бил****, не разбил;* ***тушат-тушат*** *— не потушат* рассматривались в статье Plungjan-Rakhilina 2010.

Совсем недавно внимание исследователей привлекли и другие конструкции с лексическими повторами[7] в виде аппозиции, которые получили широкое распространение в разговорной речи в самые последние годы, такие как *такая девочка-девочка* ≈ ‘весьма типичная, соответствующая стереотипу, девочка, к которой говорящий испытывает эмпатию’ (см. Giljarova 2010) или *американец такой американец* ≈ ‘соответствующий стереотипу американец, к которому говорящий не испытывает эмпатии’ (см. Vilinbakhova 2011)[8].

В определенных случаях лексически повторяющиеся элементы стоят в разных формах, жестко фиксируемых конструкцией. В одних случаях это существительные, ср.

(4) ***Учеба*** [=им] ***учебой*** [=твор]*, а отдохнуть тоже надо.*

Существенная часть конструкций представлена глаголами, ср.

(5) ***Съесть*** [=инф] ***-то*** *он* ***съест*** [=личн]*, да кто же ему даст.*

Такая конструкция иногда называется тавтологическим инфинитивом (термин О. С. Ахмановой, см. Akhmanova 1968). Она подробно разбиралась в статье Paillard and Plungjan 1993, а также, наряду с другими глагольными конструкциями с повторами, в статье McCoy 2002.

Близка к тавтолологической конструкции и конструкция типа

(6) ***Читать*** [=инф] *я эту книгу* ***не читал*** [=личн]*, но кое-что о ней слышал.*

Именно эту последнюю конструкцию мы намерены рассмотреть подробнее.

---

[7] И. Б. Левонтина несколько иронически называет такие и некоторые близкие конструкции «лексическим клонированием» (см. ее статью «Атака клонов», «Троицкий вариант», 6.07.2010).

[8] Детальную классификацию русских и иноязычных конструкций с повторами предлагает О. Ю. Крючкова (Krjuchkova 2004).

2. Микросинтаксическая конструкция типа *читать не читал (но…)*

2.1. Материал

Сначала приведем несколько примеров.

(7) ***Прочесть не прочёл**, но пролистал.*

(8) *После затяжной паузы Ломакин угадал резкое вихревое движение — **видеть не видел**, но угадал: «Петр первый» в полном соответствии с киношными образчиками небось стволом водит — полуприсев, сжав обеими руками, ловя на мушку… пустоту* (А. Измайлов).

(9) *Нет, **видеть не видел, но** вспоминал часто* (В. Ян)

(10) *Устюшкина мать Собиралась помирать. **Помереть не померла** — Только время провела* (цитируется у В. Ходасевича).

(11) ***Спать не спал,** да и они не спали, — отозвался ямщик* (Ф. М. Решетников).

(12) *Егоров **пить не пьет**, а ус в бокал макает и то к одному, то к другому моряку подсядет* (Артем Веселый).

(13) *Тут она ответила неожиданно. — **Знать — не знала, догадываться — не догадывалась…** Но вообще что-то в этом роде предчувствовала…* (В. Белоусова).

2.2. Семантика

Нетрудно увидеть, что значение этой конструкции во всех приведенных примерах одно и то же. Его можно истолковать приблизительно так:

(14) ***Z X-овать не X-овал (но…P)***:[9]
‘Z не сделал X, но Z сделал P; P — нечто менее сильное, чем X’.

Так, в (4) X — *читал*, а P — *пролистал*, в (7) X — *померла*, а Y — *время провела*. Хотя элемент толкования ‘менее сильный’ не слишком точен, его уместность можно проиллюстрировать, если поменять в цитируемых примерах

[9] Реально в этой конструкции, кроме *но*, могут фигурировать и другие уступительные союзы или частицы: *а*, *да*, *только*, *лишь*, *правда* и т. п. От рассмотрения конкретного вклада таких элементов в семантику конструкции мы сейчас воздерживаемся.

X и P местами. Такие предложения в одних случаях окажутся семантически странными, ср:

(7') #*Пролистать не пролистал, но прочёл;*

(12') #*Егоров ус макать не макает, а пьет.*

В других же случаях перестановка X и P влечет за собой смену перспективы: в предложении

(9') *Вспоминать не вспоминал, но видел иногда*

действие *видеть* воспринимается как менее сильное, чем *вспоминать*.

В силу такой семантики кажется разумным считать эту конструкцию **уступительной**.

Обратим еще внимание на то, что валентность P в этой конструкции может и не выражаться, как в примере (11) (здесь, скорее всего, имеется в виду ситуация, участник которой не спал, но делал что-то другое, в результате чего мог бы спать — например, лежал, сидел с закрытыми глазами, но вряд ли работал над статьей или плясал), или же выражаться за пределами предложения, как в (13), где P — *предчувствовала*.

Интересно отметить, что во многих случаях валентность P в этой конструкции выражается в предтексте:

(15) *Стыда потом не оберешься — да* [P]. *А погибнуть — не погибнем* (И. Ратушинская).

Регулярно этот предтекст оказывается в том же самом предложении, что и X; ср.

(16) *Номер я записала* [P]*, а позвонить* [X] *не позвонила;*

(17) *Заглянуть* [P] *к Яге в окошко можно, а входить* [X] *не входи в избушку.* (А. М. Ремизов)*;*

(18) *Ничего не скажешь, {заботится о Зинке, бельё меняет, передачи носит}* [P]*, а любить* [X] *не любит* (И. Грекова).

2.3. Грамматика и сочетаемость

Перечислим теперь грамматические и сочетаемостные особенности этой уступительной конструкции.

1.

- Оба вхождения глагола должны совпадать лексически;

- первое вхождение должно быть инфинитивом, второе личной формой;
- они должны быть согласованы в виде и залоге: **приходить не пришел; рассматриваться не рассматривали*;
- они должны быть согласованы по морфологическому варианту, т. е. принадлежать к одной субпарадигме: *прочесть не прочел* или *прочитать не прочитал*, но не **прочитать не прочел*.

2. Глагол в личной форме может стоять в настоящем, прошедшем времени или в будущем времени совершенного вида, но не в аналитическом будущем несовершенного вида: **читать я не буду читать*. Допустимо и сослагательное наклонение второго глагольного элемента, причем частица *бы* может занимать вакернагелевскую вторую позицию, ср. *Выгнать бы он его не выгнал, но премии точно лишил*, или же теснее примыкать ко второму глаголу, ср. *Выгнать он его не выгнал бы, но премии точно лишил*.
3. Лексическое заполнение конструкции практически свободно. В ней могут фигурировать по существу любые глаголы, включая семантически неожиданные

   а) стативные: *знать не знал, быть не была*,

   б) безличные: *светать не светало, смеркаться не смеркалось,*

   в) даже модальный глагол: *смочь не сможет*.

Тем самым очевидно, что конструкция (14) относится ко второму микросинтаксическому типу — нестандартным синтаксическим конструкциям. Ее специфика — повтор элементов, а не лексическая избранность.

Заметим для полноты картины, что существуют отдельные глаголы, появление которых в конструкции все-таки исключено или мало естественно. Это, в частности,

а) глагольные конгломераты *чувствовать себя* и *вести себя*: невозможно сказать что-либо вроде **Чувствовать себя он не чувствовал себя, *Чувствовать себя плохо он не чувствовал себя, *Чувствовать себя плохо он не чувствовал*;

б) глаголы с экзотическим сильным управлением: *деть, девать, деться, подеваться, задевать* (‘поместить в неизвестное место’), *запропаститься: *Деть ручку куда-то я не дел*.

в) некоторые глаголы, выражающие абстрактные отношения: *относиться (к кому-то как-либо), равняться*.

г) глаголы с искусственным или потенциальным инфинитивом: **долженствовать не долженствовал, *благодарствовать не благодарствовал*.

2.4. Конструкция *читать не читал* в ряду близких конструкций

Описание конструкции нельзя считать полным, если не очерчены ее пределы. В случае с микросинтаксическими явлениями эта задача особенно трудна. Выше мы уже видели, что, например, сочинительные конструкции с повторами могут иметь самые разные значения и, стало быть, являются разными. У рассматриваемой сейчас конструкции есть несколько близнецов, из-за которых ее идентификация оказывается отнюдь не простым делом.

Таких близнецов, как представляется автору, имеется два. Рассмотрим их по порядку.

2.4.1. Тавтологический инфинитив

Существует нестандартная синтаксическая конструкция с заметно другим значением, чем (14), которая, на первый взгляд, лексически и синтаксически устроена так же, как наша конструкция. Она выступает в примерах

(19) *Подсядь-ка, брат, к нам, не спесивься; вот у меня товарищ-ат что-то больно приуныл,* ***есть не ест и пить не пьет****; что ты станешь с ним делать…* (Д. В. Григорович)*;*

(20) *Она терпеть не может футбол, потому что там потные дяди пинают мячик.* ***Курить не курит, пить не пьет.*** *Зато обожает Борхеса и Достоевского* (Д. Емец).

Идеи о неполноте, слабости действия или состояния здесь нет совсем. Возможно, назначение инфинитива в этой конструкции состоит в том, чтобы вынести вперёд тему последующего сообщения: *Что касается курения, то он не курит: Если вы спросите, курит ли он, то я отвечу: не курит.*

Как представляется, механизм, лежащий в основе этой конструкции, похож на топикализацию в предложениях типа *Иван — о нём все и думать забыли; Иван — тот вообще не появился.*

Но если речь идет только о топикализации темы, становится непонятно, почему личный глагол обязан сопровождается отрицанием: правильно ли считать его составным элементом конструкции? Внимательное наблюдение показывает, что отрицание здесь вовсе не обязательно и не составляет части конструкции. Именно так обстоит дело в предложении (21):

(21) *Иногда бывает, что с утра встать — встал, а толком не проснулся.*

Что же это за конструкция? Это конструкция с **тавтологическим инфинитивом**, пример которой мы уже видели: *съесть-то он съест*. Она гораздо более свободна, чем та, уступительная, которую мы рассмотрели, и допускает разные типы распространения личного глагола: *Спать он сегодня почти не спал, вставать она уже понемногу встает, читать он много читает, жаловаться он никогда не жалуется.* Разумеется, во многих случаях эта конструкция сопровождается продолжением, которое вносит в содержащее ее предложение значение уступки (*съесть-то он съест,да кто ж ему даст*) , но в отличие от предыдущей конструкции эта уступка факультативна.

Особую сложность для интерпретации представляют случаи, когда в предложении встречаются два вхождения конструкций $X_{инф} + X_{личн}$, как в (22):

(22) *Болеть я не болел, но и писать — не писал.*

Смысл предложения (22) примерно таков: ‘я не болел, ощущая явления, более слабые, чем болезнь, которые проявлялись в том, что я не мог писать’. Тем самым вторая часть предложения — та самая конструкция с тавтологическим инфинитивом — выражает валентность P в нашей уступительной конструкции!

Конструкция с тавтологическим инфинитивом может выражать валентность P и в предтексте уступительной конструкции. Примером может служить несколько перифразированное предложение (21):

(21’) *Иногда бывает, что с утра встать — встал, а проснуться — не проснулся.*

Здесь вторая пара глаголов образует уступительную конструкцию, а первая пара — тавтологический инфинитив — заполняет валентность P этой конструкции.

Такая же ситуация имеет место и в примере (23):

(23) *После того как жена его прибежала от соседей и сказала, что мобилизации не будет, потому что сходку уговаривать уговаривают, а оцеплять не оцепляют, хозяин решил пойти туда вместе с дядей Сандро* (Ф. Искандер)

Тщательный просодический анализ позволит увидеть разные интонационные контуры в двух частях предложений (21’) и (23), в частности, повышение тона на первом личном глаголе (встал, уговаривают) и отсутствие такового на втором личном глаголе (проснулся, оцепляют).

В отдельных ситуациях адекватная интерпретация предложений с парой конструкций $X_{инф}$ + $X_{личн}$ может оказаться совсем не простым делом. Посмотрим, например, на предложение

(24) *Сказать никому не скажу, а забыть — не забуду.* (Г. Щербакова).

С одной стороны, можно считать, что уступительная конструкция выступает в первой части предложения (сказать не скажу), тогда вторая ее часть — тавтологическая конструкция — заполняет ее валентность P, и процесс «незабывания» оказывается слабее, чем обещание неразглашения. С другой стороны, можно считать — и, по-видимому, это естественнее, — что уступительная конструкция присутствует как раз во второй части (24), а первая часть — тавтологический инфинитив — заполняет валентность P уступительной конструкции. Если это так, предложение (24) следует признать неоднозначным.

2.4.2. Биглагольная синтаксическая фразема

Наряду с рассмотренной уступительной конструкцией существует омонимичная ей **синтаксическая фразема** — единица с фиксированным лексическим заполнением. Эта конструкция весьма частотна, ср:

(25) *Никто-то изо всей этой публики* ***знать не знал*** *о мне ровнешенько ничего!* (Ф. М. Достоевский)*;*

(26) *О мировой жизни за пределами любимого отечества он* ***ведать не ведал,*** *и позапрошлогодняя поездка в Америку мало что изменила* (А. Архангельский);.

(27) *Никогда он сюда не приезжал, ничего не просил, и никакого Бориса* ***ведать не ведал*** (О. Дивов).

(28) *Его враз кинуло в жар: бумажки-то остались на столе, и что в них было сказано, Егор и* ***знать не знал и ведать не ведал*** (Б. Васильев).

В этой синтаксической фраземе выступают в основном два синонимичных глагола — *знать* и *ведать*, причем первый из них — в обоих основных значениях (*знать1 что-нибудь* и *знать2 кого-нибудь*), а второй глагол может приобретать новое значение, параллельное *знать2*, которого вне этой синтаксической фраземы он не имеет (**ведать кого-нибудь*) — ср. примеры (27) и (28). Часто варианты этой фраземы выступают парой, как в (28). Семантическое представление этой единицы, на наш взгляд, содержит идеи усиления процесса и отстранённости субъекта высказывания от объекта глагола:

(29) *Я его знать не знал* ≈ 'Я его совсем не знал и не имел к нему отношения'.

Есть еще два-три глагола восприятия, встречающиеся в этой синтаксической фраземе: видеть, слышать, чуять, как в примерах (30)–(33):

(30) …*То ли всерьез запал на эту Ирину, то ли уязвляло его, что есть кто-то, кого он* ***знать не знает и видеть не видел****, не просто мешающий его отношениям с приглянувшейся девушкой, но как бы опускающий его.* (Е. Шкловский).

(31) *Но вот не идет у меня из головы этот шофер, которого я* ***знать не знал, видеть не видел*** (Г. Бакланов).

(32) *Никто их не знает,* ***видеть не видел****, и земля какая-то мифическая!* (Д. Липскеров).

(33) *Роясь на помойках в поисках чего-нибудь съестного, обычный бродячий пес и* ***чуять не чуял****, что судьба приведет его в теплый дом профессора медицины, где его назовут Шариком* (из рецензии на фильм «Собачье сердце, http://www.actorskino.ru/domesticfilms/1075-sobache-serdce.html).

Эти последние конструкции сродни разговорным выражениям с квазиредупликацией типа слыхом не слыхивал, нюхом не чуял и, может быть, в глаза не видел; ср. также *А вот, как выглядит чудище Поганое, Добрыня и* ***ведом не ведал, и нюхом не чуял***.

Совершенно очевидно, что разграничение всех этих конструкций — исключительно сложная задача, особенно если ее приходится решать в целях автоматической обработки текстов. Обсуждение этой проблематики выходит за рамки настоящей статьи.

Заключение

Мы рассмотрели одну любопытную нестандартную микросинтаксическую конструкцию, которая интерферирует с двумя другими микросинтаксическими единицами — другой нестандартной конструкцией и синтаксической фраземой.

Это типичная ситуация при работе с малым синтаксисом, и разумно предложить, что интерференция — отличительная черта этого фрагмента языка.[10]

References

1. *Apresjan Ju. D, Boguslavsky I.M, Iomdin L. L. Sannikov V. Z.* (2010). Theoretical Problems of Russian Syntax: Interaction of the Grammar and the Dictionary [Teoreticheskie problemy russkogo sintaksisa: Vzaimodejstvie grammatiki i slovarja]. Ed. Ju D. Apresjan. Moscow: Jazyki slavjanskikh kultur, ISBN 978-5-9551-0386-0. 408 p.
2. *Akhmanova O. S.* (1968). The dictionary of Linguistic Terms [Slovar' lingvisticheskikh terminov]. Moscow: Sovetskaja Entsiklopedija.
3. *Bulygina T. V., Shmelev A. D.* (1997). Linguistic Conceptualization of the World (on the Material of Russian Grammar) [Jazykovaja kontseptualizatsija mira (na materiale russkoj grammatiki)]. Moscow: Jazyki russkoj kultury.
4. *Fillmore Ch., Kay P. and O'Connor C.* (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of *let alone*, Language 64: 501–38.
5. *Goldberg A.* (1995) Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.
6. *Giljarova K. A.* (2010). *Takaja devochja-devochka.* The semantics of Reduplication in Russian Colloquial Speech and the Internet Language. [*Takaja devochja-devochka.* Semantika reduplikatsii v russkoj razgovornoj rechi i jazyke interneta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoj konferentsii "Dialog"], Issue 9 (16). Moscow, RGGU Publishers, pp 90–97.
7. *Goldberg A.* (2006). Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.

[10] Автор благодарит В. Ю. Апресян, прочитавшую первый вариант статьи и сделавшую ряд тонких замечаний, а также анонимных рецензентов, указавших на некоторые неточности и спорные положения. Эти замечания были учтены в окончательном тексте.

8. *Iomdin L. L.* (2003). Big problems of the minor syntax [Bol'shie problemy malogo sintaksisa], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoj konferentsii "Dialog"], Protvino, pp. 216–222.
9. *Iomdin Leonid* (2005). A Hypothesis of Two Syntactic Starts, East-West. The Second international conference on the Meaning — Text model. Eds. Ju.D. Apresjan, L. L. Iomdin. Moscow, Jazyki slavjanskoj kultury, pp. 165–175.
10. *Iomdin L. L.* (2006a). Polysemous syntactic idioms: between the vocabulary and the syntax [Mnogoznachnye sintaksicheskie frazemy: mezhdu leksikoj i sintaksisom], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoj konferentsii "Dialog"], Moscow, RGGU Publishers, pp. 202–206.
11. *Iomdin L. L.* (2006b). New observations of the syntax of Russian idioms [Novye nabljudenija nad sintaksisom russkix frazem], Obecność. Red. Bożena Chodźko, Elżbieta Feliksiak, Marek Olesiewicz. Białystok: Uniwersytet w Białymstoku. S. 247–281.
12. *Iomdin Leonid* (2007). Russian Idioms Formed with Interrogative Pronouns and their Syntactic Properties // Meaning — Text Theory 2007. Proceedings of the 3rd International Conference on Meaning — Text Theory. Wiener Slawistischer Almanach. Sonderband 69. München — Wien, 2007. S. 179–189.
13. *Iomdin L. L.* (2008). In the Depths of Microsyntax: a Lexical Class of Syntactic Idioms [V glubinakh mikrosintaksisa: odin leksicheskij klass sintaksicheskikh frazem], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoj konferentsii "Dialog"], Issue 7 (14). ISBN 978-5-7281-1022-4. Moscow, RGGU Publishers, pp. 178–184.
14. *Krjuchkova O. Ju.* (2004). Issues of Linguistic Interpretation of Lexical Reduplication in Russian [Voprosy lingvisticheskoj traktovki leksicheskoj reduplikatsii v russkom jazyke], Russkij jazyk v nauchnom osvehshchenii, No. 2 (8), pp. 63–85.
15. *McCoy, Svetlana.* Semantic and Discourse Properties of Colloquial Russian Construction of the form "X-to X, a...", Glossos, issue 3, pp. 1–26.
16. *Mel'čuk I. A.* (2012). The Language: from Sense to Text [Jazyk: ot smysla k tekstu]. Moscow, Jazyki slavjanskikh kultur.
17. *Paillard, Denis, Plungjan, V. A.* (1993). On a typf of constructions with the repetition of verbs in Russian [Ob odnom tipe konstrukcij s povtorom glagola v russkom jazyke], Russian Linguistics, 17 (3): 263–277.
18. *Plungjan V. A. and Rakhilina E. V.* (2010). Tushat-tushat — ne potushat: the Grammar of a Verbal Construction [Tushat-tushat — ne potushat: grammatika odnoj glagolnoj konstruktsii], Rakhilina E. V., ed. (2010). Construction Linguistics [Lingvistika konstruktsij], Rakhilina E. V., ed. Moscow, Azbukovnik, pp. 83–94.
19. *Rakhilina E. V., ed.* (2010). Construction Linguistics [Lingvistika konstruktsij]. Moscow, Azbukovnik. 584 p.

20. *Sannikov V. Z.* (1989). Russian Coordinative Constructions (Semantics. Pragmatics. Syntax) [Russkie sochinitel'nye konstruktsii (Semantika. Pragmatika. Sintaksis.)] Moscow, Nauka Publishers.
21. *Vilinbakhova E. L.* (2011). On the Construction of the Kind *muzh takoj muzh* in Russian (on the Material of Interner Sources) [O konstruktsii vida *muzh takoj muzh* v russkom jazyke (na materiale internet-istochnikov), Proceedings of the conference "The Russian language: constuctional and lexico-semantic approaches. [Materialy konferentsii "Russkij jazyk: konstruktsionnye I leksiko-semanticheskie podxody]. Saint Petersburg, http://iling.spb.ru/confs/rusconstr2011/pdf/Vilinbaxova.pdf.

МАГ ВЕЛ МОТ: ИЗМЕНЕНИЯ В ЯЗЫКЕ НА МАТЕРИАЛЕ БЫТОВОЙ ТЕРМИНОЛОГИИ[1]

Иомдин Б. Л. (iomdin@ruslang.ru),
Лопухина А. А. (nastya-merk@yandex.ru)

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

Панина М. Ф. (mar-fed@yandex.ru),
Носырев Г. В. (grigorij-nosyrev@yandex.ru)

Яндекс, Москва, Россия

Вилл М. В. (vill.margarita@yandex.ru),
Зайдельман Л. Я. (gde.vyhod@gmail.com),
Матиссен-Рожкова В. И. (heinin@mail.ru),
Винокуров Ф. Г. (fedor-win@ya.ru)

Российский государственный гуманитарный университет, Москва, Россия

Выборнова А. Н. (anna@179.ru),

НИУ Высшая школа экономики, Москва, Россия

Доклад продолжает исследования названий бытовых предметов в русском языке [Иомдин 2009, 2011, Iomdin et. al. 2011, Иомдин и др. 2012], проводимые на научном семинаре в Институте русского языка им. В. В. Виноградова РАН. В 2013 году предполагается подготовить к изданию проспект иллюстрированного словаря-тезауруса бытовой терминологии (СБТ). При работе над словарем обнаруживается новый языковой материал, тенденции и явления, не описанные в существующих словарях.

Ключевые слова: семантика, лексикография, словообразование, диминутивы, неологизмы, предметная лексика, частота, статистические методы

---

[1] Работа выполнена при частичной финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Язык и литература в контексте культурной динамики», гранта РГНФ №13-04-00307а и гранта НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами РФ. Авторы хотели бы также выразить признательность А. В. Темченко и И. В. Шалыминову за активное участие в подготовке материалов к докладу и рецензенту за внимательное прочтение первого варианта доклада и ценные замечания.

MAG VEL MOT: LANGUAGE INNOVATIONS IN EVERYDAY LIFE TERMINOLOGY

Iomdin B. L. (iomdin@ruslang.ru),
Lopukhina A. A. (nastya-merk@yandex.ru),
V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences, Moscow, Russia

Panina M. F. (mar-fed@yandex.ru),
Nosyrev G. V. (grigorij-nosyrev@yandex.ru),
Yandex, Moscow, Russia

Matissen-Rozhkova V. I. (heinin@mail.ru),
Vill M. V. (vill.margarita@yandex.ru),
Zajdel'man L. Ja. (gde.vyhod@gmail.com),
Vinokurov F. G. (fedor-win@ya.ru)
Russian State University for the Humanities, Moscow, Russia

Vybornova A. N. (anna@179.ru),
NRU Higher School of Economics, Moscow, Russia

The paper continues research into words denoting everyday life objects in the Russian language. This research is conducted for developing a new encyclopedic thesaurus of Russian everyday life terminology. Working on this project brings up linguistic material which leads to discovering new trends and phenomena not covered by the existing dictionaries. We discuss derivation models which gain polularity: clipped forms (*komp* < *komp'juter* 'computer', *nout* < *noutbuk* 'notebook computer', *vel* < *velosiped* 'bicycle', *mot* < *motocikl* 'motorbike'), competing masculine and feminine contracted nouns derived from adjectival noun phrases (*mobil'nik* (m.) / *mobilka* (f.) < *mobil'nyj telefon* (m.) 'mobile phone', *zarjadnik* (m.) / *zarjadka* (f.) < zarjad*noe ustrojstvo* (n.) 'AC charger'), hybrid compounds (*plat'e-sviter* 'sweater dress', *jubka-brjuki* 'skirt pants', *shapkosharf* 'scarf hat', *vilkolozhka* 'spork, foon'). These words vary in spelling and syntactic behaviour. We describe a newly formed series of words denoted multifunctional objects: *mfushkaZ*< *MFU* < *mnogofunkcional'noe ustrojstvo* 'MFD, multifunction device', *mul'titul* 'multitool', *centr* 'unit, set'. Explaining the need to compose frequency lists of word meanings rather than just words, we offer a technique for gathering such lists and provide a sample produced from our own data. We also analyze existing dictionaries and perform various experiments to study the changes in word meanings and their comparative importance for speakers. We believe that, apart from the practical usage for our lexicographic project, our results might prove interesting for research in the evolution of the Russian lexical system.

Key words: semantics, lexicography, derivation, diminutives, neologisms, everyday life vocabulary, frequency, statistical techniques

1. *Словообразование*: новые модели и новые слова

1.1. Диминутивы

Дыр бул щыл
(А. Крученых)

Как известно, названия бытовых предметов часто употребляются в форме диминутивов (во всяком случае в некоторых речевых регистрах, ср. [Земская 1981]), что свидетельствует об их освоенности. В работе [Иомдин и др. 2012] мы писали о лексикализации некоторых диминутивов (*театральная сумочка <*сумка>, шапочка <*шапка> для душа, ремешок <?ремень> для часов, половая тряпка <*тряпочка> vs. тряпочка <*тряпка> для очков* и др.) и необходимости их фиксации в качестве самостоятельных словарных единиц.

При привлечении нового материала обнаруживается активное распространение «новых диминутивов», образованных по модели «усечение и суффиксация». Интересно, что чаще всего усеченное слово с суффиксом возникает раньше[2] и соответственно более частотно, а усеченное слово без суффикса может и вовсе отсутствовать. Ср. велосипед (1887) — *велик* (1956) — *вел* (2005), *фотоаппарат* (1926) — *фотик* (2002) — *фот* (2008), *телевизор* (1933) — *телик* (1956) / *телек* (1976), *монитор* (1974) — *моник* (2004). Сравнительную частотность можно проследить по блогам (здесь приведены данные сети микроблогов Твиттер за первую неделю апреля 2013 г.): *на велосипеде* 1366, *на велике* 1094, *на веле* 90; *мой фотоаппарат* 82, *мой фотик* 78, *мой фот* 5; *монитор* 3904, *моник* 101; ср.:[3] *Продам* ***комп*** *(системник[4]+****моник****); Выкручивай газ на полную, а дальше* ***мотик*** *сам разберется, что делать, и разгонится до 60–70 км/ч; Как уговорить родаков купить* ***мот****? У меня есть в доме 3 неприкосновенные вещи:* ***моб****,* ***ноут*** *и* ***фот****.*

Обратных случаев, когда усеченное слово без суффикса возникает раньше и более частотно, чем слово с суффиксом, мало; ср. характерный пример *магнитофон* (1946), *маг* (1963) и окказиональное *магник* (1986); в современных текстах оба усеченных слова практически не встречаются[5].

---

[2] Время первой фиксации формы в НКРЯ (верхняя оценка, означающая, что слово появилось не позднее указанной даты, ср. помету «or earlier» в OED) приведено в скобках. В случаях новых и низкочастотных слов, отсутствующих в НКРЯ, приводятся примеры из библиотеки Google Books, других электронных библиотек и блогов (специфика которых, однако, такова, что точная и достоверная датировка примеров не всегда возможна). Очевидно, что нижнюю оценку установить значительно сложнее, если вообще возможно.

[3] Здесь и далее примеры с форумов, из блогов и т. п. из соображений места обычно приводим без указания конкретного источника.

[4] О словах типа *системник* см. также ниже.

[5] Не исключено, что на появление формы *маг* повлияло распространение моделей техники с соответствующими наименованиями: *Магнитофон «МАГ-Д1» разработан во «ВНАИЗ» и производился с начала 1957 года* (Сайт «Отечественная радиотехника двадцатого века).

1.2. Универбы

Еще один активный и хорошо описанный словообразовательный процесс — сокращение именных групп до одного существительного с добавлением суффикса, или компрессивная суффиксация [Земская 2007; Юй 2012]. В нашем материале интерес представляют случаи конкуренции дериватов разного рода: *тональный крем* (1987) — *тональник* / *тоналка* (2004), *зарядное устройство* (1974) — *зарядник* (2002) / *зарядка* (2009), *мобильный телефон* (1997) — *мобильник* (1998) / *мобилка* (2001). При этом интересно, что такая конкуренция, по-видимому, возможна только в случае, если мотивирующая ИГ не относится к женскому роду; см. также [Зализняк 2012].

1.3. Композиты

Профессор, снимите очки-велосипед!
(В. Маяковский)[6]

В СБТ описываемая лексика объединяется в группы слов с близким значением, в которых выделяется слово-доминанта, подобно синонимическим рядам. Некоторые возникающие при этом трудности были описаны в [Иомдин и др. 2012]. Еще одну сложную проблему представляют собой «слова-гибриды» — композиты, называющие «предмет, совмещающий в себе признаки предметов, явлений, названных мотивирующими словами» [Грамматика-80]. При работе над нашим материалом обнаружилось большое количество неологизмов такого рода, которые можно разделить на несколько групп.

1. Композит образован от названий двух разных предметов, функции и внешние особенности которых он совмещает. Наряду с такими устоявшимися в языке сочетаниями, как, например, *диван-кровать* (1966), *кресло-кровать* (1959), *плащ-палатка* (1941), в современных текстах распространены и другие, более новые композиты: *сумка-холодильник* (1991), *бюстгальтер-комбинация* (1983), *платье-свитер* (1983), *юбка-брюки* (1994), *юбка-шорты* (1989), *ложка-вилка* (2003), *стол-тумба* (1979) и другие.

2. Первый компонент композита называет гипероним (чаще — доминанту группы), второе — гипоним. Сюда относятся такие сочетания, как *сумка-авоська* (1955), *кастрюля-скороварка* (1957), *куртка-ветровка* (1986), *кресло-пуф* (1995), *карта-пропуск* (1998), *кепка-козырек* (в НКРЯ — 2010 г.; в блогах: *Можно получить один и шести призов — кепки-козырьки с ушами* (2002)), *очки-авиаторы* (в НКРЯ нет; в блогах: *Очки-«авиаторы» с желтоватыми или голубоватыми стеклами* (2004)) и др.

---

[6] В http://xaxam.livejournal.com/135535.htmlотмечена неудачность английского перевода, видимо, свидетельствующая о непонимании переводчиком оригинала: Professor, take off your bicycle glasses (Vladimir Mayakovsky, At the Top of My Voice, translated by Max Hayward and George Reavey).

В монографии [Никитина 1993] говорится: «Как показали, с одной стороны, исследования обычных бытовых текстов [Розина 1986], а с другой — психолингвистические тесты на свободную классификацию [Фрумкина, Михеев и др. 1991], гипо-гиперонимические отношения не характерны для бытовых текстов и, как правило, не встречаются в бытовых классификациях. Еще менее характерны они для текстов фольклорных». В [Fleckenstein 2001] также отмечается, что композиты, построенные по принципу «гипероним-гипоним», используются главным образом не в разговорной речи, а в «классификационных системах разного рода»: в языке торговли, науки и в официально-деловом языке. Тем не менее в области названий бытовых предметов такие номинации, как мы видим, достаточно распространены и в разговорной речи. Интересно, что порядок частей здесь отличается от принятого в фольклорных текстах и просторечии (ср. *плакун-трава*, *Ильмень-озеро*, *январь-месяц* и т. п.)

3. Компоненты композита — когипонимы, а сам композит играет роль гиперонима при отсутствии хорошего однословного эквивалента: *вилки-ложки*, *чулки-носки*, *подарки-поздравлялки* и т. п., ср. [Iomdin et. al. 2011].

4. Первый компонент композита называет предмет, второе — объект из другой области, внешне или по каким-то иным признакам напоминающий этот предмет, например: *платье-футляр*, *брюки-бананы*, *туфли-лодочки*, *шапка-носок*, *пакет-майка*, *часы-кулон*, *стол-книжка*, *кресло-груша* и другие.

Поскольку большинство таких номинаций еще не устоялись, встречаются разные варианты их орфографического оформления: через дефис (с разным порядком компонентов: *платье-рубашка* и *рубашка-платье*, *сумка-холодильник* и *холодильник-сумка* и другие), слитно (с интерфиксом: *креслокровать, креслоколяска, диванокровать, митенковарежки* и *варежкомитенки, сумкотележка, свитероплатье* и *платьесвитер, рубашкоплатье, шарфошапка* и *шапкошарф, сумкохолодильник, юбкобрюки, трусомайка* и *майкотрусы*), через слэш (*диван/кровать, кресло/кровать, плащ/накидка, ручка/указка, стол/тумба; телефон/автомат*), бэкслэш (*диван\кровать*) или знак подчеркивания (*юбка_брюки, диван_кровать, кресло_кровать, плащ_палатка, стол_тумба; тапочки_теплуши, шапка_ушанка*). Частотны и раздельные написания; по-видимому, свою роль в их распространении играет и влияние английского языка, ср. [Левонтина 2010].

Поиск в текстах разных жанров, в том числе в логах интернет-запросов, позволяет сравнить относительную частотность таких вариантов в разных группах и выявить следующие тенденции: слэш, бэкслэш и знак подчеркивания почти всегда используются только при образовании композитов первого типа; эти же композиты могут с практически равной долей вероятности писаться раздельно или через дефис (напр., в исследованных логах интернет-запросов написание

диван кровать встречается 39 445 раз[7], а *диван-кровать* — 31 213 раз). Для композитов второго типа раздельное написание предпочтительнее написания через дефис (например, *шапка ушанка* — 59 664 раза, а *шапка-ушанка* — 12 706 раз).

Интересно также, что у композитов второй группы гипероним чаще всего предшествует гипониму (98 % случаев по данным логов запросов). Композиты первой группы, появившиеся раньше других, также обнаруживают тенденцию к закреплению порядка элементов (*диван-кровать* 95 %, *кровать-диван* 5 %; *кресло-коляска* 99 %, *коляска-кресло* 1 %; *плащ-палатка* 100 %), а у неологизмов наблюдается вариативность компонентов (*брюки-юбка* 24 %, *юбка-брюки* 76 %; *шорты-юбка* 44 %, *юбка-шорты* 56 %; *варежки-митенки* 84 %, *митенки-варежки* 16 %).

Другую проблему составляет словоизменение и согласование таких композитов. Правила требуют склонения обоих компонентов (хотя в нормативных источниках и отмечается тенденция к склонению только второго компонента) и согласования по первому компоненту, см., например [Розенталь 1997]. В нашем материале и здесь обнаруживается отсутствие единообразного оформления.

Наблюдается тенденция к утрате склонения одним из элементов композита (у композитов первой группы), ср. примеры из блогов: *Полностью меблированная студия с двуспальной* ***диван-кроватью****; был дождь, а она шла без зонта с каким-то пакетом и я предложил ей укрыться* ***плащ-накидкой****; В такую погоду, да ещё и с* ***сумкой холодильник****, можете брать всё что угодно, хоть йогурты, за день ничего не случится.* Ср. характерное обсуждение «неправильного» склонения композитов в блоге: *В тексте в безумном количестве встречается понятие «кресло-коляска». Авторы упорно не склоняют первую часть, а так и пишут:* ***«кресло-коляской»****,* ***«кресло-коляски»*** *и т. д.!* Ср. также [Федорова 1998].

Данные блогов, форумов, логов запросов свидетельствуют о согласовательной вариативности композитов первого типа: *инвалидное кресло-коляска // инвалидная кресло коляска; женский плащ накидка // плащ-накидка офицерская; Приглядела себе* ***серое свитер-платье*** *из кашемира с рисунком в стиле H&M // Многие женщины не знают, с чем сочетать* ***свободный свитер-платье****; Сколько стоит* ***инвалидное кресло-коляска*** *в Туле? // Нужна* ***инвалидная кресло-коляска****;* ***Военный плащ-палатка*** *//* ***Плащ-палатка шилась*** *из водоотталкивающего габардина; Одень эту* ***классическую платье-рубашку*** *цвета*

---

[7] Для подсчета этой статистики были использованы запросы к поисковой системе Яндекс. Случайным образом было выбрано 200 000 000 запросов из поисковых логов за второе полугодие 2012 года. Логи запросов предоставляют лингвистам новый интересный материал. Пользователей поисковых систем существенно больше, чем авторов каких бы то ни было текстов, и частота исследуемых слов в запросах значительно выше, чем, например, в блогах. Пользователи обычно вводят запросы быстро и не задумываясь, что сближает их со спонтанной речью, и не рассчитывают на адресата-человека. Синтаксис запросов, безусловно, отличается от синтаксиса естественного языка, однако для лексических исследований это не кажется серьезным препятствием. Исследованные логи запросов являются собственностью компании Яндекс и поэтому недоступны извне. Те же тенденции можно проследить по открытому источнику wordstat.yandex.ru, но он содержит меньшее количество данных и его автоматическая обработка невозможна.

папайи // Белоснежное ***хлопковое платье-рубашка****; А я вот тоже недавно взяла себе* ***короткое платье-свитер, фиолетовый такой, теплый, вязаный*** (в последнем примере в препозиции к композиту прилагательное согласуется с первым словом, а в постпозиции — со вторым словом).

По всей видимости, здесь необходима работа по тщательному сбору многочисленного нового материала и принятию адекватных нормализаторских решений.

1.4. Словосочетания и новые слова

Стремительное развитие технического прогресса приводит к необходимости называть все новые сложные устройства, совмещающие в себе функции не двух, а нескольких разных. Примеры композитов с тремя компонентами носят скорее маргинальный характер, хотя встречаются и они, ср. *принтер-сканер-копир*. Здесь более распространены другие механизмы образования новых слов. Приведем несколько примеров.

Словосочетание *многофункциональное устройство* фиксируется в НКРЯ с 2004 года[8]; тогда же возникает и аббревиатура МФУ: *Персональный офисный центр для коммуникаций и делопроизводства. Это* ***МФУ*** *объединяет в себе телефон, факс, принтер, сканер и копир* («Computerworld», 2004)*; Мы уже много писали о том, что современные инженеры идут дорогой Мичурина и скрещивают технику в самых причудливых сочетаниях. Не избежали этой участи и фотопринтеры. Появились* ***многофункциональные устройства (МФУ)*** *— это принтер, сканер, копир и персональная фотолаборатория в одном флаконе* («Комсомольская правда», 2005). Вскоре от аббревиатуры образуется диминутив, который стремительно завоевывает пространство интернет-форумов, а затем и страницы прессы; это слово проникло уже и в книги, ср. *На столе расположились два монитора с логотипами «Витязь», клавиатура с мышью и* ***МФУшка*** (А. Малов, Исповедь кардера, 2010). Орфография его еще не устоялась, ср. *Моя* ***мфушка*** *работает в разы тише* (с форумов); *Epson* ***МФэушка*** *!!! Не печатает!; Встречается и редкая категория офисных черно-белых струйных эмфэушек, для которых характерно обязательное наличие факса*.

Еще один схожий пример — слово *мультитул*. В НКРЯ оно не встретилось, первые примеры относятся к 2004 году; ср. *В свое время (около года назад), я купил себе* ***мультитул****. <…> Пользовал его на сто процентов — начиная от разборки компьютеров и открывания окон в поезде до нарезания закуски и открывания бутылок.*

Сложность интерпретации и описания таких языковых единиц — в непрозрачности их внутренней формы: *мультифункциональное устройство* или

---

[8] Ср. также пример из рекламы 1997 г., где это явно еще не устойчивое, а свободное словосочетание: *Подлинно многофункциональное устройство, совмещающее возможности факса, модема, принтера, сканера, телефона и копировального аппарата* («Коммерсант-Власть», 16.12.1997).

мультитул теоретически могли бы означать совсем иные предметы. Важно определить момент, когда их значение уже устоялось и подлежит лексикографической фиксации. Со словом *мультитул*, по-видимому, это уже произошло, хотя иногда встречаются и употребления с иным значением, ср. ***Мультитул*** *Bar10der позволит приготовить любимый коктейль в любом месте и в любое время <…>. Он включает в себя приспособления для измельчения фруктов и специй, выжимания сока, процеживания, очистки цитрусовых и карвинга, два мерных стаканчика (на 1 и 0,5 унции), 4-дюймовый нож, штопор и открывалку для бутылок* («Популярная механика», 22.11.2012). А вот у слова мультиинструмент явно не сложилось (пока?) конкретное значение. Ср.: *Брелок-**мультиинструмент** выполнен из цельного куска титана, может использоваться, как средство самообороны и отвертка, также как открывалка, шестигранный ключ*; *73-летний инженер-самоучка представил свой новый проект* трактора-***мультиинструмента**, который умеет делать больше десяти различных операций: пахать, косить, сеять, чистить, поднимать грузы, поливать*; *F.Дневник — это только первый сервис в **мультиинструменте** для рыбалки от F.Гид; Квителашвили делает из гитары **мультиинструмент**, на котором исполняет и джаз, и классический рок, и фольклорные мелодии.*

В последнее время значение ‘сложный многофункциональный предмет’ все больше получает слово *центр*. Ни в одном из известных нам словарей это значение не представлено, и даже в словарной статье И. В. Галактионовой ЦЕНТР [Апресян и др. 2010] есть только *музыкальный центр* в зоне фразеологии. Словосочетание *музыкальный центр*, по-видимому, действительно представляет собой первый пример употребления слова *центр* в этом значении[9]. Однако сейчас оно встречается в самых разных словосочетаниях (пока в основном в рекламных текстах): ср. игрушку ***развивающий центр**, **игровой центр** для ванны, **центр для творчества** на колесах, игровой набор **«детский научный центр»**; **силовой центр** Torneo POWER; **многофункциональный атлетический центр**; **кухонный центр** «Петит Гурме»; **варочный центр**; **зубной центр** Sonicare FlexCare; утюг с парогенератором — универсальный **паровой центр**; многофункциональный косметический **SPA-центр**; **Интернет-центр** не сложнее домашних беспроводных роутеров;* наконец, ***универсальный центр** Festool UCR 1000*. Интересно, что это, по-видимому, не заимствование; во всяком случае, у английского слова *center* такого значения нет.

---

[9] По-видимому, это словосочетание вначале возникло как техническое наименование, ср. названия моделей: *Мелодия 101С Сетевая радиола* (1973), *Мелодия 103В Сетевой электрофон* (1975), *Мелодия 106С Сетевой музыкальный центр* (1978), однако в 1990-е годы все еще воспринималось как необычное: *В далеком 1990-ом году из магазина «Березка» на «Химфармзавод», где работали мои родители, привезли музыкальный центр «СОНИ» <…> В те времена словосочетание «музыкальный центр» было для меня непонятным и загадочным. Я представлял себе кучу аппаратуры, типа как в звукозаписывающей студии. А фигле — не магнитофон какой-нить, а целый ЦЕНТР!* (dr-batman.livejournal.com). Примеры в художественных текстах и публицистике начинают массово появляться лишь в конце 1990-х годов.

Ср. еще примеры слов с подобным значением: ***Триблоком*** *называется устройство, совмещающее в себе несколько операций. Так, обычно триблоки совмещают в себе операции мойки, розлива и укупоривания.* ***Моноблок*** *отличается от триблока, обычно, отсутствием операции мойки; Вы еще не знаете, что такое бешено набирающий популярность домашний агрегат под названием* ***мультиварка****?; Что такое* ***мультивизор****? Это телевизор со встроенной системой домашнего кинотеатра в одном «флаконе»;* ***Комбидресс*** *соединяет трусики и рубашку.*

Найти в словарях слова с подобным значением трудно или невозможно, хотя у многих людей возникает такая необходимость, в основном в целях определить нужное наименование для поиска товара; ср. примеры с форумов: *Конечно, можно было бы приобрести роскошный гарнитур для ванной комнаты, не знаю, как на самом деле называется — монолит, совмещающий в себе зеркало с подсветкой, раковину, шкафчики; Как называется девайс, совмещающий в себе расческу и машинку для стрижки волос?; Девочки, как называется прибор, который совмещает много приборов? Т.е. и мультиварка и хлебопечка и аэрогриль; Как называется девайс который совмещает в себе модем и маршрутеризатор Wi-Fi ? Хочу поставить себе модем, который ко всему прочему раздаёт Wi-Fi, но понятия не имею как его «обозвать». Подскажите плиз.*

2. Семантика: значения и толкования

> Безмен это вроде весов. На палке шар и крючок.
> Я бы нарисовать мог
> но мало места. Могу описать интересующий Вас предмет словами.
> (Д. Хармс)

В последнее время большое значение, в частности, в компьютерной лингвистике, придается созданию частотных словарей и списков слов. К сожалению, большинство имеющихся частотных списков составляются из вокабул, но не отдельных лексем (то есть слов, взятых в определенном значении). Между тем очевидно, что разные лексемы одной и той же вокабулы частотны в очень разной степени; столь же очевидно, что составление частотных списков лексем представляет собой существенно более трудную и практически не автоматизируемую задачу. Тем не менее представляется важным действовать в этом направлении. При работе над СБТ информация о частотности лексем необходима и для составления словника, и для отнесения лексем к основной или периферийной зоне, и для выявления интересных случаев эволюции значений, когда обозначения одного и того же денотата сменяют друг друга (как это произошло, например, со словами *гребёнка* и *расчёска*), и описания их в соответствующей зоне словарной статьи.

Один из методов определения частотности лексем в логах пользовательских запросов был описан в [Иомдин и др. 2012]. В настоящее время наш список наиболее частотных лексем, описывающих предметы быта, составляется следующим образом: для каждого слова из полного словника СБТ (на текущий

момент составляющего более 2000 слов) подсчитывается, входит ли оно в верхнюю сотню при ранжировании слов по частоте вхождений в различные массивы текстов на современном русском языке (за последние 40 лет), сбалансированные по жанровой принадлежности (среди них: разные подкорпусы НКРЯ, библиотека Мошкова, база данных юридических текстов Консультант Плюс, блоги, логи запросов пользователей, собственные данные анкетирования информантов и др.): вверху оказываются слова, удовлетворяющие этому критерию по всем или большинству выбранных массивов текстов. Затем из них отбираются только те слова, у которых основная доля употреблений приходится именно на интересующие нас, предметные значения. Приведем текущий вариант верхней части списка (словарные статьи этих слов лежат в основе готовящегося к изданию проспекта СБТ): *белье, брюки, бумага, бутылка, ведро, газета, джинсы, диван, документ, зеркало, карта, карточка, ключ, книга, книжка, компьютер, конверт, коробка, кошелек, крем, кресло, кровать, куртка, лекарство, мешок, нож, ноутбук, обувь, одежда, пакет, пальто, паспорт, платок, платье, плеер, ручка, сапоги, стакан, стол, стул, сумка, телевизор, телефон, туфли, учебник, холодильник, чемодан, штаны, экран, ящик*. Как представляется, такие списки могут быть полезны и для составления разного рода учебных материалов.

При этом обнаруживаются примечательные случаи изменения частотности у разных лексем одной и той же вокабулы. Среди исследований, которые мы проводим в этой области, упомянем следующие.

1. В декабре 2012 года студенты Школы анализа данных Яндекса участвовали в учебном проекте по определению времени появления в русском языке лексем из словника СБТ. Разделение вокабул на лексемы и поиск их первых употреблений они проводили, руководствуясь собственной интуицией. При этом интересно отметить ситуации, в которых участники проекта не приводили данные о некоторых значениях (по всей вероятности, ушедших из современного языка) и, наоборот, находили значения, не отраженные ни в одном словаре. Так, например, для слова *уголок* один из участников проекта (после указания, что обнаружены не все значения, и повторного поиска) привел значение ‘сигарета, самокрутка с наркотиком’ (2000), но так и не отметил таких значений, как ‘линейка в виде треугольника’ и ‘угловой предмет мебели для кухни’; у слова *маркер* было выделено значение ‘оружие в пейнтболе’.

2. В работе [Иомдин 2012] упоминались некоторые эксперименты с информантами, проводимые нашей группой для изучения наивных представлений о значениях слов и стратегий описания, основанные на изучении поведения участников языковых игр. Среди любопытных наблюдений, сделанных в ходе этих экспериментов, можно привести следующие: *канапе* (участники объясняли это слово как «маленькие бутербродики», «еда такая, кусочки на палочке», тогда как в словарях первое значение — ‘небольшой диван с приподнятым изголовьем’), *гель* (в эксперименте это ‘то, что волосы фиксирует’ или ‘то, чем моются’, а в СОШ — ‘студенистое вещество, обладающее нек-рыми свойствами твёрдых тел’), *карточка* (участник старшего поколения пытался объяснить это слово школьникам, имея в виду значение ‘фотография’, но не был понят) и др.

3. Еще один способ изучения семантики и развития полисемии — анализ словарей. Он позволяет выявить время фиксации в языке новых слов и новых значений, а также изучить трансформацию значений. Ср. эволюцию слова *визитка*, для которого некоторые словари дают в качестве единственного (СУш, МАС) или первого значения (БТС, СОШ, СЕф) «однобортный сюртук с закругленными расходящимися спереди полами, фалдами», в качестве второго «мужская ручная сумочка» (СОШ, СЕф; в ТСИ — это значение третье; в СУш и МАС вообще отсутствует) и лишь в качестве третьего значение «визитная карточка», которое в сознании современного носителя является основным (БТС, СОШ (*разг.*), СЕф; в СУш и МАС отсутствует, в ТСИ — первое, но с пометой *разг.*). Подробнее см. [Выборнова 2013 — готовится к публикации].

Авторы стремятся по возможности учитывать результаты своих исследований при практической работе над Словарем бытовой терминологии, однако, как кажется, они могут представлять и самостоятельный интерес для изучения эволюции лексической системы русского языка.

Литература

1. *Апресян* и др. 2010 — 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., Птенцова А. В., Санников А. В., Урысон Е. В. Проспект активного словаря русского языка. Отв. ред. акад. Ю. Д. Апресян. М.: «Языки славянских культур», 2010.
2. *БТС* — Большой толковый словарь русского языка / Сост., гл. ред. С. А. Кузнецов. СПб.: Норинт, 1998.
3. *Грамматика-80* — Русская грамматика Т. I. Под ред. Н. Ю. Шведовой. М.: АН СССР, 1980.
4. *Зализняк* 2012 — Зализняк А. А. Механизмы экспрессивности в языке // Смыслы, тексты и другие захватывающие сюжеты. Сборник статей в честь 80-летия И. А. Мельчука. М.: «Языки славянских культур», 2012. С. 650–664.
5. *Земская* 1981 — Земская Е. А. Русская разговорная речь. Под ред. М. В. Китайгородской, Е. Н. Ширяева. М.: Наука, 1981.
6. *Земская* 2007 — Земская Е. А. Словообразование как деятельность. Изд. 3-е. М.: КомКнига, 2007.
7. *Иомдин* 2009 — Иомдин Б. Л. Терминология быта. Поиски нормы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. *8 (15)*. М.: РГГУ, 2009. С. 127–135.
8. *Иомдин* 2011 — Иомдин Б. Л. Материалы к словарю-тезаурусу бытовой терминологии. СВИТЕР: образец словарной статьи // Слово и язык. Сборник статей к восьмидесятилетию академика Ю. Д. Апресяна. Отв. ред. И. М. Богуславский, Л. Л. Иомдин, Л. П. Крысин. М.: «Языки славянских культур», 2011. С. 392–406.

9. *Иомдин* 2012 — Иомдин Б. Л. Наивные представления о значениях слов // «Народная лингвистика»: взгляд носителей языка на язык. Тезисы докладов международной научной конференции, Санкт-Петербург, 19–21 ноября 2012 г. Отв. ред. Е. В. Головко. СПб.: Нестор-История, 2012. С. 22–24.
10. *Иомдин* и др. 2012 — Иомдин Б. Л., Лопухина А. А., Пиперски А. Ч., Киселева М. Ф., Носырев Г. В., Рикитянский А. М., Васильев П. К., Кадыкова А. Г., Матиссен-Рожкова В. И. Словарь бытовой терминологии: новые проблемы и новые методы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2012» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). М.: РГГУ, 2012. С. 213–226.
11. *Левонтина* 2010 — Левонтина И. Б. Русский со словарем. М.: Азбуковник, 2010.
12. *МАС* — Словарь русского языка: В 4-х т. / АН СССР, Ин-т рус. яз.; Под ред. А. П. Евгеньевой. М.: Русский язык, 1985–1988.
13. *Никитина* 1993 — Никитина С. Е. Устная народная культура и языковое сознание. М., Наука, 1993. 187 с.
14. *Розенталь* 1997 — Розенталь Д. Э. Справочник по правописанию и стилистике. СПб., Комплект, 1997.
15. *СЕф* — Ефремова Т. Ф. Большой современный толковый словарь русского языка. В 3-х т. М., АСТ, Астрель, 2006.
16. *СОШ* — Ожегов С. И. и Н. Ю. Шведова. Толковый словарь русского языка. М.: Азъ, 1992.
17. *СУш* — Толковый словарь русского языка / Под ред. Д. Н. Ушакова. М.: Гос. ин-т Сов. энцикл.; ОГИЗ; Гос. изд-во иностр. и нац. словарей, 1934–1940.
18. *ТСИ* — Крысин Л. П. Толковый словарь иноязычных слов. М.: Русский язык, 1998.
19. *Федорова* 1998 — Федорова О. В. Мой диван-кровать или моя диван-кровать? (Дефисно-аппозитивные словокомплексы в русском языке) // Труды международного семинара «Диалог 98» по компьютерной лингвистике и ее приложениям. Казань: Хэтер, 1998. С. 610–618.
20. *Фрумкина* и др. 1991 — Фрумкина Р. М., Михеев А. В., Мостовая А. Д., Рюмина Н. А. Семантика и категоризация. М.: Наука, 1991. 168 с.
21. *Юй* 2012 — Янь Юй. Компрессивно-стилистическая деривация в русском языке начала XXI века // Славянские языки и культуры в современном мире: II Международный научный симпозиум. М.: МГУ, 2012. С. 21.
22. *Fleckenstein* 2001 — Fleckenstein C. Сложно-составные слова в русском языке. Специфика структурного типа и выразительные возможности наименований // Slavische Wortbildung : Semantik und Kombinatorik : Materialien der 5. Internationalen Konferenz der Kommission für Slavische Wortbildung beim Internationalen Slavistenkomitee, Lutherstadt Wittenberg, 20–25 September 2001. Wittenberg, 2001. C. 201–211.
23. *Iomdin* et al. 2011 — Iomdin B., Piperski A., Russo M, Somin A. How different languages categorize everyday items // Computational linguistics and intellectual technologies. Papers from the annual international conference “Dialogue” (2011). Moscow: RSUH, 2011. P. 258–268.

24. *OED* — J. A. Simpson; E. S. C. Weiner. The Oxford English dictionary. New York: Oxford University Press, 1989.

References

1. *Apresjan V. Ju., Apresjan Ju. D., Babaeva E. E., Boguslavskaja O.Ju., Galaktionova I. V., Glovinskaja M. Ja., Iomdin B. L., Krylova T. V., Levontina I. B., Ptentsova A. V., Sannikov A. V., Uryson E. V.* (2010), Prospekt aktivnogo slovarja russkogo jazyka [Prospect of the Active dictionary of Russian]. Jazyki slavjanskih kul'tur, Moscow.
2. *Efremova T. F.* (2006), Large Contemporary Explanatory Dictionary of Russian [Bol'shoj sovremennyj tolkovyj slovar' russkogo jazyka]. AST, Astrel', Moscow.
3. *Evgenyeva A. P.* (ed.). (1981–1984), Russian Language Dictionary [Slovar' russkogo jazyka]. Russkij jazyk, Moscow.
4. *Fedorova O. V.* (1998), Hyphenated apposition word complexes in Russian [Moj divan-krovat' ili moja divan-krovat'? (Defisno-appozitivnye slovokompleksy v russkom jazyke)] // Proceedings of the international seminar «Dialog 1998» in computational linguistics and its applications. Heter, Kazan, pp. 610–618.
5. *Fleckenstein C.* (2001), Compound words in Russian. The peculiarities of the structure type and expressive possibilities of nominations [Slozhno-sostavnye slova v russkom jazyke. Spetsifika strukturnogo tipa i vyrazitelnye vozmozhnosti naimenovanij]. Slavische Wortbildung: Semantik und Kombinatorik : Materialien der 5. Internationalen Konferenz der Kommission für Slavische Wortbildung beim Internationalen Slavistenkomitee, Lutherstadt Wittenberg, pp. 201–211.
6. *Frumkina R. M., Miheev A. V., Mostovaja A. D., Rjumina N. A.* (1991), Semantics and categorization [Semantika i kategorizatsija]. Nauka, Moscow.
7. *Iomdin B. L.* (2009), Everyday life terminology. In pursuit of standards [Terminologija byta. Poiski normy]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009". Bekasovo, pp. 127–135.
8. *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.
9. *Iomdin B., Piperski A., Russo M., Somin A.* How different languages categorize everyday items. (2011), Computational Linguistics and Intellectual Technologies: Proceedings of the InternationalConference "Dialog 2011". Bekasovo, pp. 258–268.
10. *Iomdin B. L.* (2012), Naïve ideas of word meanings [Naivnye predstavlenija o znachenijah slov]. «Narodnaja lingvistika»: vzgljad nositelej jazyka na jazyk ["Folk Linguistics": Language from Speakers' Perspective]. Nestor-Istorija, St. Petersburg, pp. 22–24.

11. *Iomdin B. L., Lopuhina A. A, Piperski A. Ch., Kiselëva M. F., Nosyrev G. V., Rikitjanskij A. M., Vasil'jev P. K., Kadykova A. G., Matissen-Rozhkova V. I.* (2012), Thesaurus of Russian everyday life terminology: new problems and new techniques [Slovar' bytovoj terminologii: novye problemy i novye metody]. Computational Linguistics and Intellectual Technologies: Proceedings of the InternationalConference "Dialog 2012". Bekasovo, pp. 213–226.
12. *Krysin L. P.* (1999), Explanatory Dictionary of Loanwords [Tolkovyj slovar' inojazychnyh slov]. Russkij jazyk, Moscow.
13. *Kuznetsov S. A.* (ed.). (1998), Large Explanatory Dictionary of Russian [Bol'shoj tolkovyj slovar' russkogo jazyka]. Norint, St. Petersburg.
14. *Levontina I. B.* (2010), Russian with a Dictionary [Russkij so slovarëm]. Azbukovnik, Moscow.
15. *Nikitina S. E.* (1993), Folk spoken culture and language consciousness [Ustnaja narodnaja kul'tura i jazykovoe soznanie]. Nauka, Moscow.
16. *Ozhegov S. I., Shvedova N. Yu.* (1992), Explanatory Dictionary of Russian [Tolkovyj slovar' russkogo jazyka]. Az, Moscow.
17. *Simpson J. A., Weiner E. S. C.* (1989), The Oxford English dictionary. Oxford University Press, New York.
18. *Rozental' D. E.* (1997), A guide in orthography and stylistics [Spravochnik po pravopisaniju i stilistike]. Komplekt, SPB.
19. *Shvedova N. Yu.* (ed.). (1980), Russian grammar [Russkaja grammatika]. Vol. 1. USSR Academy of Sciences, Moscow.
20. *Ushakov D. N.* (ed.). (1934–1940), Explanatory Dictionary of Russian [Tolkovyj slovar' russkogo jazyka]. OGIZ, Moscow.
21. *Yu Y.* (2012), Compressive stylistic derivation in the Russian language in the beginning of the XXI century [Kompressivno-stilisticheskaja derivatsija v russkom jazyke nachala XXI veka]. Slavic languages and cultures in the modern world: II International Scientific Symposium. MSU, Moscow, P. 21.
22. *Zaliznjak A. A.* (2012), The expressive tools in the language [Mehanizmy ekspressivnosti v jazyke]. Meanings, texts, and other exciting things. A Festschrift to Commemorate the 80th Anniversary of Professor Igor Alexandrovic Mel'cuk. Jazyki slavjanskih kul'tur, Moscow, pp. 650-664.
23. *Zemskaja E. A.* (1981), Russian spoken language [Russkaja razgovornaja rech']. Moscow.
24. *Zemskaja E. A.* (2007), Word-building as activity [Slovoobrazovanie kak dejatel'nost']. KomKniga, Moscow.

СЕМАНТИЧЕСКИЕ РОЛИ И СЕТЬ КОНСТРУКЦИЙ В СИСТЕМЕ FRAMEBANK[1]

Кашкин Е. В. (egorkashkin@rambler.ru)
Московский государственный университет
им. М. В. Ломоносова, Москва, Россия

Ляшевская О. Н. (olesar@gmail.com)
НИУ Высшая школа экономики, Москва, Россия

Словарь русских глагольных конструкций — часть системы FrameBank, которая постоянно пополняется по данным Национального корпуса русского языка. Семантическая разметка глагольных конструкций включает а) определение значения глагола и семантических ролей (экспликаций) участников ситуации, б) формулировку семантических ограничений на заполнение валентностей, в) установление отношений между конструкциями одного глагола и между конструкциями разных глаголов в графе конструкций.

Инвентарь семантических ролей устроен иерархически, что позволяет масштабировать его размеры от десятка проторолей до множества частных определений. Инвентарь «базового уровня», описанный в работе, коррелирует с семантической классификацией глагольной лексики в НКРЯ.

В статье также формулируются принципы построения графа конструкций, отражающего как семантические переходы на множестве глагольной лексики, так и наследование / сдвиги в морфосинтаксическом оформлении конструкций.

Обсуждаются возможности практического использования FrameBank в задаче semantic role labeling, а также теоретические вопросы соотношения семантических классов глаголов, семантических ролей и семантических ограничений на заполнение валентностей.

Ключевые слова: конструкции, семантические роли, полисемия, семантические переходы, лексикография, корпусная лингвистика

---

[1] Работа выполнена в рамках программы фундаментальных исследований Президиума РАН «Корпусная лингвистика».

SEMANTIC ROLES AND CONSTRUCTION NET IN RUSSIAN FRAMEBANK

Kashkin E. V. (egorkashkin@rambler.ru)
Lomonosov Moscow State University, Moscow, Russia

Lyashevskaya O. N. (olesar@gmail.com)
NRU Higher School of Economics, Moscow, Russia

The paper reports on a research project in progress which involves a dictionary of Russian lexical constructions and a corpus tagged with FrameNet-like annotation scheme. Russian FrameBank, originally conceived as an analogue of Berkeley FrameNet, takes into account some recent approaches adopted in Construction Grammar and Russian lexical semantics, as well as certain features of the Russian lexical system and grammar.

We focus on the semantic annotation of constructions in FrameBank. First, the article describes the inventory of semantic roles used in FrameBank which correlates with the semantic classification of verbs and other predicates. Semantic roles form a hierarchy: 88 roles are classified into six clusters (those of Agent, Patient, Experiencer, Instrument, Addressee, Circumstances), which are further subdivided into some smaller groups. The hierarchical organization makes the inventory of semantic roles more flexible for use in theoretical research and computational applications (such as automatic semantic role labeling). We also show that many examples are annotated in a more appropriate way by introducing syncretic semantic roles (e. g. Instrument-Place or Result-Manner).

Second, we touch upon an ongoing project on the systematization of semantic shifts in verbal lexemes (metaphor, metonymy, and rebranding, which is argued to be a special type of a semantic shift, see, for example, [Rakhilina et al. 2010a]) and the corresponding changes in argument structure constructions (including changes of a morpho-syntactic pattern, omission of a participant which belongs to a known class, etc.). The labels for the shifts are provided, along with examples of their realization. Lexical constructions are defined on constant (lexicalized) slots, mainly verbs and other predicates in a particular meaning. Frames are thus seen as the signifié side of constructional clusters formed by synonymous predicates, aspectual pairs, etc. Since it is not uncommon for polysemous lexemes that the formal façade of constructions is inherited from sense to sense, we claim that the frame nets cannot be routed without taking into account sense relations in polysemous predicates.

The final discussion deals with the relation between semantic classes of verbs, semantic roles, and lexical/semantic constraints on the classes of participants as provided by FrameBank data.

Key words: constructions, semantic roles, polysemy, semantic shifts, lexicography, corpus linguistics

1. Русский FrameBank

Статья посвящена описанию двух разработанных классификаций: инвентаря семантических ролей и переходов между конструкциями одного глагола, которые составляют теоретическую основу семантической разметки системы FrameBank (www.framebank.ru). Прежде чем перейти к основной части изложения, охарактеризуем кратко общее устройство и назначение системы (см. также [Ляшевская, Кузнецова 2009, Lyashevskaya 2010, Lyashevskaya 2012]).

FrameBank — общедоступный онлайн-ресурс, объединяющий словарь лексических конструкций русского языка и размеченный корпус их реализаций в текстах НКРЯ. Конструкции включают предикатно-аргументные структуры глаголов, существительных, прилагательных, наречий и предикативов, а также идиомы, в которых часть элементов фиксированы, а часть представляют собой переменные (т. н. конструкции «малого синтаксиса»).

FrameBank относится к семейству FrameNet-ориентированных ресурсов (см. https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages)[2], однако и по идеологии, и по формату несколько отличается от их родоначальника, системы FrameNet, разработанной в Беркли. Если FrameNet строится вокруг фреймов — типовых ситуаций с известным набором участников и расписанными ролями, то русский FrameBank — вокруг конструкций конкретных лексем. Каждое значение лексемы соответствует своему фрейму, а близкие значения ряда лексем, описывающие типовую ситуацию, входят в общий фрейм. Во FrameNet средства выражения участников в тексте приписываются сразу всем предикатам, обозначающим данный фрейм, — во FrameBank подразумевается, что конструкция каждого предиката имеет индивидуальные особенности (ограничения на заполнение элементов, отличия в значении), даже если предикаты относятся к одному фрейму. В основе FrameNet лежит теория Frame Semantics, разработанная Ч. Филлмором в середине 1970-х гг., и идея, что сеть фреймов универсальна во всех языках, а русский FrameBank ориентирован в большей мере на исследования лексических конструкций в духе Грамматики Конструкций того же Ч. Филлмора, А. Голдберг и др. и Московской семантической школы (Ю. Д. Апресян, Е. В. Падучева и др.).

FrameBank документирует:

- русскую лексическую систему, структуру русских лексико-семантических групп и полисемии (фреймы не универсальны);

---

[2] Помимо версий FrameNet для японского, китайского, бразильского португальского, шведского, испанского и немецкого языков, клонирующих полностью или частично его структуру и формат, к семейству FrameNet-ориентированных ресурсов можно отнести современные корпусно-ориентированные словари валентностей, соответствующие компоненты WordNet-ов, а также корпуса с глубокой разметкой, отражающей аргументную структуру, кореференцию, дискурсивные стратегии и мн. др. (ср., например, тектограмматику в Prague Dependency Treebank, тестовые корпуса для машинного перевода с Compreno-разметкой ABBYY и др.).

- парадигматические отношения между значениями многозначных слов — в том, как они отражаются в системе связанных с этими значениями лексических конструкций;
- лексико-семантические ограничения на слоты конструкций;
- грамматические особенности русского языка (порядок слов, падежи, согласование и т. п.).

Ядро системы FrameBank составляют 2200 частотных русских глаголов и ассоциированных с ними конструкций и корпусных примеров. Словарь русских глагольных конструкций представляет каждую конструкцию как шаблон, в котором указаны а) морфосинтаксические характеристики элементов конструкции; б) синтаксический ранг участника; в) экспликация (роль) участника; г) семантические ограничения на заполнение слота конструкции.

Например, для глагола *резать* выделено около 30 шаблонов конструкций (обозначаемых для наглядности ярлыками-примерами), среди которых есть кластер шаблонов *Продавщица* ***режет*** *сыр, Она* ***режет*** *хлеб на тонкие куски, Василий начал* ***резать*** *хлеб длинными ломтями, Портной* ***режет*** *ткань большими ножницами,* реализующих с помощью различных моделей управления значение глагола ‘разделять объект на части давлением острого инструмента’, кластер шаблонов *Старик* ***резал*** *четки из кипариса* и *Он* ***резал*** *деревянные ложки простым ножом*, соответствующих значению ‘изготавливать что-л. посредством резьбы’, а также кластер *В боку* ***режет*** и *У него в желудке* ***резало*** (см. рис. 1), где глагол *резать* описывает определенный тип болевого ощущения.

Схема: в + Sloc Vimpers у + Sgen
Пример: У него в желудке резало.

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
Z	в + Sloc	часть субъекта физиологического ощущения	Периферия	часть тела
-	Vimpers	-	Предикат	-
T	у + Sgen	посессор	Периферия	одушевленный

Рис. 1. Шаблон одной из конструкций глагола *резать* в значении физиологического ощущения

В корпусной части ресурса FrameBank представлено около 100 примеров из НКРЯ на каждый глагол (в дальнейшем выборки планируется расширить). Разметчик должен соотнести пример с конструкцией (если нужный шаблон в системе отсутствует, разметчик пополняет словарь конструкций). Затем

определяется вариант реализации конструкции (включая нестандартные, например, при деепричастии, пассивном причастии и т. п.), в примере определяются группы, соответствующие элементам конструкции (а также синтаксические и семантические вершины групп), несовпадения в морфосинтаксическом оформлении и ограничениях на лексико-семантическую сочетаемость. Отдельно размечаются сирконстанты и модальные слова, которых шаблон, естественно, не охватывает (см. рис. 2).

-- Вот он, один из этих штукарей из Варьете ,-- послышался грозный голос над онемевшим бухгалтером . И тут же Василия Степановича арестовали.

Имя	Вершина	Группа	Заполнение вершины	Заполнение группы	Экспликация	Ранг	Семантические ограничения
X	Snom	NPnom	неопределенно-личная конструкция		агенс	Субъект	лицо / организация
Реализация	-	-			-	Не выражен	-
-	арестовать	арестовать			-	Предикат	-
Реализация	арестовать	N/A	арестовали	N/A	-	Предикат	-
Y	Sacc	NPacc	стандартный	стандартный	пациенс	Объект	лицо
Реализация	Sacc	NPacc	Василия	Василия Степановича	пациенс	Объект	лицо

№	Тип	Группа
1	сирконстант	тут же

Рис. 2. Разметка примера на шаблон конструкции «Полиция арестовала преступника». Каждому элементу конструкции отведено две строки: в верхней отражается информация из словаря, а также вариант реализации (пассивная конструкция, неопределенно-личная и т. п.), в нижней — разметка реализации в данном примере

Два компонента системы FrameNet не нашли пока отражения во FrameBank. Граф фреймов (frame grapher[3]) видится как над надстройка над графом конструкций, о котором речь пойдет ниже, а сами фреймы — как генерализация индивидуальных значений лексических конструкций. Корпус со сплошной полнотекстовой framenet-разметкой (full-text annotation) в русском ресурсе также пока отсутствует, и эта задача представляется приоритетным направлением дальнейшего развития системы.

Создаваемый ресурс выполняет как лексикографические задачи (например, предъявляет пользователю список конструкций того или иного глагола,

[3] Во FrameNet граф строится от фреймов самого частного уровня (например, фрейм, кодируемый глаголом to shop и существительным shopping) к промежуточным уровням (например, купля-продажа в перспективе продавца (продажа) VS в перспективе покупателя (купля)) и далее к еще более абстрактным (например, фрейм, общий по отношению к разнообразным вариантам купли-продажи, фрейм посессивного отношения, транзитивный фрейм и т. п.).

выдает все глаголы с выражаемой при них ролью Инструмента), так и служит инструментом для углубленного исследования русских конструкций с использованием корректно размеченных и не содержащих поискового «шума» примеров (о теоретической актуальности последней задачи ср. в частности [Рахилина 2010]).

FrameBank может быть использован и при решении прикладных задач, как размеченный вручную стандарт для машинного обучения. Наиболее тесно с нашим ресурсом связана задача автоматического распознавания семантических ролей (semantic role labeling (SRL), см. [Gildea, Jurafsky 2000, Màrquez et al. 2008, Кузнецов (в печати)]). Эта задача складывается из а) идентификации и определения ролей семантических аргументов предиката в тексте (при том, что его фрейм известен) и б) определения фрейма (значения) предиката, отсутствующего в словаре.

Решение пользовательских и компьютерных задач чувствительно к классификации семантических ролей и самих фреймов и конструкций. Ожидания пользователей об инвентаре этих единиц и круге явлений, которые они охватывают, могут быть разными. Точно так же успех задачи SRL зависит от дробности ролей и успешного «вытягивания» похожих конструкций через сеть фреймов/конструкций. Далее в статье описывается начатый в 2012 г. проект по системной семантической разметке глагольных конструкций, которая строится на иерархическом связывании конструкций и ролей. Именно этот принцип должен обеспечить гибкость в приспособлении к разным задачам. В разделе 2 обсуждается классификация семантических ролей, а разделе 3 — принципы построения графа конструкций, отражающего внутрилексемные и межлексемные семантические связи целевых глаголов.

2. Инвентарь семантических ролей

Инвентарь семантических ролей для русского языка может иметь разный состав и объем (ср. в частности [Апресян 1974/1995: 125–126], [Апресян и др. 2010: 370–377], [Падучева 2004: 587–588], а также обзор различных подходов и теоретических проблем в [Fillmore 1968, 1977, 1982], [Dowty 1991], [Лютикова и др. 2006: 17–22], [Плунгян 2011: 160–165]), что во многом определяется конкретными нуждами его использования. Следует, однако, иметь в виду, что:

- роль — это инвариант над разнообразием морфосинтаксических способов кодирования участника; так же и семантически — это генерализация функций участника в круге ситуаций, обозначаемых группой предикатов;
- роли в описании семантически близких лексем должны либо системно совпадать, либо системно различаться;
- полный инвентарь ролей должен описывать все области лексики.

Идея применения инвентаря семантических ролей к описанию больших массивов данных сама по себе не нова, ср. в частности известные проекты FrameNet, «Лексикограф», НОСС и RussNet. В упомянутых ресурсах, однако, этот инвентарь играет лишь вспомогательную роль при описании других свойств лексем и конструкций. Так, разработчики системы RussNet ориентированы в первую очередь на создание детального тезауруса русской лексики, применимого в сфере автоматической обработки естественного языка, а не на подробную классификацию и анализ семантических ролей. В современной версии системы FrameNet семантическая роль служит лишь для пояснения конфигурации участников внутри одного фрейма, и к ее называнию не предъявляется никаких требований. В результате FrameNet предлагает слишком широкий и, как кажется, потенциально неограниченный набор семантических ролей, часто и вовсе заводимых ad hoc для одного узкого фрейма — ср., например, выделение отдельного фрейма AGRICULTURE, покрывающего лексические единицы *to cultivate, to farm* и *farming*, участникам которого приписываются такие роли, как Agriculturist (тот, кто возделывает сельскохозяйственную культуру) и Food (возделываемая сельскохозяйственная культура). Неудивительно, что получить классификацию семантических ролей в онлайн-версии системы невозможно, и она не входит в число официальных компонентов FrameNet.

Проект «Лексикограф» идеологически более близок нашим задачам, однако на данный момент охватывает не все значения и тематические классы глаголов с одинаковой степенью детальности: так, в версии базы от 30.10.2010, доступной сейчас онлайн, детально разработаны глаголы физического воздействия, перемещения, звука, однако отсутствуют такие глаголы, как *видеть*, *слышать*, *понимать*, *светиться*, *греметь* и мн. др. (а для включенных в базу глаголов учтены далеко не все значения и конструкции, ср. глагол бить, для которого в «Лексикографе» имеется только два входа — «БИТЬ 1 (палкой по забору)» и «БИТЬ 2 (кого)»). В этой связи, говорить о полном инвентаре ролей не приходится.

Наш проект можно рассматривать как масштабирование идеи «Лексикографа» на больший объем данных[4]. Была поставлена задача создания инвентаря семантических ролей, строящегося на следующих принципах:

- инвентарь должен быть иерархически организован с целью создания более гибкого инструмента поиска и кластеризации: при желании, его можно свести к 5–10 проторолям, в других случаях, он может быть расширен до нескольких десятков и даже сотен ярлыков;
- интерпретация первого и второго аргумента в большей мере зависит от семантики предиката, нежели трактовка третьих, четвертых и т. д. аргументов типа Инструмента, Траектории и т. д.;

[4] При этом мы не преследуем задачу приписать каждому глаголу толкование по некоторой заданной схеме, как это делается в «Лексикографе», а сосредотачиваемся на детальном описании конструкций и связей между ними.

- инвентарь коррелирует с семантической классификацией глагольной лексики[5]: в частности, это означает, что традиционные очень широко понимаемые роли Агенса и Пациенса должны в разных группах получать разные ярлыки;
- объем роли строится по принципу прототипа и периферии: например, прототипом Пациенса является участник, претерпевающий изменение под физическим воздействием контролирующего ситуацию Агенса, периферийные случаи (пациенс нефизического процесса, пациенс, не претерпевающий изменения, пациенс, который создается в результате физического действия и пр.) получают собственные ярлыки (ср. Тема, Результат) и считаются частным случаем роли Пациенса;
- предусматривается возможность сдвоенных ролей и расщепления ролей [Апресян 1974/1995].

За основу для составления списка был взят инвентарь семантических ролей, приведенный в [Апресян и др. 2010: 370–377]. Практическая работа с имеющимися в системе шаблонами конструкций потребовала, однако, внесения в этот инвентарь ряда изменений. Помимо незначительной правки технического характера (так, вместо ярлыка «Пациенс!» мы использовали более самодостаточное наименование «Подвергающаяся воздействию часть пациенса»), в список Ю. Д. Апресяна были внесены изменения в связи с тем, что ряд содержащихся в нем ролей объединяет достаточно разнородные семантические сущности. Если роль соотносилась с несколькими семантическими классами глаголов, то она разделялась нами на несколько — например, роли Экспериенцера в нашей разметке соответствуют семантические роли Субъект восприятия *(видеть, слышать),* Субъект ментального состояния *(думать, понимать),* Субъект психологического состояния *(бояться, любить),* Субъект физиологического ощущения (*болеть, колоть в боку*) и Субъект физиологической реакции *(смеяться, тошнить).* Роль Агенса была сохранена для ядерных агентивных контекстов, но в дополнение к ней в список были включены экспликации Говорящий, Субъект поведения *(лениться, медлить),* Субъект социального отношения *(дружить, помириться),* Субъект перемещения (последняя экспликация используется для всех, не только агентивных, одноместных глаголов перемещения, коррелируя тем самым с их выделением в особый класс; агентивность глагола в этом случае однозначно устанавливается по одушевленности субъекта).

В результате для разметки шаблонов конструкций был использован список из 88 экспликаций, классифицированный по принципу семантической близости на несколько групп: блок Агенса, блок Пациенса; блок Экспериенцера; блоки Инструмента и Адресата, блок обстоятельственных характеристик. Внутри блоков можно выделить группу посессивных ролей, группы ролей Места,

---

[5] Поскольку FrameBank является «дочерним» ресурсом НКРЯ, с надстроенным слоем разметки и интегрированным словарем, он ориентирован на систему глагольных классов Основного корпуса [Kustova et al. 2009], с учетом их дополнения и расширения. Вместе с тем, сам принцип иерархического выделения ролей может быть связан с любыми другими лексическими классификациями.

Времени, Параметров, Признаков, Причины и Цели; группа Источников и Ресурсов объединяет роли из блока Агенса и Места; при максимальном сжатии инвентаря роли группы Экспериенцера можно распределить между агентивными и пациентивными ролями. Семантические роли и их блоки образуют единый граф (см. рис. 3)[6], что позволяет выбирать между разными уровнями дробности поиска, релевантными для конкретной задачи (например, найти как все шаблоны конструкций, в которых реализуются семантические роли из Блока Агенса, так и все шаблоны конструкций, где есть участник с ролью Говорящего).

Для целого ряда шаблонов конструкций оказалось невозможным приписать участнику ровно одну роль, поскольку имело место сочетание семантики двух различных ролей. В этих случаях в разметку вводились двойные семантические роли (ср. здесь [Апресян 1974/1995: 140] об отдельных примерах синкретичного выражения валентностей). Так, в контексте *обрабатывать детали* ***на станке*** речь идет об инструменте совершения действия, но одновременно этот инструмент имеет локативные свойства, поэтому в данном случае использовалась экспликация Инструмент-Место. Конструкция *Пехотинцы строились* ***клином*** описывает результат (то, что получилось в результате построения) и одновременно способ совершения действия, и в этом и подобных случаях в разметку вводилась двойная экспликация Результат-Способ. В конструкции *«Вот это фокус», —* ***удивился*** *он* участник-лицо получил сдвоенную роль Говорящий-Субъект психологического состояния. Очевидно, что сдвоенные роли присутствуют в конструкциях, где либо участник размечен морфосинтаксически нестандартно (ср. *на станке*), либо предикат относится к нескольким лексическим классам (ср. *удивился* — эмоциональное психологическое состояние и говорение).

Следует оговорить, что FrameBank предусматривает и более дробное представление ролей участников: например, для глаголов *служить* и *спасти* экспликации в стиле FrameNet «тот, кому служат» и «тот, кого спасают» будут более точными ярлыками, нежели Контрагент и Пациенс — однако, с одной стороны, такие индивидуальные ярлыки будут редко востребованы пользователями, а с другой стороны, они могут быть порождены автоматически по определенной схеме. В этой связи, основной рабочий статус в системе получает инвентарь из 88 базовых ролей[7].

---

[6] Иерархические отношения между семантическими ролями обозначены на графе сплошными линиями. Пунктирные линии соответствуют семантическим связям между ролями, не связанными непосредственным иерархическим отношением. Стрелки при ролях блока Экспериенцера показывают семантическую близость этих ролей к блоку Агенса или к блоку Пациенса. О-блок объединяет в себе шесть групп обстоятельственных ролей, которые традиционно не сводят к одной гиперроли. Пространственное расположение ярлыков (например, сверху vs. справа) относительно ярлыков ролей верхнего уровня в иерархии не несет какой-либо смысловой нагрузки.

[7] Это число не является абсолютным и, безусловно, со временем будет меняться. В частности, не исключено, что потребуется расширение инвентаря при разметке конструкций имен прилагательных и существительных.

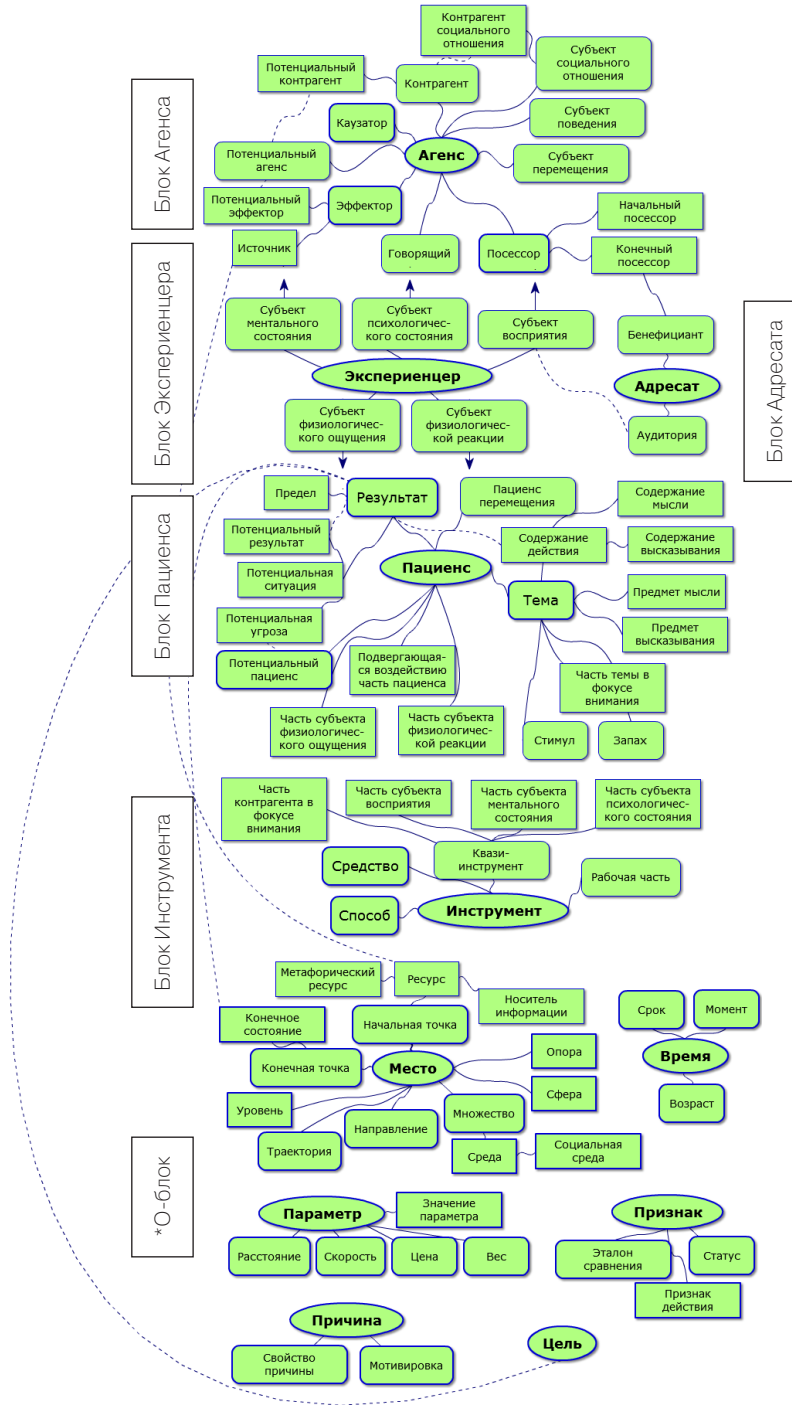


Рис. 3. Иерархия семантических ролей системы «ФреймБанк»

3. Граф глагольных конструкций

3.1. Полисемия глагольной лексики и ее подача в системе FrameBank

Разработка семантической разметки системы FrameBank предполагает также системный анализ полисемии глагольной лексики и создание поискового инструмента при исследованиях в этой области. На решение этой задачи и нацелена ведущаяся на данном этапе работа. Речь идет не только об анализе отдельных семантических полей (ср. [Кустова 2004] о глаголах физического воздействия и лексемах с экспериенциальной семантикой, [Падучева 2004] о фазовых и бытийных глаголах, глаголах восприятия, эмоций, звучания, речи, перемещения, и мн. др.), но и о создании такого ресурса, который предоставлял бы информацию о системных закономерностях семантических переходов глагольной лексики — причем о таких закономерностях, которые формулируются в терминах не только наиболее базовых противопоставлений вроде «конкретный предмет» vs. «абстрактное имя», но и более детальной классификации семантических ролей и семантических ограничений.

В нашей работе мы опираемся на теоретический опыт школы Е. В. Падучевой в осмыслении глагольной полисемии (см. [Падучева 2004] и др.), а также на опыт проекта Базы данных по многозначным качественным прилагательным и наречиям русского языка ([Рахилина и др. 2009], [Карпова и др. 2010], [Карпова и др. 2011]), призванного решить аналогичную задачу для признаковой лексики. Естественно, набор используемых нами ярлыков для семантических переходов не является копией аналогичного инвентаря из базы данных прилагательных и наречий — в первую очередь в силу значительно большей вариативности конструкций у глагольной лексики по сравнению с признаковой лексикой.

При разметке системы FrameBank для каждого из глаголов строится семантическая сеть, иллюстрирующая направления и типы переходов между всеми шаблонами его конструкций. Прежде всего, выделяются внутрифреймовые и межфреймовые связи. Связи между конструкциями, относящимися к одному значению глагола и, соответственно, к одному фрейму, маркируют введение нового участника, замену участника при переносе фокуса внимания, мену коммуникативного статуса и морфосинтаксического оформления и т. п. явления. Межфреймовые связи соединяют конструкции, относящиеся к разным значениям глагола.

Кроме того, различаются связи, маркирующие семантический переход, и формальные связи, маркирующие изменение или наследование формального паттерна конструкции. Примером семантического перехода является метафора — например, при переходе от конструкции *Гром* ***гремит*** к конструкции *Директор* ***гремел****, обличая пороки.* Вместе с тем, для конструкции *Директор гремел…* устанавливается отношение точного формального наследования (Snom V) как с конструкцией *Гром гремит*, так и с конструкцией *Друзья говорили до утра*, обозначающей прототипическую речевую ситуацию.

3.2. Типы переходов между конструкциями одного глагола

Ниже приводится пилотная классификация типов переходов между конструкциями одного глагола, выделенных по результатам обработки 10% глагольной лексики в системе «Фреймбанк»; разметка связей между конструкциями разных глаголов еще предстоит. Классификация включает различные комбинации изменения/сохранения плана выражения и плана содержания конструкций.

А1. Мена морфосинтаксического оформления участника: *занес этот факт в протокол → занес в протокол, что судья обрывает его;* в т. ч. в зависимости от типа лексического заполнения элемента: *он занес ногу за порог → через плетень → на ступеньку брички;*

А2. Мена статуса участников, диатетический сдвиг: *протираю стол от пыли → протираю пыль на столе*;

А3. Невыражение участника, относящегося к известному классу: *наши следы занесло снегом → наши следы занесло*

А4. Невыражение участника, дейктически или ситуативно известного: *он выписал все адреса из справочника → он выписал все адреса*

А5. Невыражение участника, неопределенного (неважного) в ситуации: *птицы летят на юг → летят птицы*

А6. Добавление участника: *вода собирается — вода собиралась каплями*; частный случай добавления — гибрид двух конструкций: *рыбку занесло из речки, рыбку занесло в протоку → рыбку занесло из речки в соленый океан*

А7. Мена участников (перенос фокуса с одного участника на другого): *он выписал все адреса из справочника → он выписал все адреса в тетрадь*

Б. Добавление неядерного участника ситуации: в производную конструкцию эксплицитно добавляется участник, не предусмотренный прототипом фрейма: *Птица летит → Птица летит за кормом. Вахтер выписал пропуск → Вахтер выписал мне пропуск.*

В1. Специализация значения фрейма, связанная с невыражением одного из участников: *Иван пьет чай → Антон пьет*; *Мы говорили о прошлом → Ребенок уже говорит* (‘умеет говорить’).

В2. Идиоматизация значения, связанная с введением в конструкцию новых лексических констант (в частности, вместо переменных-участников): *Он опустил кулак → Он опустил руки* (‘перестать действовать, потеряв надежду’).

Г1. Метонимия: смежный участник. Используется при замене участника на другого, смежного участника в пределах одного фрейма: *слушать музыку → слушать Баха*

Г2. Метонимия: перераспределение коммуникативных акцентов между участниками ситуации (при диатетическом сдвиге, А2): *Суд слушает дело → В суде слушают дело.*

Г3. Метонимия: сдвиг домена: *Я вас **любил*** (по МАС, ‘чувствовать сердечную склонность к лицу другого пола’) → *Она своих девочек очень **любит*** (‘чувствовать глубокую привязанность к кому-л., быть преданным кому-л.’).

Г4. Метонимия: смежный класс. Используется при переходе глагола в смежный с исходным тематический класс, ср. *Вечером, сидя за чаем, Семён Семёнович со скучающим видом* ***слушал*** (глагол восприятия) *жену, которая что-то записывала на бумажке…* [И. А. Ильф, Е. П. Петров (1935)] → *Хотя мэр Москвы по стилю своего публичного поведения, безусловно, принадлежит к людям, готовым* ***слушать*** (‘принимать во внимание слова, просьбы, советы’, глагол ментального состояния) *москвичей*. [«Известия» (2001)] → *Нет человека, властного над ветром, умеющего удержать ветер, особенно когда этот ветер в голове. Не хотят* ***слушать*** (‘подчиняться распоряжениям, следовать советам, слушаться’, глагол поведения) *старших — пусть идут. Пусть хлебнут горя своей золотой ложечкой.* [М. Успенский (1995)]

Д. Метафора. Используется при смене таксономического класса какого-либо из участников ситуации, сопровождающемся сдвигом значения глагола: *В парке борются <u>два парня</u>* → *На выборах борются <u>две партии</u>*.

Е. Ребрендинг. Понимается нами в соответствии с [Бонч-Осмоловская и др. 2009], [Рахилина и др. 2009], [Рахилина и др. 2010а], [Рахилина и др. 2010б], [Карпова и др. 2011] как семантический переход лексемы в другой таксономический класс, основанный на механизме импликатуры (т. е. результат семантического перехода является следствием или выводом из исходного значения), ср. переход *Солдат <u>стреляет</u>* → *У него <u>стреляет</u> в голове.*, где происходит явная смена таксономического класса глагола *стрелять* (глагол физического воздействия → глагол болевого ощущения), а производное значение осмысляется как вероятный результат действия, подразумеваемого исходным значением (субъект испытывает такое ощущение, как будто в боку происходит действие стрелять).

Ж. Другие, более далекие и менее прозрачные переходы. Ср., например, *выступить <u>из толпы</u>* → *выступить <u>на совещании</u>*.

З. Грамматикализация: выветривание значения в случае, когда глагол принимает роль лексической функции: *являться, выступать* (*свидетелем*), *обратить* (*внимание*), *питать* (уважение) и т. д.

Нередко глагольные значения связаны не непосредственно, а через цепочку «посредников» в значениях других предикатов. Ср., например, конструкции *занести письмо домой* и *Войдет и занесет такую чушь…* ‘начать нести (чушь)’, которые связаны через посредство разных значений глагола *нести*; конструкцию *собраться с силами*, которая очевидно соотносится с конструкцией *собрать силы(, чтобы встать)*.

4. Заключение. О соотношении семантических ролей участников, семантических ограничений, классов глаголов

Разметка семантических ролей участников конструкций и систематизация переходов между конструкциями позволит выявить закономерности системы полисемии глагольной лексики в ее связи со свойствами конструкций, в которых реализуется конкретный глагол. Так, с использованием базы можно будет выявить, с одной стороны, типы переносов, характерных для глаголов какого-либо исходного семантического класса (и свойства соответствующих конструкций), с другой стороны, типы переносов (и свойства конструкций), результатами которых являются глаголы заданного класса — ср., например, перенос в семантическую область речи из областей перемещения (***Летят*** *птицы* → ***Летит*** *молва*), психологического состояния (*Парень* ***волновался*** → *«Не догонит!» —* ***волновался*** *парень*), физического воздействия (*Хозяйка* ***отрезала*** *кусок хлеба* → *«И слышать об этом не хочу!» —* ***отрезала*** *хозяйка*) и др.

Обсуждение такого рода явлений поднимает и теоретические вопросы о соотношении семантических ролей участников, семантических ограничений на заполнение валентностей, а также глагольных классов. В частности, для метафорических переносов (по определению предполагающих изменение семантических ограничений на заполнение хотя бы одной валентности) в базе обнаруживаются следующие возможности:

- Смена семантического класса глагола и семантических ролей участников, ср. ***Летят*** *птицы* (перемещение, Субъект перемещения) → *Время* ***летит*** (скорость перемещения, Время) и ***Летят*** *птицы* (перемещение, Субъект перемещения)→ ***Летит*** *молва* (речь, Содержание высказывания).
- Сохранение семантического класса глагола и семантических ролей участников. Ср., например, переходы ***Летят*** *птицы* → *Конь* ***летит****, ветер свистит в ушах* и ***Летят*** *птицы* → *Не раз он* ***летел*** *кубарем* (во всех примерах глагол лететь относится к классу глаголов перемещения, а субъект получает семантическую роль Субъект перемещения).
- Смена семантического класса глагола при сохранении паттерна семантических ролей: *Мальчик* ***ест*** *хлеб* (уничтожение, Агенс + Пациенс) → *Мошки* ***едят*** *лошадей* (физическое воздействие, Агенс + Пациенс)[8].

Детальное исследование этих вопросов предполагается сделать возможным с использованием системы FrameBank. Перспективным направлением

[8] В этой связи встает вопрос о регулярности соответствия между семантической классификацией глаголов и приписываемыми их аргументам семантическими ролями. Действительно, в большинстве случаев глаголам разных классов в системе приписываются разные наборы ролей. Но в ряде случаев классификация глагольной лексики может быть и более дробной (что отражается, например, в моделях метафорических сдвигов), как в рассматриваемом примере, где глаголы уничтожения по сути являются подклассом глаголов физического воздействия (ср. также глаголы перемещения, внутри которых выделяется подкласс глаголов падения, но семантической ролью субъекта в любом случае является Субъект перемещения) — однако семантические характеристики самих участников и отношения между этими участниками в таких случаях, как представляется, очень близки и могут быть сведены к одинаковым наборам ролей.

развития системы предполагается и создание графа фреймов (аналогичного имеющемуся во FrameNet), который бы послужил дополнением к разрабатываемому сейчас графу конструкций и содержал эмпирический материал для обсуждения связи фреймов с семантическими ролями, семантическими ограничениями и классами глаголов.

Литература

1. *Апресян Ю. Д.* Избранные труды, том I. Лексическая семантика. М., 1995. 1-е изд.: М., 1974.
2. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Санников В. З.* Теоретические проблемы русского синтаксиса: взаимодействие грамматики и словаря. М., 2010.
3. *Апресян Ю. Д., Палл Э.* Русский глагол — венгерский глагол. Управление и сочетаемость. Будапешт, 1982.
4. *Бонч-Осмоловская А. А., Рахилина Е. В., Резникова Т. И.* Глаголы боли: лексическая типология и механизмы семантической деривации // Концепт боль в типологическом освещении / под ред. В. М. Брицына, Е. В. Рахилиной, Т. И. Резниковой, Г. М. Яворской. Киев, 2009. С. 8–27.
5. *Карпова О. С., Резникова Т. И., Архангельский Т. А., Кюсева М. В., Рахилина Е. В., Рыжова Д. А., Тагабилева М. Г.* База данных по многозначным качественным прилагательным и наречиям русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М.: РГГУ, 2010. С. 163–168.
6. *Карпова О. С., Рахилина Е. В., Резникова Т. И., Рыжова Д. А.* Оценочные значения ребрендингового типа в признаковой лексике (по материалам Базы данных семантических переходов в качественных прилагательных и наречиях) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). М.: РГГУ, 2011. С. 292–304.
7. *Кузнецов И. О.* Автоматическое выделение глагольных актантов: теоретическая основа и актуальные подходы // НТИ. Сер. 2 (в печати).
8. *Кустова Г. И.* Типы производных значений и механизмы языкового расширения. М., 2004.
9. *«Лексикограф»*. Электронный ресурс: http://lexicograph.ruslang.ru
10. *Лютикова Е. А., Татевосов С. Г., Иванов М. Ю., Пазельская А. Г., Шлуинский А. Б.* Структура события и семантика глагола в карачаево-балкарском языке. М., 2006.
11. *Ляшевская О. Н., Кузнецова Ю. Л.* Русский фреймнет: к задаче создания корпусного словаря конструкций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 306–312.

12. *Падучева Е. В.* Динамические модели в семантике лексики. М., 2004.
13. *Плунгян В. А.* Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира. М., 2011.
14. *Рахилина Е. В., Карпова О. С., Резникова Т. И.* Модели семантической деривации многозначных качественных прилагательных: метафора, метонимия и их взаимодействие // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 420–425.
15. *Рахилина Е. В., Резникова Т. И., Карпова О. С.* Семантические переходы в атрибутивных конструкциях: метафора, метонимия и ребрендинг // Лингвистика конструкций / Отв. ред. Е. В. Рахилина. М., 2010а. С. 398–455.
16. *Рахилина Е. В., Резникова Т. И., Бонч-Осмоловская А. А.* Типология преобразования конструкций: предикаты боли // Лингвистика конструкций / Отв. ред. Е. В. Рахилина. М., 2010б. С. 456–540.
17. *Рахилина Е. В.* (ред.) Лингвистика конструкций. М., 2010.
18. *Dowty, D. R.* (1991), Thematic proto roles and argument selection, Language 67, pp. 547–619.
19. *Fillmore Ch. J.* (1968), The Case for Case, in Bach E. and Harms (Ed.), Universals in Linguistic Theory. New York, pp. 1–88.
20. *Fillmore Ch. J.* (1977), The case for case reopened, in Cole P., Sadock J. M. (eds.), Grammatical Relations, Acad. Press, New York, pp. 59–81.
21. *Fillmore Ch. J.* (1982), Frame semantics, Linguistics in the morning calm: Selected papers from the SICOL-1981, Hanship, Seoul, pp. 111–137.
22. *FrameNet*. Электронный ресурс: http://framenet.icsi.berkeley.edu
23. *Gildea D., Jurafsky D.* (2000), Automatic labeling of semantic roles, Proc. of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), pp. 512–520.
24. *Kustova Galina I., Olga N. Lashevskaja, Elena V. Paducheva, and Ekaterina V. Rakhilina* (2009), Verb taxonomy: from theoretical lexical semantics to practice of corpus tagging, in Lewandowska B., K. Dziwirek (eds.), Cognitive Corpus Linguistics Studies. Frankfurt: Peter Lang.
25. *Lyashevskaya O.* (2010), Bank of Russian Constructions and Valencies // Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010. Valletta: ELRA, 2010. P. 1802–1805.
26. *RussNet:* тезаурус русского языка.
Электронный ресурс: http://project.phil.spbu.ru/RussNet

References

1. *Apresjan Ju. D.* (1995), Selected papers, Vol. 1, Lexical Semantics [Izbrannye trudy, tom I. Leksicheskaja semantika], Jazyki Russkoj Kul'tury, Vostochnaja Literatura, Moscow.
2. *Apresjan Ju. D., Boguslavskij I. M., Iomdin L. L., Sannikov V. Z.* (2010), Theoretical issues of Russian syntaxs: the interrelation between grammar and vocabulary [Teoreticheskie problemy russkogo sintaksisa: vzaimodejstvie grammatiki i slovar'a], Jazyki slavjanskih kul'tur, Moscow.
3. *Apresjan Ju. D., Pall E.* (1982), Russian verb — Hungarian verb. Government and combinability [Russkij glagol — vengerskij glagol. Upravlenie i sochetaemost'], Tankyonvkiado, Budapest.
4. *Bonch-Osmolovskaja A. A., Rakhilina E. V., Reznikova T. I.* (2009), Pain verbs: lexical typology and mechanisms of semantic derivation [Glagoly boli: leksicheskaja tipologija i mehanizmy semanticheskoj derivatsii], in The concert of pain from a typological point of view [Kontsept bol' v tipologicheskom osveschenii], Vidavnicjij Dim Dmitra Burago, Kiev, pp. 8–27.
5. *Dowty, D. R.* (1991), Thematic proto roles and argument selection, Language 67, pp. 547–619.
6. *Fillmore Ch. J.* (1968), The Case for Case, in Bach E. and Harms (Ed.), Universals in Linguistic Theory. New York, pp. 1–88.
7. *Fillmore Ch. J.* (1977), The case for case reopened, in Cole P., Sadock J. M. (eds.), Grammatical Relations, Acad. Press, New York, pp. 59–81.
8. *Fillmore Ch. J.* (1982), Frame semantics, Linguistics in the morning calm: Selected papers from the SICOL-1981, Hanship, Seoul, pp. 111–137.
9. *FrameNet.* An online resource, available at: http://framenet.icsi.berkeley.edu
10. *Gildea D., Jurafsky D.* (2000), Automatic labeling of semantic roles, Proc. of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), pp. 512–520.
11. *Karpova O. S., Reznikova T. I., Arkhangel'skij T. A., Kjuseva M. V., Rakhilina E. V., Ryzhova D. A., Tagabileva M. G.* (2010), A database of polysemous qualitative adjectives and adverbs in Russian [Baza dannyh po mnogoznachnym kachestvennym prilagatel'nym I narechijam russkogo jazyka], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intelleltual'nye tehnologii: po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog"], Moscow, pp. 163–168.
12. *Karpova O. S., Rakhilina E. V., Reznikova T. I., Ryzhova D. A.* (2011), Evaluative meanings of the rebranding type in qualitative lexemes [Otsenochnye zhachenija rebrendingovogo tipa v priznakovoj leksike], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intelleltual'nye tehnologii: po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog"], Moscow, pp. 292–304.

13. *Kustova G. I.* (2004), Types of figurative meanings and mechanisms of linguistic broadening [Tipy proizvodnyh zhachenij i mehanizmy jazykovogo rasshirenija], Jazyki slavjanskoj kul'tury, Moscow.
14. *Kustova Galina I., Olga N. Lashevskaja, Elena V. Paducheva, and Ekaterina V. Rakhilina* (2009), Verb taxonomy: from theoretical lexical semantics to practice of corpus tagging, in Lewandowska B., K. Dziwirek (eds.), Cognitive Corpus Linguistics Studies. Frankfurt: Peter Lang.
15. *Kuznetsov I. O.* (forthcoming), Automatic extraction of verb arguments: theoretical grounds and state of the art [Avtomaticheskoe vydelenie glagol'nyh aktantov: teoreticheskaja osnova i aktual'nye podhody], Scientific and Technical Information. Ser. 2. Information Processes and Systems [Nauchnaja i tehnicheskaja informacija. Ser. 2. Informatsionnye processy i sistemy].
16. *Lexicographer* [Leksikograf].
An online database, available at: http://lexicograph.ruslang.ru
17. *Ljutikova E. A., Tatevosov S. G., Ivanov M. Ju., Pazel'skaja A. G., Shluinskij A. B.* (2006), Event structure and verb semantics in Karachaj-Balkar [Struktura sobytija i semantika glagola v karachaevo-balkarskom jazyke], IMLI RAN, Moscow.
18. *Lyashevskaya O.* (2010), Bank of Russian Constructions and Valencies, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, pp. 1802–1805.
19. *Lyashevskaya O. N., Kuznetsova Ju. L.* (2009), Russian FrameNet: constructing a corpus-based dictionary of constructions [Russkij Frejmnet: k zadache sozdanija korpusnogo slovarja konstruktsij], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intelleltual'nye tehnologii: po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog"], Moscow, pp. 306–312.
20. *Màrquez L., Carreras X., Litkowski K. C., Stevenson S.* (2008), Semantic role labeling: an introduction to the special issue, Computational Linguistics, Vol. 34–2, pp. 145–159.
21. *Paducheva E. V.* (2004), Dynamic patterns in lexical semantics [Dinamicheskie modeli v semantike leksiki], Jazyki slavjanskoj kul'tury, Moscow.
22. *Plungian V. A.* (2011), Introduction to grammatical semantics: grammatical meanings and grammatical systems of the world's languages [Vvedenie v grammaticheskuju semantiku: grammaticheskie znachenija i grammaticheskie sistemy jazykov mira], RSUH, Moscow.
23. *Rakhilina E. V., Karpova O. S., Reznikova T. I.* (2009), Patterns of semantic derivation in polysemous qualitative adjectives: metaphor, metonymy, and their interaction [Modeli semanticheskoj derivatsii mnogoznachnyh kachestvennyh prilagatel'nyh: metafora, metonimija i ih vzaimodejstvie], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'juternaja lingvistika i intelleltual'nye tehnologii: po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog"], Moscow, pp. 420–425.

24. *Rakhilina E. V., Reznikova T. I., Karpova O. S.* (2010a), Semantic shifts in attributive constructions: metaphor, metonymy, and rebranding [Semanticheskie perehody v atribultivnyh konstrukcijah: metafora, metonimija i rebrending], in Linguistics of Constructions [Lingvistika konstrukcij], Azbukovnik, Moscow, pp. 398–455.
25. *Rakhilina E. V., Reznikova T. I., Bonch-Osmolovskaja A. A.* (2010b), Typology of constructional changes: pain predicates [Tipologija preobrazovanija konstrukcij: predikaty boli], in Linguistics of Constructions [Lingvistika konstrukcij], Azbukovnik, Moscow, pp. 456–540.
26. *Rakhilina E. V.* (ed., 2010) Linguistics of Constructions [Lingvistika konstrukcij], Azbukovnik, Moscow.
27. *RussNet:* a Russian language thesaurus.
Available at: http://project.phil.spbu.ru/RussNet

ДИСКУРСИВНАЯ ТАКСОНОМИЯ

Кибрик А. А. (aakibrik@gmail.com)

Институт языкознания РАН;
МГУ им. М. В. Ломоносова, Москва, Россия

Один из центральных разделов теории дискурса — дискурсивная таксономия, то есть выявление оснований, по которым дискурсы делятся на типы. Таких оснований несколько, и они нередко смешиваются. Основные среди них — это модус, жанр и функциональный стиль. Различие по модусу касается типа носителя: устный vs. письменный. Жанры связаны с типами коммуникативных целей, признаваемых тем или иным дискурсивным сообществом, и характеризуются стандартными схемами. Функциональные стили выделяются в связи с различными сферами человеческого бытия. Есть и другие дискурсивные таксономии — так, очень важны типы изложения, характеризующие не целые дискурсы, а отдельные их фрагменты, или пассажи. Каждая из дискурсивных таксономий имеет рефлексы в области грамматических, лексических и иных локальных языковых выборов. Такие выборы являются равнодействующей всех факторов, связанных с дискурсивными таксономиями. Хотя разные дискурсивные таксономии в принципе независимы, выделяемые по разным основаниям типы дискурса могут иметь сходные рефлексы — это касается, например, письменного модуса и официального функционального стиля.

DISCOURSE TAXONOMY

Kibrik A. A. (aakibrik@gmail.com)

Institute of Linguistics RAS;
Lomonosov Moscow State University, Moscow, Russia

Among the central issues in the theory of discourse is discourse taxonomy, that is elucidation of the parameters classifying discourses into types. There are several such parameters, and they are often cofused. The main ones include mode, genre, and functional style. The distinction in mode concerns the medium: spoken or written. Genres are related to the typical communicative goals, acknowledged by discourse communities, and are characterized by standard schemata. Functional styles are identified in connection with the various domains of human existence. There are other discourse taxonomies as well, in particular, quite important is the distinction between types of presentation that characterize not whole discourses but their fragments, or passages. Each discourse taxonomy reflects upon grammatical, lexical and other local linguistic choices. Such choices are a resultant of all factors stemming in discourse taxonomies. Even though discourse taxonomies are in principle independent from each other, discourse types established on the basis of different parameters may have similar properties. For example, the written mode and the official functional style have similar reflexes in the linguistic structure.

БОЛЬШЕ ЕДИНИЦЫ: РУССКИЕ ИДИОМЫ С КОМПОНЕНТОМ *ОДИН/ЕДИН*[1]

Киселева К. Л. (xenkis@mail.ru),
Вознесенская М. М. (voznes-masha@yandex.com),
Козеренко А. Д. (akozerenko@mail.ru)

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

Рассматривается фрагмент фразеологической системы русского языка, содержащий идиомы с компонентом *один/един*, такие, как *номер один, один на [всем] [белом/божьем] свете, одним махом, одно название, все как один, все до единого, игра в одни ворота, дудеть в одну дуду, одного поля ягоды, ставить на одну доску, как одна копеечка* и многие другие. Показано, что в разных типах идиом (к примеру, в идиомах типа *одно название* и типа *одной ногой*) представлены разные лексические значения *один*. Соответственно, отличается и тот вклад, который этот компонент вносит во фразеологическую семантику, и те модели внутренней формы, с которыми он соотносится. Присутствие компонента *один/един* в составе идиомы создает предпосылки для такой модели внутренней формы, которая является той или иной операцией над множеством. На этом основании были выделены группы идиом со значением «единичность», «минимальный ресурс», «совпадение», «сходство», «унификация», «пустое множество», «псевдоисчерпание», «сведение множества к элементу» и др.

Ключевые слова: фразеология русского языка, семантика, модели внутренней формы, числа

MORE THAN ONE: RUSSIAN IDIOMS WITH *ODIN/EDIN* COMPONENT

Kiseleva K. L. (xenkis@mail.ru),
Voznesenskaja M. M. (voznes-masha@yandex.com),
Kozerenko A. D. (akozerenko@mail.ru)

V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

[1] Работа выполнена при поддержке РГНФ, гранты 12-34-10413, 12-04-12054, 12-04-12055, и Министерства образования и науки РФ, заявка 2012-1.1-12-000-3004-052 (Разработка интерактивной системы «Корпус текстов по русской фразеологии»).

The paper deals with a part of Russian phraseology:, the idioms containing the *odin/edin* ('one', 'single') lexical component, e.g. *vse kak odin, odin-edinstvennyj, vse do edinogo, odnoj levoj, ni odna zhivaja dusha, iz odnogo testa* etc. (English equivalents for: 'one and all', 'all alone', 'all down to the last one', 'with one hand tied behind one's back', 'not a one living soul', 'cut from the same cloth'). We observe that, first, the meaning of the idioms containing *odin/edin* depends on the meaning of the word *odin* in this context (ex. in *smekh odin* and *v odin prisest* we have two different lexical meanings of *odin*). Second, we try to classify these idioms according to the inner form model that we see in each case. For example, *vse do odnogo* is based on the model labeled "exhaustion" while the similar idiom *vse kak odin* is based on another model, labeled "matching". Apart from suggesting several classes of idioms depending on their inner form model, we show that the presence of the component *odin* systematically brings two semantic effects to the meaning of the idioms: uniqueness, oneness, wholeness vs. insufficiency, poorness, lameness.

Key words: Russian phraseology, semantics, idioms, inner form models, lexicography, numbers

0. В данной работе мы продолжаем изучение чисел и операций с ними в составе русских идиом, начатое в (Вознесенская и др. 2013)[2]. Объектом описания на этот раз служат фразеологические единицы, содержащие компонент *один* (также в форме *един-*). *Один* представляется нам интересным для изучения и значимым компонентом в структуре идиом не только в силу его фразеологической продуктивности, но и потому, что это «самое первое число, начало цифрового ряда, обращает внимание на внутреннюю «счетность» объекта, на то что он *один* и *един*, на *единство одного,* состоящего из *единиц» (*Арутюнова 2005: 15).

Слово один обладает развитой многозначностью. Так, в (МАС 1999) у него выделяются девять значений: собственно числительное, пять адъективных значений и три местоименных. К адъективным относятся:

- 'без других, в отдельности, в одиночестве';
- 'никто другой или ничто другое, кроме; единственный';
- 'тот же самый, тождественный, одинаковый';
- 'целостный, неделимый, единый';
- *один* в сочетании с *другой* для противопоставления.

К местоименным относятся:

- *один* в сочетании с предлогом *из* — для выделения единичного предмета из ряда;
- *один* в сочетании с *другой* — при перечислении и противопоставлении;
- 'какой-то, некий'.

---

[2] Мы опускаем перечисление работ, посвященных числу во фразеологии. Упомянем лишь статью (Dobrovol'skij, Piirainen 1998), где рассматривается символьная функция чисел в идиомах различных языков.

И. А. Мельчук в работе «Поверхностный синтаксис русских числовых выражений» (1985) различает четыре адъективные лексемы один, выражающие значения «некоторые», «в одиночку», «только» и «один и тот же» (цит. по Арутюнова 2005: 15).

Мы рассмотрим различные семантические группы идиом, содержащих компонент *один/един*. Можно предположить, что (1) в каждой группе идиомы с компонентом *один/един* будут иметь некий общий семантический признак; (2) компонент *один/един* присутствует в идиомах в разных значениях и (3) идиомы могут быть объединены в группы на основе общей модели внутренней формы, поскольку внутренняя форма — «не случайная характеристика, а регулярный фактор, организующий определенные фрагменты фразеологической системы» (Баранов, Добровольский 1998).[3]

1. Единичность: *один-единственный* vs *один-одинешенек*

Наиболее близок к базовому количественному значению слова «один» тот тип, который мы здесь обозначили как «единичность». Идея единичности может быть реализована по-разному: в фокусе может находиться единственный элемент, обладающий некоторым свойством *(один-единственный; один бог знает, одному богу ведомо),* либо единичность определяется через отсутствие других элементов множества *(один-одинешенек; один как сыч; один как перст)*.[4]

Идея единичности оценивается положительно и отрицательно в зависимости от контекста:

(1) *Ты понял меня? я знаю о ней все, даже ее имя. Ты не можешь, ты не должен знать это имя, ее имя знаю только я — **один на всем свете**. Ты просчитался: Вета, ее зовут Вета, я люблю женщину по имени Вета Акатова.* (Саша Соколов. Школа для дураков)

(2) *«От любви бывают дети. Ты теперь **один на свете.** Помнишь песню, что, бывало, я в потемках напевала? Это — кошка, это — мышка. Это — лагерь, это — вышка. Это — время тихой сапой убивает маму с папой».* (И. Бродский. Представление)

---

[3] Приводимые толкования написаны в формате, принятом в (ФОС 2009), или заимствованы с некоторыми модификациями из этого словаря.

[4] Интересна структура идиом *один-единственный* и *один-одинешенек*, которая представляет собой своеобразную «автодефиницию»: второй компонент *(единственный, одинешенек)*, будучи однокоренным дериватом слова *один*, проясняет, уточняет значение первого компонента. Особый случай составляют идиомы с редупликацией компонента *один*: *один на один, один в один, один к одному, одно к одному, один за одним*. В данной статье они исключены из рассмотрения.

К первой подгруппе примыкает идиома *номер один*, во внутренней форме которой заключена идея ранжирования элементов по количественному признаку, так что элемент, занимающий первое место, интерпретируется как уникальный в качественном отношении. *Номер один* означает, что (кто-л./что-л.) в наибольшей степени обладает некоторым свойством среди других элементов множества: *певица (враг, театр...) номер один*. Применительно к сущностям типа «проблемы» эта идиома указывает, что элемент, о котором идет речь, является самым важным или самым сложным в данной ситуации: *загадка (вопрос, задача...) номер один*.

2. Минимальный ресурс: быстро, легко... *мало*

2.1.

Прототипическим способом реализации значения минимального ресурса служат идиомы, представляющие собой сочетания существительного с прилагательным «один» в косвенных падежах: *одним/единым духом; одним/единым махом; одним ударом; одним движением руки; на одном/едином дыхании; в/за один присест*. Эти идиомы означают, что некоторое действие было совершено очень легко и/или быстро, что описывается как такой способ действия (ср. распространенный в этих конструкциях творительный падеж), который потребовал от субъекта минимального (буквально единичного) ресурса. В качестве ресурса может выступать множество движений *(одним махом, одним движением руки, одним росчерком пера; в один присест)*, множество «квантов» дыхания *(одним духом, на одном дыхании)* или множество частей тела, способных выступать в качестве инструмента *(одним мизинцем)*.

2.2.

Разновидностью идиом со значением минимального ресурса можно считать случаи, в которых имеет место неполнота пары. О том, что идея неполноты пары имеет прямое отношение к числу «один» как началу числового ряда, говорится в (Степанов 1989: 61): «Если, таким образом, русск. *один* содержит и.-е. *od-/l*ed-, то к тому же корню естественно отнести англ. odd ‘нечетный, странный’. <…> Первоначальное значение русского и английского слов в таком случае одно и то же — ‘оставшийся вне пары, без пары, стоящий особняком’». Когда речь идет о парных сущностях (глаза, ноги, ворота на спортивном поле…), ситуация неполноты пары оценивается отрицательно, как ущербная, поскольку в норме в ней должны быть задействованы оба члена пары. Это может иметь разные семантические следствия: значение минимального используемого ресурса *(одной левой, хоть/хотя [бы] одним глазком [взглянуть])*, минимального времени, затраченного на выполнение действия *(одна нога здесь, другая там)*, неполноты ситуации *(одной ногой; одной ногой*

в могиле), или нарушения нормы *(игра в одни ворота, улица с односторонним движением)*:

ХОТЬ/ХОТЯ [БЫ] ОДНИМ *ГЛАЗКОМ* [ВЗГЛЯНУТЬ… (на кого-л./что-л.)] — выражение очень сильного желания увидеть что-л./кого-л. недоступного в данный момент для субъекта, в форме согласия даже на минимальный визуальный контакт с кем-л./чем-л., при котором задействован только один орган зрения из пары.

(3) *И вот по мере того, как я облазил всю доступную мне здешнюю округу и одно лишь Коргозеро оставалось белым пятном, все больше и больше меня туда тянуло и хотелось* ***хоть одним глазком взглянуть*** *и на деревню, и на озеро.* (А. Варламов. Падчевары)

ОДНОЙ *НОГОЙ* (кто-л. где-л.) начать переход из одного состояния в другое или из одного места в другое и почти завершить его, что осмысляется как наполовину завершенное единичное движение при ходьбе.

(4) *То, что президент одной* ***ногой уже*** *в строю, стало ясно, когда кремлевская пресс-служба сообщила о встрече главы государства со своим помощником.* (Корпус Публ.)[5]

***ИГРА* В ОДНИ ВОРОТА** конфликтная ситуация, в которой один из ее участников находится в заведомо слабом, неблагоприятном положении, что сравнивается с матчем, в котором голы забиваются только одной команде.

(5) *«Папаратце», между тем, бегали вокруг и снимали эту так называемую беседу, матч в одни ворота, ибо мерзавцы не давали Тане и рта раскрыть.* (В. Аксенов. Остров Крым)

2.3.

В тех случаях, когда указание на то, что речь идет об одном элементе, вступает в противоречие с нормальной мощностью данного множества, возникает семантическое следствие «мало, сильно меньше обычного». Такой семантический механизм реализуется в идиомах *одна извилина; жить одним днем; на один зуб*. Так, идиома *на один зуб* в одном значении может интерпретироваться как ‘мало, и это плохо’ *(Мише такая порция* ***на один зуб****)*, а в другом — как ‘мало, т. е. легко, и это хорошо’:

(6) *— Сколько их? — Вольдемар передернул затвор автомата. — Человек двадцать. — На один зуб. Дай гляну.* (А. Михайлов. Капкан для одинокого волка)

---

[5] Здесь и далее атрибуция ряда примеров отсылает к корпусам текстов, созданным в отделе экспериментальной лексикографии Института русского языка им. В. В. Виноградова РАН.

3. Два в одном: совпадение, сходство, тождество

Для идиом данной группы можно предположить, что компонент *один/един* употребляется в значении ‘тот же самый, тождественный; одинаковый’ (МАС 1999). Три подгруппы, которые выделены ниже, соответствуют разным контекстам, в которых происходит сопоставление двух или более элементов множества. В первом случае имеет место случайное или намеренное совпадение (временное, качественное, формальное и др.). Во втором сходство двух и более элементов является результатом их пространственной или иной близости, общего происхождения и условий. Наконец, в третьем случае идиома описывает саму операцию унификации двух и более элементов, при котором они «насильственно» приобретают сходство. Следует отметить, что, собственно, слово *один* в идиомах данной группы обозначает вовсе не элемент множества, а как раз само множество как нечто цельное, что находит отклик в одной из трактовок числа «один»: «В текстах наиболее древнего типа 1 или вовсе не появляется или встречается крайне редко. При этом кажется оправданным предположение, что 1 означало, как правило, не столько первый элемент ряда в современном понимании числового ряда, сколько целостность, главной чертой которой выступает нерасчлененность» (Топоров 2010: 137).

3.1. Совпадение

В идиомах *в один голос; дудеть в одну дуду; на одно лицо; один черт* имеет место совпадение двух событий по некоторому признаку. В разных значениях идиома *в один голос* выражает совпадение одинаковых реплик во времени (7) или совпадение независимо высказанных мнений (8):

(7) *— Последние находки в Сахаре и Месопотамии позволяют думать, что в далекие времена на Земле побывали пришельцы из космоса. — Может быть, те самые марсиане? —* ***в один голос*** *ахнули дамы.* (В. Аксенов. Апельсины из Марокко)

(8) *Специалисты* ***в один голос*** *предсказывали победу чемпиону мира.* (Корпус Публ.)

В идиоме *дуть/дудеть в одну дуду* совпадение может интерпретироваться как совпадение мнений разных людей, совпадение высказываний одного субъекта в разные моменты времени или совпадение «вектора» разных действующих сил, что в любом случае осмысляется как использование (совместное или многократное) одного и того же музыкального инструмента:

(9) *Вот какая странность: утонченный литературовед, человек брокгауз-ефроновской эрудиции и ведущий некогда популярной телепередачи «Вокруг смеха», точно сговорившись,* ***дуют в одну дуду****.* (Корпус Публ.)

(10) *Куплю дом в деревне, буду как Толстой. <…> Там же в деревне и вторую книгу напишу. Совсем другую, ни у кого даже мысли не будет «ах, он повторяется!», у меня не будет кризиса «второго альбома», как у всяких поп-групп, и я не стану как Карлос Сантана* ***дудеть в одну дуду*** *годами.* (ХудТексты Инт.)

(11) *Экономика стимулирует человека денежкой. И хорошо бы, чтобы это стимулирование денежкой совпадало со стимулами, которые дает культура. Чтобы они* ***дули в одну дуду****. Так, например, получается немец, которому совесть не позволяет гайку недокрутить.* (Огонек)

3.2. Сходство по смежности

Внутренняя форма идиом этой группы строится вокруг пространственной метафоры: близость в пространстве (прошлая или нынешняя) переинтерпретируется в актуальном значении идиомы как взаимозависимость, общность происхождения, сходство свойств, единство задач и т. п. К ядру этой группы относятся такие идиомы, как *в одной каше вариться (с кем-л.); в одном котле вариться (с кем-л.); под одной крышей; в одной лодке; в одном флаконе; дышать одним воздухом; в одной связке (с кем-л.); звенья одной цепи; одного поля ягода; одним миром мазаны; из одного теста, одна сатана*:

(12) *Гибридность власти находит отражение и в том, что она включает в себя представителей всех политических сил.* ***В одной лодке*** *уживаются, казалось бы, несовместимые личности — либералы и левые, державники и популисты.* (Корпус Публ.)

В свою очередь отказ от пространственной близости с кем-либо в ситуации, когда эта близость вообще не предполагает никакого взаимодействия, может интерпретироваться как крайне отрицательное отношение говорящего к другому субъекту и отказ иметь с ним что-л. общее, ср. выражение *я рядом/на одном поле… срать не сяду* (с кем-л.).

Происхождение из одного источника, наличие общего прошлого чаще интерпретируется отрицательно:

(13) *— Тетя Марта! Слышишь? — Слышу, сынок, — отзывается тетя Марта. — Все вы* ***одного поля ягоды*** *— жулики и бездельники!* (Н. Думбадзе. Я, Бабушка, Илико и Илларион)

(14) *По Толстому, Анна, Вронский и общество —* ***одним миром мазаны****. Общество отторгает Анну, лишь следуя формальным правилам поведения, а на самом-то деле поощряет прелюбодеяние и пустые бесовские страсти.* (Корпус Публ.)

Пространственная метафора реализована также в идиоме *две стороны одной [и той же] медали*. Эта идиома означает, что две различных сущности, часто противоположные друг другу, имеют один источник и представляют собой нечто единое; причем это единство описывается как единство двух поверхностей плоского предмета:

(15) — *Шульгин ошибался в своих оценках преимуществ фашизма перед большевизмом* (сегодня-то ясно, что это ***две стороны одной медали***). (Столица)

Если общее происхождение и пространственная близость в текущей ситуации оцениваются по-разному, то наличие общего вербального средства коммуникации осмысляется однозначно положительно, как общность позиций и взглядов:

(16) *Современная литература мне интересна потому, что я хочу узнать тональность сегодняшней жизни. У меня дочь, у меня растет внук, я хочу* ***говорить*** *с ними* ***на одном языке***. (Корпус Публ.)

3.3. Унификация

В данном случае речь идет об операции сведения разнородных элементов в одно множество, причем возможны два варианта: либо происходит игнорирование существенных различий *(мерить на одну мерку; ставить на одну доску; валить в одну кучу)*, что часто означает уравнивание хорошего и плохого, либо разнородные элементы в результате насильственной унификации становятся похожими *(под одну гребенку [стричь…]; привести к одному знаменателю)*:

(17) — *Если вас, Прасковья Федоровна, возможно, и обидел кто-нибудь из мужчин, то это еще не означает, что теперь всех их нужно* ***на одну мерку****, — с мягкой укоризной сказал Истомин.* (А. Калинин. Любя и враждуя)

(18) *Программа партии образца зимы-весны 1993 года воспринимается не иначе как набор лозунгов, без разбору* ***сваленных в одну кучу***. (Корпус Публ.)

(19) *«Весовые категории» различных субъектов РФ столь разнятся, что* ***стричь всех под одну гребенку*** *нельзя.* (Корпус Публ.)

Отдельно следует рассматривать идиому *поставить в один ряд*, которая совместима с контекстами как положительной, так и отрицательной оценки.

ПОСТАВИТЬ В ОДИН РЯД (кого-л./что-л. с кем-л./чем-л.) уподобить кого-л. (что-л.) другим людям (неодушевленным сущностям), объединив

их в одну группу, поскольку и о первых, и о вторых можно думать одинаково, как бы поместив в единую последовательность.

(20) *Партия кричала на него, топала на него сталинскими сапогами: «Если ты проявишь нерешительность, то* ***поставишь*** *себя* ***в один ряд*** *с выродками, и я сотру тебя в порошок!»* (В. Гроссман. Все течет)

(21) *Открытие в Якутии крупного месторождения алмазов* ***поставило*** *СССР* ***в один ряд*** *с крупнейшими производителями этих драгоценных камней в мире.* (Корпус Публ.)

4. Другие операции с множеством

4.1. Пустое множество

Пустота некоторого множества во фразеологии часто выражается через отрицание того, что существует хотя бы один элемент, входящий в это множество. К идиомам, которые строятся по модели «указание на отсутствие хотя бы одного представителя некоторого множества», относятся *не сказать ни единого слова; ни одна/единая живая душа; ни одной/единой [живой] души; ни одна собака; ни единый; без [единой] копейки [денег]; сна ни в одном глазу*. Аналогичным образом один может функционировать и за пределами фразеологии: «Если же имеется в виду *один* как член определенного множества, то отрицательный признак приписывается *всему* множеству (всем членам множества или, иначе, каждому члену множества): *Ни один русский не любит порядка* (Арутюнова 2005: 15).

4.2. Псевдоисчерпание

Идиомы типа *все до единой копейки; все до единого/одного* образуются, согласно (Баранов, Добровольский 2008), по модели псевдоисчерпания, когда в результате перебора всех элементов некоторого множества утверждается, что в этом множестве нет исключений: «Внутренняя форма идиомы *все до одного* основывается на нестандартном выражении идеи исчерпанности некоторого множества по определенному параметру». Модель псевдоисчерпания является одним из вариантов тавтологии в внутренней форме идиом (см. Баранов 2010). Тавтологичность усиливается, если в левом контексте уже присутствует *все* как указание на совпадение рассматриваемого множества с множеством элементов, обладающих некоторым свойством:

(22) *И если хиппи ездят в плацкарте* (3-м классе)*, жрут какую-то дрянь, не имеют денег, то бедные совьетики, которыми набиваются поезда, которые по-скотски ездят в плацкарте, они что,*

тоже хипуют? Тогда у нас в стране все хипуют, ***все до одного,*** *кроме номенклатуры.* (В. Нарбикова. Шепот шума)

4.3. Отождествление множества с одним элементом

К данной разновидности конструкций с компонентом *один/един* относятся идиомы *всё одно; все едино; [все] как один [человек].* В перечисленных идиомах компонент *один/един* употребляется значении ‘целостный, неделимый, единый’ (МАС 1999).

(23) *Здесь Чевенгурский ревком опустил голову* ***как один человек****: из бумаги исходила стихия высшего ума, и чевенгурцы начали изнемогать от него, больше привыкнув к переживанию вместо предварительного соображения.* (А. Платонов. Чевенгур)

Идиома *[все] как один [человек]* может быть истолкована так: ‘у всех людей, изначально образующих некоторое множество, имеется общее для всех них свойство, что описывается как неразличимость разных элементов множества с точки зрения этого свойства’.

Более сложный случай представляет собой идиома *как одна копейка/копеечка.* Во-первых, она может употребляться применительно не только к денежным суммам, но и к промежуткам времени, и к другим сущностям, воспринимаемым как важный и требующий точности ресурс. Во-вторых, она употребляется к контекстах субъективно большого количества (денег, лет, ресурса), но сравнивает его с наименьшей единицей счетного множества. В результате возникают два семантических эффекта в актуальном значении. Первый отмечен в МАС: ‘целиком, полностью (о больших суммах денег)’; иными словами, количество осмысляется как некоторая цельная сущность. Другой эффект — большое количество, будучи единицей *(одной копейкой)*, воспринимается как нечто незначительное, обыденное. Это парадоксальное совмещение двух оценок отмечено в словаре (Лубенская 2004): идиомы *как одна копейка/копеечка и как одну копейку/копеечку* используются «with a phrase denoting a sum of money that is either objectively, or perceived by the speaker to be, large», а в качестве переводных эквивалентов приводятся, среди прочего, «…no less» и «spend as if it were nothing».

(24) *Есть свидетели, что он прокрутил в селе Мокром все эти три тысячи <…> разом,* ***как одну копейку.*** (Ф. Достоевский. Братья Карамазовы)

(25) *Часовой на посту утерял оружие — трибунал и десять лет,* ***как одна копеечка,*** *а для примера могут и еще что-нибудь пострашнее выдать.* (В. Конецкий. Вчерашние заботы)

Модель внутренней формы, при которой имеет место установление отношения между элементом и множеством, реализована также в идиоме *один за всех, все за одного*. Здесь мы видим не отождествление, как в предыдущих примерах, а симметричное отношение, при котором выделенный элемент *(один)* относится к множеству *(за всех)* так же, как множество *(все)* относится к выделенному элементу *(за одного)*.

4.4. Сведение сущности к ее «минимальной» реализации

Так устроены идиомы *одно воспоминание осталось; одно название; смех один; одни кости*. В них нечто описывается как не более чем образ/ярлык/впечатление, связанное с кем-л./чем-л., причем эта форма «призрачного» присутствия осмысляется как отрицание самой сущности:

(26) — *О, Боже,* — *вздохнула Ариадна.* — *От моей талии* ***останется одно воспоминание***. (С. Антонова. Несерьезные размышления о жизни)

(27) *Я вот крещусь, а вообще-то какая я верующая? Только* ***название одно***. *Арсений говорит: ты, говорит, даже смысла литургии не знаешь.* (Е. Козловский. Душный театр)

Идиомы такого типа, как нам представляется, содержат компонент *один* в значении ‘никто другой или ничто другое, кроме; единственный; имеющийся без наличия кого-л., чего-л. другого’ (МАС 1999).

5. Выводы

Во фразеологии как системе реализованы два полюса семантики *один/един*: единство, целостность, полнота, уникальность vs недостаток, неполнота, ущербность, одиночество. В основе семантических групп идиом с компонентом *один/един* лежат разные модели внутренней формы, коррелирующие с разными лексическими значениями слова один и во многом предопределяющие актуальное значение идиом. Модели внутренней формы идиом, содержащих многозначный компонент *один/един*, представляют собой разного рода операции над множеством (сведение множества к элементу, сравнение элементов, исчерпание элементов множества, перебор, указание на пустое множество, указание на наличие определенного свойства у всех элементов множества и т. п.), поэтому идиомы с этим компонентом служат удобным объектом для выработки языка толкований, отсылающего к таким операциям.

Литература

1. *Арутюнова Н. Д.* Проблема числа // Логический анализ языка. Квантификативный аспект языка. М.: Индрик, 2005. С. 5–21.
2. *Баранов А. Н.* Еще раз о факторах идиоматичности: тавтология и онимизация // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 20–24.
3. *Баранов А. Н., Добровольский Д. О.* Внутренняя форма идиом и проблема толкования // Известия АН. Серия литературы и языка, том 57, 1998, № 1. С. 36–44.
4. *Баранов А. Н., Добровольский Д. О.* Аспекты теории фразеологии. Москва, Знак, 2008.
5. *Вознесенская М. М., Киселева К. Л., Козеренко А. Д.* «Тоже мне, бином Ньютона»: операции с числами в составе русских идиом. // Логический анализ языка. Числовой код в разных языках и культурах. Москва, 2013 (в печати).
6. *Лубенская С. И.* Большой русско-английский фразеологический словарь. М.: АСТ Пресс книга, 2004.
7. *МАС* — Словарь русского языка в 4-х томах. Под. ред. А.П. Евгеньевой. М.: Русский язык, 1999.
8. *Степанов Ю. С.* Счет, имена чисел, алфавитные знаки чисел в индоевропейских языках // ВЯ 1989, № 4. С. 46–72.
9. *Топоров В. Н.* Мировое дерево. Универсальные знаковые комплексы. Т. 2, М.: Рукописные памятники древней Руси, 2010.
10. *ФОС* — Баранов А. Н., Вознесенская М. М., Добровольский Д. О., Киселева К. Л., Козеренко А. Д. Фразеологический объяснительный словарь русского языка. М.: Эксмо, 2009.
11. *Dobrovol'skij D., Piirainen E.* Symbole in Sprache und Kultur. Studien zur Phraseologie aus kultursemiotischer Perspektive. // International Journal of Lexicography, Vol. 4, 1998. P. 169–186.

References

1. *Arutjunova N. D.* Number-related issues [Problema chisla], Logicheskij analiz jazyka. Kvantifikativnyj aspekt jazyka, [Logical analysis of language. Quantificational aspect of language], Moscow, 2005, P. 5–12.
2. *Baranov A. N.* Once more on idiomaticity factors: tautology and onimisation [Eshche raz o faktorah idiomatichnosti: tavtologija I onimizatsija], Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"], Bekasovo, 2010, P. 20–24.
3. *Baranov A. N., Dobrovolskij D. O.* (1998), Inner form of idioms and problem of definition [Vnutrennjaja forma idiom I problema tolkovanija], Izvestija AN. Serija literarury i jazyka [News of AS. Literature and language series], Vol. 57, P. 36–44.
4. *Baranov A. N., Dobrovolskij D. O.* (2008) Aspekty teorii fraseologii [Aspects of phraseology theory], Znak, Moscow.
5. *Baranov A. N., Voznesenskaja M. M., Dobrovolskij D. O., Kiseleva K. L., Kozerenko A. D.* (2009) Frazeologicheskii ob"jasnitel'nyi slovar' russkogo iazyka [Explicative phraseological dictionary of Russian language], Eksmo, Moscow.
6. *Dobrovol'skij D., Piirainen E.* (1998) Symbole in Sprache und Kultur. Studien zur Phraseologie aus kultursemiotischer Perspektive, International Journal of Lexicography, Vol. 4, P. 169–186.
7. *Lubensky S. I.* (2004) Bolshoj russko-anglijskij frazeologicheskij slovar [Russian-English Dictionary of Idioms], AST Press kniga, Moscow.
8. *Russian* language dictionary in 4 volumes (1999), Ed. Jevgenjeva A. P., Russkij jazyk, Moscow.
9. *Stepanov Ju. S.* (1989) Counting, names of numerals, alphabetic symbols of numerals in Indo-European languages [Schet, imena chisel, alfavitnyje znaki chisel v indojevropejskih jazykah], Voprosy jazykoznanija [Issues of linguistics], № 4, P. 46–72.
10. *Toporov V. N.* (2010) Mirovoe derevo. Universalnyje znakovyje kompleksy [World tree. Universal sets of symbols], Rukopisnyje pamjatniki drevnej Rusi, Moscow.
11. *Voznesenskaja M. M., Kiseleva K. L., Kozerenko A. D.* "Tozhe mne, binom Njutona": operations with numbers in Russian phraseology ["Tozhe mne, binom Njutona": operatsii s chislami v sostave russkih idiom], Logicheskij analiz jazyka. Chislovoj kod v raznyh jazykah i kulturah, [Logical analysis of language. Numeral code in different languages and cultures], Moscow (in print).

ПОЛИПРЕДИКАТИВНЫЕ КОНСТРУКЦИИ С *ТО ЧТО* В НЕПУБЛИЧНОЙ УСТНОЙ РЕЧИ[1]

Коротаев Н. А. (n_korotaev@hotmail.com)

Российский государственный гуманитарный университет, Москва, Россия

Рассматриваются русские разговорные конструкции, в которых в качестве формального средства связи между предикациями выступает сочетание беспредложной формы *то* и формы *что*, не интерпретируемой как союзное слово. На основании просодических и семантико-синтаксических факторов демонстрируется, что в непубличной устной речи данное сочетание регулярно функционирует как слитный союзный комплекс. В частности, *то что* может использоваться в тех контекстах, в которых употребление расчлененного сочетания местоимения *то* и союза *что* семантически неуместно или даже грамматически недопустимо. Выдвигается предположение о том, что в ряде случае сложный союз *то что* заменяет простое *что*.

Ключевые слова: устная речь, русский язык, корпусный анализ, полипредикативные конструкции, просодия, синтаксис, семантика предикатов, условия употребления, союзы, соотносительные местоимения

CLAUSES-COMBINING WITH *TO CHTO* IN SPOKEN RUSSIAN

Korotaev N. A. (n_korotaev@hotmail.com)

Russian State University for Humanities, Moscow, Russia

In spoken Russian discourse complement clauses introduced by a combination of *to* (originally — a correlative pronoun in nominative or accusative case) and *chto* (complementizer) may exhibit specific features that are not possible in standard written speech. Based on the data from several spoken corpora, the present study claims that *to chto* is regularly used as a compound complementizer. In prosodic terms, *to chto* is often pronounced together with the subordinate clause, while *to*-pronoun usually adheres to the main predicate, a strong intonation boundary appearing between it and the *chto*-clause. In semantic terms, to *chto*-constructions may violate the condition of 'givenness' that presumably licenses the use of the correlative pronoun *to* in standard speech. In syntactic terms, *to chto* may be used with

[1] Работа выполнена при поддержке РГНФ, грант № 12-04-00258.

predicates that require a different case (genitive, instrumental) or a prepositional phrase. Also, coordination of *chto*-clauses and *to chto*-clauses are possible, and *to chto*-clauses appear in contexts with other correlative pronouns in the main clause (like *takoj*).

Key words: spoken discourse, Russian language, corpus linguistics, clause-combining, prosody, syntax, semantics of predicates, conditions of use, complementizers, correlative pronouns

1. Постановка задачи

Форма языковых выражений в существенной степени обусловливается характером коммуникации. Это, в частности, касается и синтаксиса сложного предложения: известно, что в неподготовленном устной речи полипредикативные структуры устроены во многом иначе, чем в кодифицированном письменном дискурсе (см., inter alia, Земская и др. (1981), Miller, Weinert (1998), Thompson (2002)). В настоящей работе рассматривается один из подобных случаев, а именно — особенности поведения сочетания *то что* в непубличной устной речи. Нас будут интересовать полипредикативные конструкции, обладающие следующими свойствами:

(i) связь между клаузами оформляется (в том числе) при помощи последовательности *то что*;
(ii) *то* употреблено в форме, соответствующей местоимению *ТО* в именительном или винительном падеже без предлога;
(iii) *что* не является союзным словом, т. е. не может быть проинтерпретировано как один из актантов предиката зависимой клаузы.

Задача исследования — уточнить природу последовательности *то что* в конструкциях, удовлетворяющих указанным условиям. Материалом работы послужили (а) три корпуса звучащей речи, снабженные просодической разметкой: «Рассказы о сновидениях», «Рассказы сибиряков о жизни» и «Веселые истории из жизни» (в сумме — более 3 с половиной часов звучания, около 30 000 словоупотреблений; корпуса и их описания доступны по ссылке http://spokencorpora.ru); (б) подкорпус устной непубличной речи Национального корпуса русского языка (http://ruscorpora.ru; 1 226 003 слова). Всего в данных источниках было обнаружено немногим более 200 релевантных примеров.

В целях предварительного обсуждения рассмотрим следующие примеры[2]:

---

[2] Все примеры приводятся в исходной нотации: со стандартными знаками препинания или слэшами в НКРЯ, в виде дискурсивной транскрипции в корпусах с просодической разметкой. В последних приподнятые точки соответствуют паузам, слэши — направлению движения тона в акнцентированных словоформах, апострофы — гортанным смычкам. Подробнее о принципах используемой транскрипции см. Кибрик, Подлесская (ред.) (2009). В заголовках примеров из корпусов звучащей речи приводится сокращенное название корпуса (РоС — «Рассказы о сновидениях», ВИ — «Веселые истории из жизни») и номер текста.

(1) PoC, 071n
··(0.4) и когда я= ···(0.6) с-сам /\просыпаюсь,
мне как бы /\кажется,
···(0.6) то что й-я ··(0.1) во /–\снее ещё.

(2) НКРЯ, из коллекции НКРЯ, 2008
Потом ещё… Люди говорят / то что эмо режут вены и много плачут.

(3) ВИ, 38
самое /интересное оказалось –то,
что \действительно в-в /Англии порядка /–семидесяти ·· этих
сервис-центров,,,

(4) НКРЯ, из коллекции Ульяновского университета, 2007
А сейчас его еще бесит то, что мы ни на что не реагируем.

В (3) и (4) представлены стандартные конструкции, в которых *что* является союзом, вводящим подчиненную клаузу, а *то* — соотносительным местоимением, форма которого, в частности, указывает на синтаксическую позицию зависимой предикации в общей структуре сложного предложения[3]. Оформление синтаксической связи между предикациями в обоих этих примерах, как кажется, вполне соответствует литературной норме.

В конструкциях (1) и (2) дело обстоит иначе: в них использованы синтаксические средства, не входящие в канон кодифицированной речи. Именно примеры подобного рода являются основным объектом работы. Мы надеемся показать, что в них сочетание *то что* не состоит из местоимения и союза, как в (3) и (4), а выступает в роли неразложимого союзного комплекса, по ряду свойств сближающегося с простым союзом *что*. Для подтверждения этой точки зрения мы рассмотрим свидетельства интонации и набор семантико-синтаксических факторов.

2. Свидетельства интонации

В примерах (1) и (по всей вероятности) (2) элементы *то* и *что* объединены просодически. Они произносятся без внутренней границы, интонационно примыкая к зависимой клаузе. В транскрипте (1) этот факт отражен в делении на строки; кроме того, предшествующая сочетанию *то что* словоформа *кажется* произносится с пограничным акцентом и за ней следует продолжительная пауза — 0,6 секунды. В записи примера (2) содержится значительно

[3] Отметим, что для настоящего исследования в целом несущественно, о какой именно синтаксической позиции идет речь: о подлежащем (4), прямом дополнении (2), зависимом при прилагательном (3) или при глаголе с экспериенцером в дательном падеже (1) и др.

меньше просодической информации, но на слитное произнесение *то что* предположительно указывает расположение слэша перед, а не после *то*. В (3) и (4), напротив, сочетание *то, что* интонационно расчленено.

Данное различие, как представляется, в известной степени отражает и различие в синтаксической структуре: слитное произнесение свойственно неразложимому сочетанию *то что*, наличие интонационной границы скорее сопутствует синтаксической разъединенности. Отметим, что в корпусах звучащей речи на 8 примеров со слитным *то что* нам встретилось лишь 2 примера с расчлененным *то, что* (один из них представлен в (3)).

Впрочем, в общем случае интонационный критерий не достаточен. Во-первых, он практически неприменим на материале НКРЯ. Доступ к аудиоверсиям текстов, включенных в НКРЯ, отсутствует, а используемая разметка необязательно соответствует фонетической реальности: слэши во многих случаях расставлены лишь для «удобства чтения» — см. Гришина (2005).

Во-вторых, даже достоверно установленной интонационной слитности недостаточно для того, чтобы признать сочетание *то что* неразложимой языковой единицей. Так, в корпусах звучащей речи встречаются примеры, в которых отсутствуют интонационные границы между соотносительным местоимением и союзным словом:

(5) PoC, 024z
···(0.6) ’и /дома ···(0.8) она мне \/пок<u>а</u>зывает,
···(0.8) то что –ум<u>е</u>ет.

По просодическому оформлению пример (5) весьма близок к (1), однако это не влияет на синтаксическую интерпретацию *то* как соотносительного местоимения. В данном случае интонационное единство *то* и *что* обусловлено определенной стратегией распределения коммуникативной нагрузки между составляющими полипредикативной конструкции. Подобная стратегия может быть использована и в той ситуации, когда зависимая клауза с союзным словом соотносится не с местоимением, а с полнозначным именем (подробнее см. Подлесская (2008)):

(6) PoC, 069n
··(0.4) и /увидел из /темнот<u>ы</u>
маленькую тень которая \м<u>о</u>лится.

3. Семантико-синтаксические факторы

Как показано в предыдущем разделе, данных интонации оказывается недостаточно для ответа на основной вопрос: является ли последовательность *то что* в конструкциях (1) и (2) отдельной языковой единицей, выполняющей функцию подчинительного союза, или же в этих примерах, несмотря на их просодические особенности, все же представлено сочетание соотносительного

местоимения *то* с союзом что. В данном разделе, отвлекаясь от интонационного оформления, мы представим ряд других аргументов в поддержку первой гипотезы. Для этого мы проследим, в какой степени обнаруженные примеры следуют известным ограничениям на употребление местоимения *то* в сложноподчиненных конструкциях.

3.1. Контраст

Типичным контекстом появления соотносительного местоимения *то* в сложноподчиненном предложении является ситуация логического подчеркивания, выделения сентенциального актантного зависимого. Природа такого выделения описывается в литературе по-разному. Например, в Шведова (ред.) (1980: §2796) отмечается, что местоимение *то* в подобных контекстах «употребляется как экспрессивное (выделительное) средство»; в Валгина (2000: 108.3) дополнительно конкретизируется, что оно необходимо «при противопоставлении и связанном с ним отрицании»; в Кобозева (2013) указывается на выделенность содержащейся в зависимой клаузе пропозиции в множестве / ряду других событий. Представляется, что в целом набор этих значений достаточно точно описывается понятием *контраста* — см., например, Янко (2001).

В тех случаях когда конструкция имеет контрастивное значение, статус *то* как соотносительного местоимения не вызывает сомнений — см. выше (3) с выделением при помощи *САМЫЙ*, а также следующий пример, в котором реализована конструкция с противопоставлением:

(7) НКРЯ, из коллекции НКРЯ, 2008
Меня так умилила Лин / которую удивило не то / что человек сидит / вывесив ноги наружу из окна / а то / что у человека руки в феньках по локти…

Вместе с тем, в примерах (1) и (2) контраста не ощущается: информация о том, что показалось рассказчику в (1), или о том, какие слова хочет передать говорящая в (2), не получает какого-либо дополнительного выделения.

3.2. Данное vs. новое

Помимо контрастивной семантики, влияние на необходимость / допустимость местоимения *то* в изъяснительных конструкциях оказывает лексическое значение вершинного предиката. В Кобозева (2013) высказывается предположение, что в литературной речи появление *то* в контекстах без контраста связано с противопоставлением «данное vs. новое». Местоимение употребляется в том случае, когда содержащаяся в зависимой клаузе информация дана в предтексте или доступна в общем для участников коммуникации фонде знаний о мире. Когда же в зависимой клаузе содержится новая информация, то в вершинной предикации невозможно — именно этим, по мнению автора,

обусловливается недопустимость соотносительного местоимения в объектных конструкциях с такими глаголами, как *считать, бояться* (в значении «полагать возможным и оценивать эту возможность негативно»), *воображать* (ментальное конструирование) и т. д.

Анализ устных данных, однако, показывает, что во многих конструкциях с *то что* это ограничение нарушается. Так, в примере (1) *то что* употреблено при глаголе *казаться*, в примере (8) — при глаголе *считать*:

(8) НКРЯ, из материалов Санкт-Петербургского университета, 2006
Почему ты считаешь то / что ты помнишь / а я не помню.

Оба этих глагола, согласно Кобозева (2013), входят в число предикатов, допускающих *что*-зависимые и не допускающих *то, что*-зависимых. Эти наблюдения подтверждаются и корпусными методами: предварительный поиск по основному и газетному подкорпусам НКРЯ не выдает интересующих нас контекстов вида *считать то, что*; *казаться то, что*[4]. Таким образом, появление примеров вида (1) и (8) можно считать характерной особенностью неформального устного общения. При этом семантика предикатов, судя по всему, изменений не претерпевает: и в (1), и в (8) зависимая клауза выражает не данную, а новую информацию; меняется лишь синтаксическое поведение сочетания *то что*.

Отдельный класс представляют случаи употребления *то что* при глаголе *сказать*:

(9) НКРЯ, из материалов корпуса «Один речевой день», 2007
А выходить нам… чем от вокзала пешком идти / мы отсюда напрямик пройдем // Я из-за чего тебе и сказал / то что нам… пейсят второй / пейсят шестой / пейсят девятый //

В отличие от предикатов мнения, предикаты речи вполне могут присоединять зависимые, выражающие данную информацию. Именно такой случай и представлен в примере (9): говорящий явно сообщает хорошо известные факты. Как бы то ни было, в литературной речи глагол *сказать* — вне зависимости от значения параметра «данное vs. новое» — не присоединяет сентенциальные дополнения с *то, что*[5].

3.3. Контексты, требующие предлога или формы, отличной от «то»

Как следует из примеров, рассмотренных в предыдущем разделе, в устной непубличной речи допускается употребление сочетания *то что* в контекстах,

---

[4] Речь не идет о конструкциях с «малыми клаузами» (small clauses) вида *Считаем важным (то), что…* Для них, очевидно, действуют иные правила.

[5] Напомним, что мы анализируем только конструкции, в которых *что* выступает в роли союза, а не союзного слова.

в которых в литературной норме использование местоимения *то* перед *что*-придаточным семантически неуместно. Уже одно это, на наш взгляд, является достаточно весомым аргументом в пользу того, чтобы рассматривать в этих случаях *то что* как единый союзный комплекс.

Еще очевиднее отход *то* от его местоименной природы проявляется в примерах вида (10)–(12), в которых употребление местоимения *ТО* в форме именительного или беспредложного винительного падежа не допускается моделью управления вершинного предиката:

(10) НКРЯ, из материалов Саратовского университета, 2004
Он боится то / что я скроюсь и не расплачусь с ним и он будет выплачивать за меня деньги //

(11) PoC, 114n
и \извин<u>и</u>лась перед этой старушкой,
/–ну-у ··(0.1) *попросила /прощ<u>е</u>ния*,
то что я так ··(0.3) \захл<u>о</u>пнула,

(12) НКРЯ, из материалов корпуса «Один речевой день», 2007
Ээ / мы ээ договаривались то что в пятницу с мамой придем…

Прономинальная трактовка *то* в этих примерах более чем затруднительна: в (10) местоимение должно было бы принять форму *того*, в (11) — *за то*, в (12) — *о том*. Такая же ситуация наблюдается и в примере (2) с глаголом говорить, который допускает использование соотносительного местоимения, но не в форме винительного падежа без предлога, а в предложном падеже *(о том)*. В нашем материале зафиксированы также аналогичные примеры с предикатами надеяться, сомневаться, *(приснился) сон, ныть* и др.

В то же время интерпретация *то что* как своего рода эквивалента простого союза что кажется вполне уместной: во всех рассмотренных выше примерах (пожалуй, за исключением (3) и (7)), в том числе и в тех, в которых расчлененное *то, что* в литературной речи недопустимо, использование одиночного что возможно, а зачастую — и желательно.

3.4. Дальнейшее сближение с ЧТО

Приведем еще два факта, свидетельствующих о том, что сочетание *то что* способно функционировать как практически полный аналог союза *что*. Во-первых, нам встретились примеры, в которых клаузы, вводимые *что* и *то что* сочиняются в рамках одного сложноподчиненного предложения:

(13) PoC, 127n
···(0.5) Вот /мн<u>е</u> приснилось,

что я \поступ<u>и</u>л тоже в Академию \–ФСБ вот эту вот,
···(0.5) и то что там с /вокзала чего-то \уезж<u>а</u>ли люди.

Во-вторых, в некоторых случаях употреблению *то что* не препятствует даже наличие отдельного соотносительного местоимения в главной предикации — см. примеры (14) с *такая ситуация* и (15) с *такое дело*:

(14) НКРЯ, из материалов Саратовского университета, 2004
… я те говорила / в прошлом году у меня получилась такая ситуация / то что вот девочка девочка у меня работала и прост-напросто повыдёргивала деньги из кассы и ушла / понимаешь чё / просто ушла //

(15) НКРЯ, из коллекции НКРЯ, 2008
Ну да / бывает такое дело / то что некоторые психи / ну / позеры.

Отметим, что в личном общении нам также встречались примеры вида *думаю о том / то что…; дело в том / то что…* .

4. Выводы и вопросы

В неподготовленной устной речи последовательность *то что* в полипредикативных конструкциях регулярно ведет себя не как сочетание соотносительного местоимения с союзом, а как неразрывная союзная единица. Об этом свидетельствуют интонационные свойства конструкций, а также факты семантико-синтаксической сочетаемости: *то что* встречается, в частности, в тех случаях, когда стандартное расчлененное *то, что* недопустимо грамматически. Парадоксальным образом, более сложная по внутренней форме единица «захватывает» некоторые употребления простого союза *что* — как в контекстах с сентенциальными актантами, так и в некоторых других типах полипредикативных конструкций.

Мы видим следующие перспективы развития настоящей работы.

1) На текущий момент произведен только качественный анализ полученного материала. Между тем только полная разметка корпуса позволить оценить относительную частотность рассмотренных типов.
2) Сфера непубличной устной речи была выбрана для первоначального анализа именно потому, что в ней ожидалось обнаружить набольшее число релевантных случаев. В дальнейшем было бы интересно сравнить частотность выявленного употребления *то что* в корпусах публичной и непубличной речи.
3) Нуждается в уточнении вопрос о том, до какой степени семантико-синтаксические факторы объединения *то что* в сложный союзный комплекс соответствуют интонационным свойствам конструкций. Для разрешения этого вопроса потребуется привлечение большего числа звучащих текстов.

4) Безусловно интересен вопрос о том, насколько рассмотренные употребления *то что* связаны с возрастом, полом и социальными характеристиками говорящих. Уже по имеющимся данным напрашивается предварительный вывод о том, что некоторые говорящие практически не используют *то что* в качестве союза, а некоторые, напротив, делают это на регулярной основе. Было бы любопытно проследить, как в речи информантов второго типа распределены между собой конструкции со слитным *точто*, с расчлененным *то, что* и с простым *что*.

Литература

1. *Валгина Н. С.* (2000) Синтаксис современного русского языка. М.
2. *Гришина Е. А.* (2005) Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. М.: Индрик, с. 94–110.
3. *Земская Е. А., Китайгородская М. В., Ширяев Е. Н.* (1981) Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. М.: Наука.
4. *Кибрик А. А., Подлесская В. И.* (ред.) (2009) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
5. *Кобозева И. М.* (2013) Условия употребления «то» перед придаточным изъяснительным с союзом «что» // Olga Inkova (éd.). Du mot au texte. Études slavo-romanes. Bern: Peter Lang, с. 131–150.
6. *Подлесская В. И.* (2008) Фразовая акцентуация в относительных предложениях: анализ корпусных данных // Фонетика и нефонетика. К 70-летию Сандро. В. Кодзасова. М.: ЯСК, с. 427–445.
7. *Шведова Н. Ю.* (ред.) (1980) Русская грамматика. М.: Наука.
8. *Янко Т. Е.* (2001). Коммуникативные стратегии русской речи. М.: ЯСК.
9. *Miller, J., Weinert, R.* (1998) Spontaneous spoken language: Syntax and discourse. Oxford: Clarendon Press.
10. *Thompson, S. A.* (2002) “Object complements” and conversation // Studies in Language, 26–1, с. 125–164.

References

1. *Grishina E. A.* (2005) Spoken Russian in Russian National Corpus [Ustnaja rech' v Natsional'nom korpuse russkogo jazyka], in Natsional'nyj korpus russkogo jazyka: 2003–2005 [Russian National Corpus: 2003–2005]. Moscow: Indrik.
2. *Janko T. E.* (2001). Kommunikativnye strategii russkoj rechi [Communicative strategies of Russian speech]. Moscow: Languages of Slavonic Culture.
3. *Kibrik A. A., Podlesskaja V. I.* (eds.) (2009) Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Languages of Slavonic Culture.
4. *Kobozeva I. M.* (2013) Conditions on the use of a correlative pronoun "to" before complement clauses with "chto"-complementizer [Uslovja upotreblenija "to" pered pridatochnym iz"jasnitel'nym s sojuzom "chto"], in Olga Inkova (ed) Du mot au texte, Études slavo-romanes [From word to text, Slavo-Romance studies]. Bern: Peter Lang, 131–150.
5. *Miller, J., Weinert, R.* (1998) Spontaneous spoken language: Syntax and discourse. Oxford: Clarendon Press.
6. *Podlesskaja V. I.* (2008) Pitch accent placement in relative-clause sentences: a corpus-based study [Frazovaja aktsentuatsija v otnositel'nyx predlozhenijax: analiz korpusnyx dannyx], in Fonetika i nefonetika. K 70-letiju Sandro V. Kodzasova [Phonetics and non-phonetics. A 70th birthday Festshcrift for Sandro V. Kodzasov]. Moscow: Languages of Slavonic Culture.
7. *Shvedova N. Ju.* (ed.) (1980) Russkaja grammatika [Russian grammar]. Moscow: Nauka.
8. *Thompson, S. A.* (2002) "Object complements" and conversation // Studies in Language, 26–1, 125–164.
9. *Valgina N. S.* (2000) Sintaksis sovremennogo russkogo jazyka [The syntax of contemporary Russian]. Moscow.
10. *Zemskaja E. A., Kitajgorodskaja M. V., Shirjaev E. N.* (1981) Russkaja razgovornaja rech'. Obshchie voporsy. Slovoobrazovanie. Sintaksis [Conversational Russian speech. General issues. Word formation. Syntax]. Moscow: Nauka.

КОМПЕНСАЦИЯ КОММУНИКАТИВНЫХ СТИМУЛОВ В ЭМОЦИОНАЛЬНОМ ДИАЛОГЕ[1]

Котов А. А. (kotov@harpia.ru)

НИЦ «Курчатовский институт», Москва, Россия

Ключевые слова: коммуникативные стимулы, теория вежливости, направление взгляда, невербальная коммуникация

COMPENSATION OF COMMUNICATION STIMULI IN THE EMOTIONAL DIALOGUE

Kotov A. A. (kotov@harpia.ru)

National Research Center “Kurchatov Institute”, Moscow, Russia

An utterance is generated as an expression of an internal communication stimulus. As indicated in the theory of politeness, contradicting tendencies may interfere with the expression of an initial stimulus, in particular an initial face threatening act may be modified by the strategies of negative and positive politeness. Basing on the observations on a multimodal emotional corpus we argue that a certain number of expressive cues in a similar way compensate and modify an initial communication stimulus. (a) A speaker may compensate the changes in gaze direction through gestures, showing iconic gestures when looking aside, and closing gestures when looking at the addressee. We show that “looking aside” is usually combined with addressed gestures (demonstration, iconic gestures). (b) Smiles may also compensate the definitiveness of the main utterance. We show that smiles usually appear in the postposition to an utterance and reduce face threatening in the situations of failure or doubtful proposal — in these cases smiles do not express pleasure and are not connected to jokes.

Keywords: communication stimuli, politeness theory, gaze direction, nonverbal communication

[1] Работа поддержана грантом РФФИ No. 11-06-00301 «Когнитивный анализ семантики слова (компьютерно-корпусный подход)».

1. Введение

Высказывание (и шире — коммуникативное поведение) является результатом внешнего выражения человеком некоторых внутренних коммуникативных стимулов. Процесс выражения коммуникативного стимула можно представлять по-разному: как преобразование одного исходного коммуникативного стимула в речь, либо как конфликт между несколькими противоположными коммуникативными стимулами. На основе данных мультимодального видеокорпуса мы приведём аргументы за то, что коммуникативное поведение является компромиссом между конфликтующими тенденциями: стремлением выразить исходный коммуникативный стимул и противоположным стремлением подавить его выражение.

Среди различных взглядов на процесс построения высказывания в лингвистике, пожалуй, одним из самых известных является положение классической генеративной грамматики о том, что предложение порождается из абстрактного символа *S*, к которому применяются правила развёртывания синтаксической структуры ($S \rightarrow NP + VP$), в результате формирующие предложение на естественном языке. Таким образом, каждый порождённый элемент предложения является «потомком» единственного исходного символа *S*.

Несколько иной взгляд на моделирование поведения используется в биологии и этологии, где поведение живого существа описывается с помощью комбинации противоположных тенденций: приближения и удаления (approach / withdrawal) или четырёх базовых инстинктов (агрессии, бегства, секса и голода). Предполагается, что в поведении животного одновременно проявляются сразу несколько тенденций, и каждый поведенческий акт является компромиссом между разными тенденциями. Например, голодное животное может конкурировать за еду со своим соперником, балансируя при этом между проявлениями агрессии и бегства. При этом степень голода, привлекательность добычи, агрессивность соперника, а также целый ряд других факторов — влияют на выбор животным агрессии или бегства в качестве основной поведенческой стратегии в конкретной ситуации. Конкурирующие тенденции не полностью вытесняют друг друга, их элементы могут чередоваться и объединяться, формируя сложные формы поведения. Таким образом, в противоположность многим лингвистическим теориям, этология описывает поведение животного как результат конкуренции и совмещения элементов противоположных тенденций.

В области искусственного интеллекта для моделирования действий робота аналогичная архитектура конкуренции стимулов была предложена М. Минским [Minsky, 1988] и воплощена в целом ряде проектов по созданию виртуальных компьютерных персонажей, например, в архитектуре CogAff [Sloman, Chrisley, 2003]. Хотя компьютерные архитектуры, аналогичные CogAff, моделируют этологические принципы поведения живых существ, в рамках этих моделей часто отсутствует представление о том, что поведение может объединять противоположные тенденции: считается, что при конкуренции

стратегий поведения победившая стратегия «захватывает» исполнительные механизмы агента, пока она не будет вытеснена другой стратегией. Таким образом, поведенческие элементы двух конкурирующих стратегий чередуются, но не смешиваются.

В области лингвистики примером подхода, объединяющего противоположные тенденции в одном высказывании, является теория вежливости [Brown, Levinson, 1987]. Теория вежливости рассматривает различные высказывания, «угрожающие социальному лицу» адресата или лицу самого говорящего. К таким высказываниям могут относиться просьбы (при просьбе мы «командуем» адресатом, нанося ущерб его социальному лицу) или высказывание мнений адресантом (высказывая спорное мнение, адресант угрожает социальному лицу адресата, а высказывая ошибочное мнение — сам теряет лицо). Для сохранения социального лица высказывание обогащается элементами позитивной вежливости (например, комплимент адресату может смягчить просьбу) или негативной вежливости, которая скрывает угрозу социальному лицу адресата с помощью косвенных речевых актов (*Не могли бы Вы передать мне соль*) или оговорок (*Закрой окно, если можно. / Кажется, он уже ушёл*) [Brown, Levinson, 1987: 145].

Теория вежливости предлагает интересное расширение для архитектур систем автоматического синтеза речи: как оказывается, некоторые элементы высказывания появляются в речи, чтобы компенсировать или подавить выражение основного коммуникативного стимула. Как можно ожидать, ещё более сложная ситуация может наблюдаться в коммуникативном поведении, где жесты и движения глаз могут находиться в сложной взаимосвязи с исходным коммуникативным стимулом и с содержанием высказывания. Вместе с тем, задача синтеза коммуникативного поведения (наряду с синтезом высказывания) оказывается актуальной для лингвистики в связи с необходимостью создания компьютерных персонажей и мобильных роботов, общающихся с пользователем как с помощью речи, так и с помощью жестов, мимики и изменения направления взгляда. Для решения этих прикладных задач архитектура синтеза коммуникативного поведения должна быть расширена, чтобы учитывать не только прямое выражение основного коммуникативного стимула, но и конфликтные тенденции и производные коммуникативные стимулы, проявляющиеся в поведении и речи.

2. Анализ простых форм конфликтов стимулов на основе мультимодального корпуса

Мы анализируем элементы коммуникативного поведения на основе Русскоязычного эмоционального корпуса [Kotov, 2009]. В составе корпуса значительную часть материала составляют видеозаписи устных университетских экзаменов. Часто это достаточно напряжённые эмоциональные ситуации, где можно найти простые примеры коммуникативных колебаний, вызванных противоположными тенденциями.

Наиболее простой случай — это колебания движений рук при передаче предметов адресату. В примере (20080717-c07) в начале экзамена информант передаёт собеседнику (экзаменатору) письменную работу и тестовые карточки, которые использовались в эксперименте. При этом информант совершает поступательные колебания рукой: (а) когда вытягивает руку, чтобы передать работу адресату, (б) когда кладёт работу на стол экзаменатора и возвращает руку, (в) когда повторно протягивает руку, чтобы взять карточки и объяснить их роль в эксперименте.

(а) (б)

Рис. 1. Информант демонстрирует поступательные колебания в жестах, (а) передавая письменную работу экзаменатору, возвращая руку и (б) повторно протягивая руку за работой

В этом примере мы находим сравнительно простой для интерпретации случай: информант должен передать работу экзаменатору, но возникающие сомнения проявляются в жестах. Колебания (обратные движения руки) возникают во время основного жеста и как бы противостоят его выполнению.

Мы хотели бы указать на аналогию между рассматриваемым примером и механизмом оговорок в теории вежливости. Во всех этих случаях действует некоторый механизм ***компенсации исходного коммуникативного стимула***, состоящий в том, что при выражении основного коммуникативного стимула говорящий также будет демонстрировать в речи и поведении элементы, снижающие категоричность исходного коммуникативного стимула и противостоящие его прямому выражению.

3. Компенсация при изменении направления взгляда

Направление взгляда — это внешний параметр поведения, который может управляться самыми разными внутренними стимулами. С одной стороны, глаза служат для восприятия, поэтому говорящий может направлять взгляд на адресата или на обсуждаемый объект. С другой стороны, направление взгляда позволяет выражать эмоции, причём некоторые эмоциональные знаки требуют прямого зрительного контакта — как строгий или жалостливый взгляд, а другие эмоции заставляют отводить взгляд — как это происходит под

действием смущения или задумчивости. Наконец, отводя взгляд, говорящий может демонстрировать особые коммуникативные знаки, ориентированные на адресата — стрелять глазами, закатывать глаза и т. д.

В корпусе мы наблюдаем множество примеров того, как информанты во время рассуждений при ответе на вопрос уходят от прямого зрительного контакта с адресатом — они обращают свой взгляд вбок или вверх, при этом часто продолжая отвечать на вопрос. В ситуации экзамена потеря коммуникативного контакта с адресатом — достаточно опасная стратегия, поэтому человек, сдающий экзамен, старается компенсировать потерю зрительного контакта другими средствами, например, жестами. Рассмотрим следующий пример, где подряд наблюдаются две противоположных компенсаторных тенденции.

(а)

— Иллокуция — это речевой акт, рассматриваемый с точки зрения его внеязыковой цели. А иллокутивная сила — это характеристика вот такого акта.
Заканчивая ответ, информант смотрит вбок, при этом активно жестикулируя левой рукой по направлению к адресату.

(б)

Закончив ответ, информант переводит глаза на адресата, при этом натягивая ворот свитера на нижнюю часть лица.

Рис. 2. Компенсация жестами изменения направления взгляда (изображение намеренно искажено)

Как мы видим из этого примера, действия руками последовательно компенсируют коммуникативный эффект от изменения направления взгляда. В первом фрагменте информант уводит взгляд, что, по-видимому, вызвано умственным напряжением — попыткой вспомнить и сформулировать ответ на вопрос. Потеря зрительного контакта может рассматриваться как реакция избегания (withdraw), поэтому для сохранения контакта с экзаменатором информант активно жестикулирует левой рукой, демонстрируя противоположную тенденцию приближения (approach). Во втором фрагменте, после завершения ответа, информант прямо смотрит на экзаменатора, чтобы оценить был ли

принят его ответ. Такое действие может рассматриваться как слишком явное приближение (approach) и опять компенсируется действиями рук: информант прячет нижнюю часть лица в ворот свитера, тем самым демонстрируя реакцию избегания (withdraw).

Таким образом, в указанном примере мы можем рассматривать действия рук как средство компенсации коммуникативных стимулов, управляющих направлением взгляда. Чтобы выйти за рамки единичного примера, посмотрим, как связано изменение направление взгляда с жестами рук в составе всего корпуса.

Съёмка материала корпуса произведена обычной видеокамерой, что не позволяет разметить все изменения направления взгляда. Однако в корпусе размечаются случаи, когда информант переводит взгляд вбок или вверх на время около 1 секунды или более. Рассмотрим все действия рук, которые информанты выполняют одновременно с переводом взгляда вбок.[2]

Таблица 1. Жесты рук, выполняемые одновременно со взглядом вбок (с точностью 700 мс)

Действие, выполняемое глазами	**Действие, выполняемое руками**	**Число случаев**
Взгляд вбок	манипулирует	26
	иконический жест	24
	демонстрирует	14
	указывает	9
	подпирает	6
	контакт	5
	другое	3
	скрещивает	3
	закрывает	2
	трёт	2
	чешет	2
	стучит	1
Всего		97

Всего в корпусе присутствует 2866 тэгов «взгляд вбок», при этом в 97 случаях их начало совпадает с началом какого-либо действия руками. Как видно из таблицы, наиболее часто перевод взгляда вбок комбинируется с жестами, направленными на адресата. Прежде всего это относится к тэгам «иконический жест» и «демонстрирует» (так в корпусе размечаются случаи, когда информант показывает адресату какой-либо предмет или открывает ладони в направлении адресата).

---

[2] Одновременность в данном случае означает, что разница между моментами начала этих двух действий не превышает 700 миллисекунд.

Таким образом, если изменение направления взгляда наступает одновременно с каким-либо действием руками (одновременность здесь является косвенным подтверждением общности коммуникативного стимула), то в большинстве случаев это связано с действием механизма компенсации: либо демонстрация жеста заставляет адресата уходить от прямого зрительного контакта, либо наоборот, необходимость отвести взгляд заставляет компенсировать это жестом, направленным на адресата. И в том и в другом случае компенсация связана с желанием избежать слишком сильного приближения (approach).

4. Сглаживание категоричности

Механизмы компенсации можно обнаружить и в мимическом оформлении высказывания. Человек может демонстрировать различные мимические паттерны в разные моменты своего коммуникативного поведения: (а) в качестве реакции на слова собеседника во время его высказывания или сразу после него, (б) перед началом произнесения собственного высказывания, (в) в середине своего высказывания, при обнаружении ошибок, при автокоррекции, в паузах хезитации, а также (г) по окончании собственного высказывания.

Среди различных мимических действий традиционно наибольшее внимание было уделено исследованиям улыбки и смеха, причём выделение коммуникативных функций смеха вызывало большие споры. Общепринятый взгляд состоит в том, что смех является реакцией на остроту, следовательно, смех должен следовать за высказыванием собеседника. Противоположный взгляд на функции смеха состоит в том, что смех является реакций на снятие собственного напряжения. Так, З. Фрейд в своей классической работе [Freud, 1905] указывает, что смех проявляется в результате выражения в остро́те подавленных сексуальных или агрессивных тенденций адресанта. Прямое выражение этих тенденций запрещено моралью, но острота представляет возможность выразить эти желания в социально приемлемой форме, а снятие напряжения высвобождает энергию удовольствия. Таким образом, согласно Фрейду, реакция смеха или улыбки должна, скорее, сопровождать высказывание говорящего, а не отвечать на высказывание собеседника.

Ещё одной функцией смеха может быть снижение агрессивности и категоричности высказывания [Kotov, 2008]. На такую функцию человеческого смеха указывают этологи, которые подчёркивают, что гомолог человеческого смеха у некоторых видов обезьян маркирует игровое поведение или является знаком подчинения, направленным на снижение агрессии адресата [Butovskaja, 2004: 77]. Если смех у человека обладает сходной функцией, то он также должен скорее следовать за словами говорящего, а не за словами собеседника.

Смех блокирует возможность говорить, поэтому в качестве элемента исследования мы выбрали улыбку — она может свободно пересекаться как с собственными словами говорящего, так и со словами адресата. Мы зафиксировали

подкорпус, в котором присутствуют 144 улыбки.[3] После этого мы обратили внимание на моменты начала и конца улыбок: нас интересовало совпадение этих моментов с речью говорящего или с речью адресата.

Говорящий начинает улыбаться на чужих словах в 18 случаях (2,54% высказываний собеседника заканчиваются улыбкой адресанта), тогда как на своих словах адресант начинает улыбаться намного чаще — в 54 случаях (улыбкой заканчиваются 3,7% высказываний говорящего). Для момента окончания улыбки различие оказывается несущественным.

Таблица 2. Соответствия моментов начала и конца улыбки словам адресанта и словам собеседника (на подкорпусе рассмотрены случаи 144 улыбок)

	…на чужих словах*	**…на своих словах****
Адресант начинает улыбаться…	**18** (12,5% всех улыбок, 2,54% высказываний)	**54** (37,5% всех улыбок, 3,7% высказываний)
Адресант заканчивает улыбаться…	**28** (3,9% высказываний)	**45** (3,14% высказываний)
	* всего в избранном подкорпусе 707 высказываний собеседника	** всего в избранном подкорпусе 1435 высказываний говорящего

Эти данные подтверждают гипотезу о том, что коммуникативный стимул улыбки чаще связан с собственными словами, а не со словами собеседника. Таким образом, алгоритм генерации коммуникативного поведения для компьютерного агента должен порождать улыбку в составе общего механизма порождения высказывания. Рассмотрим теперь конкретные примеры улыбок в конце собственной фразы. Мы могли бы ожидать примеров, когда говорящий улыбается в конце собственной шутки, чтобы подтолкнуть к улыбке адресата. Однако в корпусе мы видим противоположные примеры: говорящий обычно улыбается в конце высказываний, где он неуверенно отвечает на вопрос или где он обращает к собеседнику сомнительную просьбу:

(1) **Преподаватель** (запутывает студента)*: Смотрите, это обозначение, сколько здесь зависимых переменных, независимых и побочных, так?*
Студент: *Да.*
П: *Похоже на правду?*
С: *Похоже. Да.*
П: *Похоже*[4]*. Нет, не похоже, это не то.*
С: *Почему не похоже?* (улыбка/смех)

---

[3] В подкорпус были выделены видеофрагменты, в которых полностью размечены все высказывания участников коммуникации.

[4] Эхолалическое повторение.

П: *Ну, потому что здесь не идёт вообще никакого указания на то, сколько здесь зависимых и побочных* [*переменных*]. (20080717-c16)

(2) **Студент:** *Не проще ли просто поставить* [*мне зачёт*]*? Чтобы больше таких бездарей, как я, не видеть на пересдаче* (улыбка/смех) (20081229-a3)

Конечно, материал корпуса обладает спецификой — это не общение в расслабленной дружеской атмосфере, стимулирующей шутки, а взаимодействие в напряжённой ситуации, где можно ожидать стратегий сглаживания конфликта и противоречий. Вместе с тем, полученные данные подтверждают гипотезу о том, что улыбка (в одной из своих функций) регулярно появляется в конце собственного высказывания и служит для снижения его категоричности. В терминах теории вежливости это ситуации потери социального лица говорящим (когда говорящий демонстрирует неуверенный ответ или признаёт свою ошибку) и ситуации атаки на социальное лицо адресата (когда говорящий обращает к нему некорректную просьбу).

Таким образом, улыбка также является внешним проявлением механизма компенсации, стремящегося скорректировать исходный коммуникативный стимул высказывания.

5. Подходы к моделированию коммуникативного поведения

Если синтаксическая структура высказывания может быть построена с помощью правил развёртывания из одной отправной точки, то прагматика высказывания (и, как следствие, его семантика) является компромиссом между конкурирующими тенденциями. Вне рамок нашего рассмотрения остался случай, при котором коммуникативное поведение конструируется за счёт нескольких не связанных друг с другом стимулов. Однако даже если рассмотреть более простой случай, при котором всё поведение является результатом выражения одного коммуникативного стимула, то можно увидеть тенденции, которые возникают при попытке выражения этого стимула и компенсируют его проявления.

Эти механизмы возникают на самых разных уровнях коммуникативного поведения:

(a) как известно из теории вежливости, попытка выразить действие, угрожающее лицу адресата, приводит к появлению в высказывании элементов позитивной или негативной вежливости, которые снижают категоричность исходного стимула — то есть выражение исходного стимула в речи приводит к появлению компенсирующих элементов в речи;

(б) необходимость высказать сомнительное суждение или сомнительная просьба, адресованная адресату, заставляют говорящего компенсировать категоричность высказывания, например, с помощью улыбки в постпозиции к высказыванию — то есть выражение исходного стимула в речи приводит к появлению компенсирующих элементов в мимике и жестах;

(в) изменение направления взгляда заставляет адресанта компенсировать с помощью жестов слишком явный прямой зрительный контакт или, наоборот, уход от зрительного контакта — то есть выражение исходного стимула с помощью одних коммуникативных действий приводит к компенсации за счёт других коммуникативных действий.

Механизм синтеза коммуникативного поведения из двух (или более) коммуникативных стимулов, конечно, усложняет теоретическую и прикладную компьютерную модели. Вместе с тем, по аналогии с теорий вежливости количество компенсаторных действий, появившееся в поведении, может использоваться для выражения степени сомнения или волнения компьютерного персонажа, а тип используемых компенсаторных действий может добавлять персонажу индивидуальности.

Литература

1. *Brown P., Levinson S. C.* (1987) Politeness: Some Universals in Language Usage (Studies in Interactional Sociolinguistics), Cambridge.
2. *Butovskaja M. L.* (2004) Body language: nature and culture [Jazyk tela: priroda i kultura], Scientific World, Moscow.
3. *Freud S.* (1905) Der Witz und seine Beziehung zum Unbewußten, Franz Deuticke, Leipzig — Wien.
4. *Kotov A. A.* (2008) Functions of laughter in a dialogue: another view at the classic problem [Funktsii smeha v dialoge: eshchë odin vzgljad na klassicheskuju problemu], Human being in past and present: behavior and morphology [Chelovek v proshlom i nastojashchem: povedenie i morfologija], Institute of Ethnology and Anthropology RAS, pp. 31–48.
5. *Kotov A. A.* (2009) Patterns of emotional reactions in communication: problems of corpora studies and application to computer agents [Patterny emotsional'nyh kommunikativnyh reaktsij: problemy sozdanija corpusa i perenos na kompjuternyh agentov], Computer linguistics and intellectual technologies, Issue 8 (15), RSUH, Moscow, pp. 211–218.
6. *Minsky M. L.* (1988) The Society of Mind, Touchstone Book, New-York, London.
7. *Sloman A., Chrisley R.* (2003) Virtual Machines and Consciousness, Journal of Consciousness Studies, vol. 10, No 4–5, pp. 133–172.

ТЕЛО И ЕГО ЧАСТИ В РАЗНЫХ ЯЗЫКАХ И КУЛЬТУРАХ (ИТОГИ НАУЧНОГО ПРОЕКТА)

Крейдлин Г. Е. (gekr@iitp.ru),
Переверзева С. И. (P_Sveta@hotmail.com)

Институт лингвистики РГГУ, Москва, Россия

В статье излагаются результаты работы над научным проектом, целью которого является построение знаковых репрезентаций тела и телесности в разных языках (английском, арабском (египетский диалект), литовском, немецком и хинди) и соответствующих языках жестов. Даётся характеристика телесных объектов и их содержательных объединений, описываются основные признаки телесных объектов и значения этих признаков наряду с их языковыми именами. Указываются также способы их языкового отображения и жесты с участием тех или иных частей тела — прежде всего те, которые выражают отношения между людьми. В заключение статьи подводятся итоги проведённых исследований, относящихся к прикладной невербальной семиотике.

Ключевые слова: тело, телесный (соматический) объект, семиотическая концептуализация, сопоставительный анализ, язык, жест, признак, значение признака, прикладная невербальная семиотика

HUMAN BODY AND ITS PARTS IN DIFFERENT LANGUAGES AND CULTURES (THE RESULTS OF THE SCIENTIFIC PROJECT)

Krejdlin G. E. (gekr@iitp.ru),
Pereverzeva S. I. (P_Sveta@hotmail.com)

Russian State University for the Humanities, Moscow, Russia

The paper presents the main results of a project aimed at constructing semiotic representations of human body and corporality in different natural languages (English, Arabic (the Egyptian dialect), Lithuanian, German and Hindi) and the corresponding body languages. The lexical system of a body language consists of gestures (in a broad sense of the word), i.e. gestures proper (manual gestures, gestures of legs, etc.), postures, meaningful glances, touches and some other semiotic classes of units. The primary

directions of the project are (1) to describe somatic objects and their significant combinations; (2) to describe major classes of these objects, such as the human body itself, body parts, bones, biological liquids; (3) to examine the features of these objects and their values as well as those of their names; (4) to exhibit different kinds of gestures with somatic objects, among them those expressing human relationships. We also focus on some results in the field of applied nonverbal semiotics, i.e. (a) description of Russian symptomatic gestures performed by a patient in a conversation with a doctor. These gestures may serve to characterize a patient's disease; (b) semantic analysis of Russian phraseological units with names of somatic objects; (c) exploration of meaning and functional characteristics of the so-called Bible somatisms — linguistic expressions in the Bible texts with names of somatic objects as well as of the gestures; (d) analysis of theatrical corporeal behavior.

Key words: body, corporal (somatic) object, semiotic conceptualization, comparative analysis, language, gesture, feature, value, applied nonverbal semiotics

1. Введение

Проект «Тело и его части в разных языках и культурах», о котором пойдёт речь в настоящей работе, является естественным продолжением проекта «Части тела в русском языке и русской культуре», о котором нам уже приходилось рассказывать и писать.[1] Разработанные в нём формат и метаязык описания тела и телесных объектов послужили отправной точкой для предпринятого нового исследования.

Основным понятием, общим для обоих проектов, было понятие **семиотической концептуализации тела и телесности**. Отражая типовые взгляды обычных, не искушённых в биологии или медицине людей на человеческое тело, семиотическая концептуализация моделирует их представления о теле и телесности, выраженные в знаках данного языка и культуры. «Строительным материалом» для семиотической концептуализации являются обслуживающие её знаки не только естественного языка, но и телесных кодов, или языков тела, а также, возможно, единицы некоторых других знаковых кодов. При построении моделей тела и телесности для языков и культур, отобранных нами, главное внимание мы уделили значениям и употреблениям релевантных слов и словосочетаний, а также значениям и употреблениям жестов разных семиотических классов и лексическим особенностям их номинаций.

В качестве материала для анализа были взяты следующие языки: английский, арабский (египетский диалект), литовский, немецкий и хинди — и соответствующие им языки тела. Изучалось то, как тело представлено в вербальном

[1] См. Крейдлин 2010; Крейдлин, Переверзева 2010а. Участниками обоих проектов были студенты (А. С. Сидорова, А. В. Семёнова, Э. Е. Заришева), аспиранты (Е. А. Клыгина, Л. А. Хесед) и преподаватели РГГУ (Г. Е. Крейдлин, А. Г. Кадыкова, С. И. Переверзева) и ВШЭ (А. Б. Летучий), а также научный сотрудник Института славяноведения (П. М. Аркадьев).

и невербальных знаковых кодах этих языков и культур, и сравнивались выразительные возможности этих кодов.

Решение поставленных в проекте задач было разбито на несколько этапов.

1. С самого начала был составлен список основных телесных, или, иначе, соматических, объектов (далее: СО) и их типовых естественно-языковых имён. Затем объекты и имена были распределены по классам, таким, как тело (отдельный класс), части тела (голова, рука), части частей тела (ноздри, пальцы), внутренние органы (печень, почки, желудок), особые места на/в человеческом теле (подмышки, пупок, ложбинка), линии (пояс, талия), инородные объекты (горб, прыщ, синяк) и др.
2. Составленные для каждого языка списки подверглись дальнейшей обработке. В частности, имена СО объединялись в синонимические ряды, и каждый элемент ряда был снабжён пометами, характеризующими его сферу употребления или отдельные смысловые или стилистические особенности слова.

Материалами и источниками данных для предполагаемого сопоставительного анализа знаковых систем послужили печатные тексты и электронные корпуса. Учитывались данные разных словарей, включая жестовые; использовались видеоматериалы записанных нами академических лекций преподавателей РГГУ разных гуманитарных профессий. Видеозапись, разумеется, сочеталась с визуальным наблюдением за реально исполняемыми жестами; при их анализе мы опирались не только на наши собственные наблюдения, но и на видеоматериал, собранный по нашей просьбе некоторыми коллегами и студентами российских и зарубежных вузов.

2. Результаты исследования

Работа над проектом велась по нескольким линиям. Это (I) характеристика самих СО, их типов и знаковых репрезентаций в разных языках; (II) описание телесных признаков и их значений, а также способов их вербального и невербального знакового отображения; (III) анализ средств невербального отображения отношений между людьми (для каждого из рассматриваемых языков тела); (IV) исследования в области прикладных аспектов невербальной семиотики.

2.1. Характеристика СО, их типов и знаковых репрезентаций в разных языках

В ходе анализа СО подробно были описаны некоторые их классы, или типы, — в частности, кости (на материале русского и английского языков). Показано, что класс «кости» разбивается на два крупных семантических подкласса,[2] один из которых с языковой точки зрения состоит из элементов, являющихся

[2] См. Крейдлин 2013.

костями, — поскольку люди говорят о них именно как о костях, костных объединениях и сочленениях. Это *кость*, *скелет*, *ребро*, *хрящ*, *ключица* и некоторые другие. Второй подкласс с языковой точки зрения представляет собой, прежде всего, части тела или части таких частей; иными словами, в большинстве коммуникативных актов люди рассматривают имена таких СО как обозначения частей тела или частей таких частей. Между тем центральным, если не единственным, формо- и структурообразующим элементом в этих СО является кость, и по этой причине такие слова, как *локоть*, *лоб*, *колено*, *таз*, *скула* и некоторые другие, интерпретируются в ряде контекстов как кости, то есть имеют другие значения или другой тип употребления. Таким образом, имена костей обладают разветвлённой структурой многозначности и множественной референцией. Понятие множественной референции в дальнейшем было перенесено и на другие типы СО.

Параллельно с русским материалом изучалось то, как класс «кости» представлен в английском языке. Например, слово *bones* обозначает не только ‘множество костей’, но и ‘скелет’; кроме того — по-видимому, из-за несоответствия эталонов красоты в англосаксонской и русской культуре, — оно означает также ‘черты лица’, ср. *beautiful bones of face* ‘красивые (благородные) черты лица’. И наконец, слово *bones* часто используется иронически, ср. *She will take care of your bones* ‘Она позаботится о вас’ (букв. ‘Она позаботится о ваших костях’); такая же ирония прослеживается и в сложных словах с частью *bone*.

Помимо типов СО, анализировались и отдельные СО, такие, как щёки, пупок, переносица, волосы, глаза и ряд других.

При построении семиотической концептуализации щёк были описаны их структурные, физические и функциональные признаки и проанализированы их языковые отображения.[3] Особое внимание обращалось на смысловые и культурные выделенные значения (*values*) этих признаков, которые имеют свои выражения в каждом из исследуемых языков.

Под смысловой выделенностью значения признака мы понимаем ту дополнительную смысловую нагрузку, которую несёт языковая единица, служащая средством выражения этого значения, — обычно это информация об обладателе объекта с этим признаком. Например, высказывание *У него кривые руки* характеризует человека как плохо делающего что-то или не умеющего делать нечто нужное или полезное; таким образом, сочетание *кривые руки* является языковым выражением выделенного значения признака «форма рук».[4] Под культурной выделенностью значения мы имеем в виду его важность для правильного осмысления культурного контекста употребления самого признака.

---

[3] См. Крейдлин, Летучий 2010.

[4] Смысловая выделенность значения признака «форма рук», которое передаётся сочетанием *кривые руки*, — это не коннотация слова *кривой*. У этого слова вообще нет коннотации, если, конечно, не считать, что у слова *кривой* в сочетании *кривые руки* есть особое значение. Ср. выражения *кривые ноги* и *кривые зубы*, передающие значения (*values*) признаков, соответственно, «форма ног» и «форма зубов». Указанные значения не обладают свойством смысловой выделенности.

Так, значение /иметь бороду/ для признака «наличие волосяного покрытия на лице» (у мужского лица) ещё в недавнем прошлом было культурно выделенным, поскольку бороду носили лишь определённые социальные группы мужчин: религиозные деятели, художники, писатели.

Выделенными являются и некоторые значения цвета и формы щёк, ср. *розовые щёки*, *белые щёки*, *впалые щёки*, поскольку эти сочетания говорят не только о цвете и форме щёк, но и о физическом здоровье или нездоровье человека, а также о некоторых свойствах его тела.

Подробно описывались также семантика и синтаксис трёх важных классов жестов с участием щёк, а именно жестов-ударов (хлопнуть/ударить по щеке, похлопать по щеке или пощёчина), жестов-касаний (трепать по щеке или поцеловать в щёку), и жестов-щипков (ущипнуть за щёку). Показано, что среди ударов жесты занимают особое место, выражая разные эмоции, как негативные (ср. ударить по щеке и пощёчина), так и позитивные (ср. похлопать по щеке). Разумеется, не любой удар по щеке является знаковым, например, бьют по щеке (щекам) в драке или приводя людей в чувство.

Интересными оказались некоторые характеристики СО «пупок» и его признаков: (а) имя *пупок* обладает множественной референцией, обозначая не только часть живота, но и определённое место на нём; (б) существует важный параллелизм в семантике слов *пупок* и *пуп*, ср. значение слова *пупок* как ‘центрального места на человеческом теле’ и слова *пуп* как ‘центра Земли’. Кроме того, интерес представляют (в) культурная символизация пупка и его дисфункции. Так, пупок является маркером генетической связи материнского тела с телом ребёнка; кроме того, пупок мыслится как средоточие жизненных сил человека, а такая операция над ним, как завязывание, ассоциируется с началом, «завязыванием» жизни человека. Кроме того, завязывание пупка служит жизненно важной цели, а именно препятствует проникновению болезней в тело человека, то есть пупок символически представляется как одно из тех мест (наряду с глазами, коленями, сгибом локтя и некоторыми другими), через которые в организм человека проникают разные болезни.

Информация о дисфункции пупка, прежде всего о его болезнях, важна для адекватного перевода с одного языка на другой и для построения типологии соответствующих имён в разных языках. В Интернете есть много текстов, в которых люди (в основном матери младенцев) интересуются, как лечить такую болезнь, как *мокнущий пупок* (выражение *мокнущий пупок* означает ‘плохое заживление пупочной ранки, когда она постоянно мокнет’). Название данной болезни по-английски звучит, однако, как *weeping navel*, то есть буквально ‘плачущий пупок’, — если этого не знать, то перевести на английский сочетание *мокнущий пупок* весьма затруднительно.

Мы изучили и такой «экзотический» СО, как «переносица»: сформулировали толкование слова *переносица* — ‘твёрдая часть носа между глазами, расположенная ближе всех других частей носа ко лбу’, — в которое включены две важные характеристики переносицы. Это физический признак (точнее, значение этого признака «твёрдый») и структурный признак — «местоположение». При описании переносицы выделена важная — и, по-видимому, свойственная

многим культурам, — характеристика, соответствующая её основной функции, а именно, переносица является местом, на котором люди носят очки. Таким образом, переносица входит в класс типовых мест на теле или в теле человека, за которыми закреплены строго определённые функции; обнаружено, что таких мест не очень много.

При изучении значения и употребления ряда русских жестов, таких, как, например, жестовых ударений, а также слова *переносица* и его иноязычных эквивалентов мы столкнулись с необходимостью ответить на ряд вопросов, относящихся к области словообразовательной семантики. В частности, это вопросы о значениях приставки *пере-* и правилах её комбинирования с корнем *-нос-* и — шире — о связи значений разных приставок с русскими именами телесных объектов: приставки *за-* в словах *запястье*, *залысина*, *затылок*, приставки *под-* в словах *подбородок*, *подмышка*, *поджелудочная <железа>* и приставки *пред-* в *предплечье*, *предсердие*. Показано, что значение приставки *пере-* в слове *переносица* соответствует смысловому компоненту ‘между глазами’, то есть переносица мыслится как некоторое место/часть носа, расположенное между глазами. Показательны в этой связи английские и литовские эквиваленты переносицы, соответственно, *bridge of nose* (букв. ‘мост носа’, то есть мост, тянущийся через нос от одного глаза к другому) и *tarpuakis* (букв. ‘межглазье’). Иное представление о переносице отражено в её французском имени — *radix de nez*, букв. ‘корень носа’; оно показывает, что переносица мыслится как основание носа — место, от которого нос строится. Таким образом, во французском языке важно не горизонтальное измерение переносицы, а вертикальное. Идея вертикальности отображается и в другом литовском обозначении переносицы — слове *viršunosė*, обозначающем буквально ‘верхнюю часть носа’.

Анализ семантики словообразовательных, в частности, приставочных, морфем в составе имён СО послужил нам стимулом для продолжения изучения нетривиальных связей такой области лингвистики, как приставочное словообразование, с кинесикой — наукой о жестах, жестовых процессах и жестовых системах и составной частью невербальной семиотики[5].

2.2. Описание телесных признаков, их значений и способов их вербального и невербального знакового отображения в разных языках

Второе направление наших исследований было связано с анализом признаков СО и их значений. Анализируя признаки объектов мира, мы считаем принципиально важным отталкиваться всегда от той предметной области, к которой данные объекты принадлежат (в нашем случае это СО). Иными словами, мы изучаем не признак «цвет» как таковой и не признак «форма» и т. д., а признаки «цвет волос», «форма носа» и т. д. Выбранная для анализа предметная область диктует как сам набор признаков, так и их возможные, прежде всего, выделенные, значения.

---

[5] См. об этом в работе Крейдлин 2012.

Все признаки СО были разделены на четыре больших группы: (а) классификационные (такие, как, например, объединение СО в типы или иные группы, ср. сочетания *опорно-двигательная система*, *вестибулярный аппарат*), (б) структурные, (в) физические и (г) функциональные.

К структурным признакам относятся, например, признаки «биологическая парность» и «семиотическая парность».

Были выделены два типа биологической парности — перцептивная и когнитивная. Первая отличает видимые части тела, такие, как плечи, руки и ноги; об их парности человек узнаёт из личного опыта и наблюдений. Когнитивные биологические пары образуют СО, парность которых для человека неочевидна: он узнаёт о ней в ходе обучения в школе или при болезни СО. К таким СО относятся, в частности, лёгкие и почки.

У слов, референты которых входят в перцептивные пары, числовая характеристика устроена несколько иначе, чем у слов, референты которых образуют когнитивные пары. Мы одинаково легко и с более или менее одинаковой частотой употребляем слова руки и рука, ноздри и ноздря, ухо и уши между тем как употребления слов типа лёгкие и почки, то есть слов, обозначающих когнитивные биологические пары, встречаются на порядок чаще, чем употребления слов лёгкое и почка[6].

СО, образующие биологическую пару (не важно, является эта пара перцептивной или когнитивной), являются семиотически парными, если: (1) стандартное языковое обозначение самой пары имеет форму множественного числа (ср. руки, ноги, глаза); (2) каждый член пары имеет имя, которое может быть либо (а) формой единственного числа, соответствующую имени пары (ср. рука, нога, глаз), либо (б) сочетанием такой формы со словом, выражающим местоположение обозначаемого объекта относительно вертикальной (ср. ноги — левая и правая нога, глаза — левый глаз и правый глаз) или — реже — относительно горизонтальной оси тела (ср. губы — верхняя и нижняя губа, веки — верхнее и нижнее веко); (3) и у данного СО, и у биологически парного ему СО имеются имена, хорошо освоенные данным языком и не являющиеся терминами. Таковы, например, сочетания левая рука и правая рука.[7]

Производным от понятий семиотической парности и семиотической пары является понятия **семантически** и **культурно выделенного члена семиотической пары** (для данного языка и данной культуры). Например, в русском языке левая рука выделена в семиотически парном объекте «руки» относительно смысла ‘лёгкость выполнения некоторого действия’. Об этом говорит, например, оборот сделать что-л. одной левой (при неправильном *сделать что-то одной правой/только правой). В мусульманских культурах левая рука является культурно выделенной: она считается нечистой, ей, например,

---

[6] Так, согласно НКРЯ (Основной корпус со снятой омонимией, 28.03.2013) слово *лёгкое* встречается примерно в 4 раза реже, чем слово *лёгкие* (21/86), тогда как *рука* и *руки* употребляются примерно с одинаковой частотой (4156/4588).

[7] Подробно о понятиях биологической и семиотической парности, а также о биологических, языковых и культурных причинах, обуславливающих выделенность того или иного члена пары см., например, в работе Крейдлин, Переверзева 2010б.

не принято давать или брать подарки. В русской же культуре за левой рукой такие культурные коннотации не закреплены.

При анализе СО и повседневного телесного поведения человека были сформулированы два важных принципа, регулирующих такое поведение, — **принципы физиологического удобства** и **пространственного удобства**. Оба они говорят о том, какой СО человек обычно выбирает в качестве инструмента действия. Первый гласит: если человек хочет совершить некоторое действие и в его распоряжении есть несколько СО, которые он может использовать в качестве инструментов для этого действия, то он выбирает тот СО, который является физиологически наиболее удобным. Например, если человек исполняет жест **закрыть глаза руками**, он закрывает правой рукой правый глаз, а левой рукой — левый; иначе ему неудобно. Второй принцип говорит о том, что если человек намеревается совершить некое действие с предметом, расположенным в его личном пространстве, и использовать для этого один из релевантных СО, то он выбирает тот СО, который в данный момент находится ближе всего к данному предмету.

Различие членов пары по выделенности и формулировка данных принципов важны не только для теоретической, но и для практической лексикографии, так как они позволяют сократить тексты лексикографических описаний.

Из физических признаков были подробно рассмотрены признаки «звуки, <издаваемые> СО», «цвет СО» и некоторые другие. На множестве телесных звуков была определена серия важных противопоставлений, образующих основу для семантических классификаций. Среди них: (1) звук, издаваемый данным СО / звук, для которого он является вместилищем; (2) уникальный / неуникальный телесный звук; (3) телесный звук, издаваемый человеком в нормальном / изменённом состоянии; (4) звук, в норме слышимый только данным человеком / слышимый и другими людьми; (5) реальный / имажинальный (то есть воображаемый, представляемый) телесный звук и ряд других.[8] Рассматривались также противопоставления на множестве лексических единиц, связанных с телесными звуками.

При анализе признака «цвет СО» (на материале русского, английского, немецкого, французского и итальянского языков) было показано, что этот признак играет важную роль не только для характеристики СО, но и для описания внешнего облика, некоторых личных качеств и физического или эмоционального состояния человека. Отметим здесь наиболее важные результаты:

(1) для разных языков СО чувствительны к признаку «цвет» по-разному. Для одних СО он является конституирующим, то есть без указания значения (или нескольких значений) этого признака описание СО, такого, как, например, глаза, волосы, кожа, будет неполным и неточным, Для других (например, для лица, рук или губ) указание на цвет нужно только в определенных коммуникативных или социальных ситуациях. Для третьих (подбородок, поясница или ступни) цвет в текстах не указывается вовсе или указывается исключительно редко;

---

[8] Полный список и анализ вводимых противопоставлений на множестве звуков см. в статье Крейдлин, Переверзева 2011.

(2) даже в пределах одной западноевропейской культуры языковые отображения выделенных значений признака «цвет данного СО» различаются. Так, в английском языке одно из выделенных значений признака «цвет глаз» передается словом *pink* ‘розовый’: при этом сочетание *pink eyes* характеризует человека, больного конъюнктивитом. В русском языке сочетание розовые глаза для описания человека применяется очень редко;

(3) обнаружена не свойственная другим исследуемым европейским языкам особенность многих немецких композитов, которая состоит в том, что они, будучи по морфологической структуре обозначениями цвета СО (например, *Himbeerzunge* букв. ‘малиновый язык’, где *Himbeer* — ‘малина’, а *Zunge* — ‘язык’), семантически более насыщенны. Так, в приведенном примере слово *Himbeerzunge* обозначает не только цвет, но и особенность поверхности языка — его бугристую структуру, напоминающую структуру малины; иными словами, оно означает следующее: ‘язык, по своей бугристой поверхности и цвету похожий на малину’.

Перейдём теперь к особенностям концептуализации тела и его частей в рассматриваемых нами языках.

При анализе особенностей арабской семиотической концептуализации тела были получены (среди прочего) следующие результаты:

(1) составлен список типовых арабских имён основных СО и показано, что они по многим признакам отличаются от русских аналогов;

(2) выявлено существенное различие в употреблении арабских имён СО, а также их признаков и значений признаков, в составе свободных сочетаний и в составе фразеологизмов; арабские единицы сравнивались при этом с соответствующими единицами русского языка. Например, устойчивое сочетание *ghariqa hatta: al-adqa:n* (букв. ‘погрузиться по подбородки’) соответствует русскому обороту со значением ‘увязнуть по уши’, а фразеологизм *raghma ’anfihi* ‘против чьего-л. желания’ (букв. ‘против чьего-л. носа’) не имеет никаких соответствий в русской соматической фразеологии;

(3) показано, что арабские названия СО нередко образуются от корней с широким значением, изначально не связанным с тематическим полем тела и телесности, а потому они могут передавать информацию о других фрагментах наивной картины мира. Например, слово *Zahrun* ‘спина’ означает также ‘заднюю сторону какого-либо объекта вообще’, в частности, ‘оборотную сторону листа’, тогда как корень этого слова, *Zhr*, переводится как ‘представать, появляться’, то есть со смыслом ‘спина’ не связан. Отметим, что смысл ‘представать, появляться’ в русской культуре и русской картине мира соотносится скорее с другим СО, а именно лицом (а не со спиной, как в арабском языке).

В ходе построения фрагмента литовской семиотической концептуализации тела и телесности было показано, что в литовском языке не находят отражения смысловые противопоставления, отмеченные для русских слов *тело*, *корпус*

и *плоть*.[9] Зато в литовском языке есть разные слова, которые соответствуют двум русским лексемам *плечи* — *плечо1* (‘плоские горизонтальные части тела по обе стороны от шеи’, лит. *petis*) и *плечо2* (‘часть руки от плеча1 до локтя’, *žastas*).

Подробно изучались разнообразные и часто не характерные для русского метонимические соотношения между отдельными значениями имён СО (например, слово *plaukai* обозначает не только волосы на голове и на теле человека, но и шерсть животных и её масть). Литовскому языку присущ также целый ряд фразеологизмов, не свойственных русскому языку. Например, выражение *visokio plauko*, буквально означающее (с учётом семантики падежной формы) ‘разных волос’ или ‘разных мастей’, переводится на русский как *всевозможные*, *разношёрстные*. При описании литовских обозначений волос обнаружено, что, хотя волосы и в литовском, и в русском относятся к типу телесных покровов, признаки волос в этих языках отличаются характером выделенных значений.

Разнообразие способов представлений тела в литовской и русской картинах мира исследовалось, в частности, на материале особого рода текстов — лечебных заговоров. Литовские лечебные заговоры о теле и телесности, как правило, строятся по особым формулам, которые можно назвать *формулами телесности*. Это, по большей части, ритуализованные, обычно клишированные, выражения с именами СО, обращённые к высшей силе или каким-то предметам, наделяемым магической силой, и служащие средством избавления людей от разных заболеваний. Обнаружено, что разные СО «освоены» такими текстами в разной степени. Максимальная освоенность у СО «кровь»: литовцы говорят, что *кровь бежит*, *течёт*, *выманивает душу из тела*, её просят остановиться. Кровь не должна *видеть солнца* или *показываться свету*; её нужно *завязать*, *запереть*, *удерживать в жилах*. Интересно, что среди литовских эпитетов крови, встречающихся в лечебных заговорах, отсутствует привычная для русских характеристика *красная* — кровь бывает только ‘огненная’ (*ugniavystas*) и ‘прозрачная’ (*skaidrus*).

Помимо анализа литовских выражений и текстов изучалась языковая концептуализация СО «голова», «сердце» и «язык», а именно, описывались семантически и культурно выделенные значения признаков этих СО. Так, было показано, что в современном литовском названии языка — слове *liežuvis* (от *liežti* ‘лизать’), в отличие от старолитовского названия, нет переносных значений ‘язык’ и ‘речь’: они в современном литовском передаются при помощи основы *kalba*, ср. глагол *kalbėti* ‘говорить’.

2.3. Анализ способов и средств невербального отображения отношений между людьми (для данного языка тела и культуры)

Третье направление включает в себя исследования отдельных способов невербального и смешанного, вербально-невербального, отображения межличностных отношений. Изучались семантические классы дружеских и любовных

[9] Семантический анализ этих слов см. в работах Крейдлин 2010, Крейдлин 2014 (в печати).

жестов русской и англо-саксонской культуры, инвариантом которых является выражение хороших чувств жестикулирующего к адресату. Речь идёт, в частности, об описании физической реализации, семантики и закономерностей употребления таких русских жестов и классов жестов, как **объятия**, **поцелуи**, дружеские **шлепки** и **похлопывания**. Для их характеристики важны, прежде всего, такие различительные признаки, как форма, манера и место реализации жеста, степень психологической близости субъекта и адресата, сила, с которой жест исполняется, и ряд других.

Изучались смысловые различия между языковыми конструкциями, описывающими такие жесты или способы их исполнения. В качестве примера укажем на различия между сочетаниями *обнять чьи-то плечи* и *обнять кого-то за плечи*: в семантике предложного сочетания, по сравнению с беспредложным, содержится дополнительная информация: это идея захвата, присвоения и удерживания адресата жеста при себе в течение краткого времени.

Из любовных жестов-**поцелуев** особенно подробно описаны жестовые лексемы, объединённых названием *поцелуй в нос*.[10] Кроме того, были описаны такие классы жестов, как **шлепки** и дружеские **похлопывания**. Было введено важное понятие телесной этики, которое помогает описанию многих телесных знаков, выражающих этикетные смыслы, такие, как ‘прилично’, ‘неприлично’ и т. п., а также культурных сценариев, отражающих типовые русские этикетные нормы телесного поведения.

2.4. Исследование прикладных аспектов невербальной семиотики

Четвёртое направление исследования образуют работы, связанные с прикладными аспектами невербальной семиотики. Их общая цель — раскрыть важные и вместе с тем неочевидные связи между невербальной семиотикой и другими областями знаний.

Среди них: (а) выполненное на материале одного конкретного коммуникативного жанра — диалога врача с пациентом во время первичного приема больных описание русских симптоматических жестов, служащих важными и надёжными, как считают медики, показателями патологических состояний людей. Такое исследование, по мнению врачей, может существенно помочь начинающим медикам при постановке диагноза; (б) разработка методики анализа русских фразеологических единиц с именами СО (так называемых фразеологических соматизмов). Была сформулирована гипотеза, говорящая о том, что для полного и точного описания семантики таких фразеологизмов необходимо знание свойств многих составляющих семиотической концептуализации тела и телесности. Эта гипотеза была подтверждена при описании русских фразеологизмов со словом *голова*[11],

---

[10] О жестах-**поцелуях**, в частности описание жеста **поцелуй в нос**, см. в работе Крейдлин, Переверзева 2013 (в печати).

[11] Полный список и анализ вводимых противопоставлений на множестве звуков см. в статье Крейдлин, Переверзева 2011.

а также фразеологизмов языка хинди со словом *зубы* и их переводных эквивалентов в целом ряде языков; (в) раскрытие смысловых и функциональных особенностей библейских соматизмов, то есть выражений естественных языков, включая имена жестов, встречающихся в библейских текстах. Определена роль тела и других СО (главным образом, глаз, рук, коленей) в семиотическом акте молитвы и описан класс основных религиозных жестов — жесты **коленопреклонения**; (г) анализ сценического телесного поведения. Речь идёт о невербальных единицах и моделях, свойственных некоторым известным в театральном мире системам и направлениям режиссуры. В частности, было введено понятие невербального театра и показано, в чём состоят отличительные особенности ряда невербальных театров в их сравнении с некоторыми другими видами сценических направлений.

3. Заключение

Подытоживая сказанное, подчеркнём ещё раз, что важность и актуальность построения семиотической концептуализации тела и телесности в разных языках и культурах состоит в том, что оно (а) открывает возможности построения разных типологий тела и телесности и показать те точки, в которых эти типологии дополняют друг друга или противоречат друг другу; (б) уточняет и дополняет набор признаков и их значений, характеризующих тело и различные явления телесности. Кроме того, описание тела и телесности в разных языках и культурах открывает перспективы (в) создания моделей устной коммуникации людей, в которой существенно используются и взаимодействуют вербальные и невербальные знаковые коды, и (г) усовершенствования общих стратегий и инструментов перевода устных и письменных текстов с одного языка на другой с учётом содержащейся в них вербальной и невербальной информации. Этим и смежным с ними проблемам мы собираемся посвятить нашу коллективную монографию под условным названием «Образ тела в языке и культуре».

Литература

1. *Козеренко, Крейдлин* 2011 — Козеренко А. Д., Крейдлин Г. Е. Фразеологические соматизмы и семиотическая концептуализация тела // Вопросы языкознания, № 6, 2011. С. 54–66.
2. *Крейдлин* 2010 — Крейдлин Г. Е. Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 1: тело и другие соматические объекты // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). М.: Изд-во РГГУ, 2010. Вып. 9 (16). С. 226–234.
3. *Крейдлин* 2012 — Крейдлин Г. Е. Невербальная семиотика и словообразование: точки соприкосновения // Актуальные проблемы словообразования. IV. Кемерово: Кемеровский ГУ, 2012.

4. *Крейдлин* 2013 — Крейдлин Г. Е. Соматические объекты и некоторые их типы: проблемы лингво-семиотического описания // Материалы Гротовских чтений. Сборник научных статей. М., 2013 (в печати).
5. *Крейдлин* 2014 — Крейдлин Г. Е. Библейские соматизмы: квалификация и оценка // Хвала и хула в языке и коммуникации (сборник статей). М.: РГГУ, 2014 (в печати).
6. *Крейдлин, Летучий* 2010 — Крейдлин Г. Е., Летучий А. Б. Части тела в русском языке и в невербальных семиотических кодах. II. Щёки // Русский язык в научном освещении, № 1 (19), 2010. С. 222–235.
7. *Крейдлин, Переверзева* 2010а — Крейдлин Г. Е., Переверзева С. И. Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 2: Признаки соматических объектов и их значения // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.)). М., 2010. Вып. 9 (16). С. 235–240.
8. *Крейдлин, Переверзева* 2010б — Крейдлин Г. Е., Переверзева С. И. Части тела и их имена в русском языке: биологическая и семиотическая парность частей тела // II Международная конференция «Русский язык и литература в международном образовательном пространстве: современное состояние и перспективы», Гранада, 8–10 сентября 2010 г. Том II. Доклады и сообщения. Гранада, 2010. С. 2064–2069.
9. *Крейдлин, Переверзева* 2011 — Крейдлин Г. Е., Переверзева С. И. Основные противопоставления на множестве телесных звуков // Вестник РГГУ. 2011. № 11 (73). С. 80–101.
10. *Крейдлин, Переверзева* 2013 — Крейдлин Г. Е., Переверзева С. И. Дружеские и любовные жесты. II. Поцелуи // Лингвистика для всех. Летние лингвистические школы 2009, 2010 и 2011. М.: МЦНМО, 2013 (в печати).

References

1. *Kozerenko A. D., Krejdlin G. E.* (2011), Phraseological somatisms and semiotic conceptualization of the body [Frazeologicheskije somatizmy i semioticheskaja kontseptualizatsija tela], Voprosy jazykoznanija [Questions of Linguistics], no. 6, pp. 54–66.
2. *Krejdlin G. E.* Body in a dialog: semiotic conceptualization of the body (results of the project). Part 1: Body and other somatic objects [Telo v dialoge: semioticheskaja kontseptualizatsija tela (itogi projekta). Chast' 1: telo i drugije somaticheskija ob'jekty]. Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Moscow, 2010, pp. 226–234.
3. *Krejdlin G. E.* Bridges between nonverbal semiotics and the theory of word-formation [Neverbal'naja semiotika i slovoobrazovanije: tochki soprikosnovenija]. Aktual'nyje problem slovoobrazovanija. IV [Actual problems of the theory of word-formation. IV]. Kemerovo, 2012.

4. *Krejdlin G. E.* Somatic objects and come of their types: the problems of linguo-semiotic description [Somaticheskije ob'jekty i nekotoryje ih tipy: problem lingvo-semioticheskogo opisanija]. materialy Grotovskih chtenij Sbornik nauchnyh statej [Materials of Grot conference]. Moscow, 2013, to appear.
5. *Krejdlin G. E.* Bible somatisms: qualification and assessment [Biblejskije somatizmy: kvalifikatsija i otsenka]. Hvala i hula v jazyke i kommunikatsii [Praise and disgrace in language and communication]. Moscow, 2014, to appear.
6. *Kreydlin G. E., Letuchiy A. B.* (2010), Body parts in the Russian language and in nonverbal semiotic codes. II. Cheeks [Chasti tela v russkom jazyke i v neverbal'nyh semioticheskih kodah. II. Shcheki], Russkii iazyk v nauchnom osveshchenii [Russian language in scientific coverage], no.1, pp. 222–235.
7. *Kreydlin G. E., Pereverzeva S. I.* Body in a dialog: semiotic conceptualization of the body (results of the project). Part 2: Features of somatic objects and their values [Telo v dialoge: semioticheskaja kontseptualizatsija tela (itogi projekta). Chast' 2: Priznaki somaticheskih ob'jektov ii h znachenija]. Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Moscow, 2010, pp. 235–240.
8. *Kreydlin G. E., Pereverzeva S. I.* Body parts and their names in Russian: biological and semiotic pairs of body parts [Chasti tela i ih imena v russkim jazyke: biologicheskaja i semioticheskaja parnost' chastej tela]. II mezhdunarodnaja konferentsija "Russkij jazyk i literature v mezhdunarodnom obrazovatel'nom prostranstve: sovremennoje sostojanije i perspektivy" [II International conference "Russian language and literature in the International Education: modern state and perspectives"]. Granada, 2010, pp. 2064–2069.
9. *Kreydlin G. E., Pereverzeva S. I.* (2011), Corporeal sounds: basic oppositions [Osnovnyje protivopostavlenija na mnozhestve telesnyh zvukov], Vestnik RGGU [RGGU Bulletin], no. 11(73), pp. 80–101.
10. *Kreydlin G. E., Pereverzeva S. I.* Gesture of friendship and love. II. Kisses [Druzheskije i l'ubovnyje zhesty. II. Potselui]. Lingvistika dl'a vseh. Letnije lingvisticheskije shkoly 2009, 2010 i 2011 [Linguistics for everybody. Summer Schools of Linguistics in 2009, 2010 and 2011]. Moscow, 2013, to appear.

СЕМАНТИЧЕСКИЕ МЕХАНИЗМЫ ФОРМИРОВАНИЯ АДВЕРБИАЛЬНЫХ ВЫРАЖЕНИЙ НА БАЗЕ ОТГЛАГОЛЬНЫХ СУЩЕСТВИТЕЛЬНЫХ[1]

Кустова Г. И. (galinak03@gmail.com)

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

В статье рассматривается репрезентативный класс абстрактных существительных — отглагольные имена (номинализации) — на одном из этапов процесса грамматикализации, а именно — на этапе превращения в адвербиальные выражения. Номинализации, вообще говоря, наследуют свойства глаголов (в частности, набор актантов), но для адвербиалов, сформированных на базе номинализаций, существуют системные ограничения на выражение валентностей исходной номинализации: у одного типа адвербиалов субъект системно не выражается (при этом он совпадает с субъектом главной предикации): *Пассажиры ходили по перрону **в ожидании** поезда* (**в ожидании поезда пассажирами*); у другого типа адвербиалов субъект обычно (хотя и не обязательно) выражается (при этом он не совпадает с субъектом главной предикации): *Приехал **по приглашению дяди***; некоторые номинализации образуют оба типа адвербиалов (с разными предлогами): *при выборе товара* vs. *по выбору клиента*. Таким образом, адвербиалы имеют более ограниченные возможности по сравнению с исходными номинализациями, поскольку у номинализаций в принципе выразимы все глагольные валентности: *ожидание поезда пассажирами = кто ожидает чего*; *приглашение племянника дядей = кто пригласил кого* и т. д. В этом отношении адвербиалы аналогичны деепричастным и причастным формам глаголов, у которых тоже имеются системные запреты (или системные требования) относительно выражения тех или иных валентностей.

Ключевые слова: номинализация, адвербиал, валентность, грамматикализация

ADVERBIAL EXPRESSIONS BASED ON VERBAL NOUNS

Kustova G. I. (galinak03@gmail.com)

V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

[1] Работа выполнена при поддержке РГНФ, проект № 11-04-00223а и № 11-04-00488а.

The paper discusses a stage of abstract noun grammaticalization — namely, transformation into adverbial expressions, cf. *v ozhidanii* 'waiting', *pod okhranoj* 'under protection', *po priglasheniju* 'by invitation', *v blagodarnost'* 'in gratitude'. Two types of such adverbials are distinguished: 1) the agent of the adverbial is not expressed (*Passazhiry khodili po perronu* ***v ozhidanii*** *poezda* 'The passengers were strolling along the platform waiting for the train'); 2) the agent of the adverbial is necessarily expressed (*Prijekhal* ***po priglasheniju djadi*** 'came by invitation of his uncle'). In contrast to adverbials, nominalizations can express all arguments.

Key words: nominalization, adverbial, valence, grammaticalization

В статье рассматриваются обстоятельственные выражения типа *в надежде что сделать; под охраной кого; по приглашению кого* и т. п. (а также близкие к ним производные предлоги типа *под давлением кого / чего* и под.), сформировавшиеся на базе отглагольных имен (номинализаций). Номинализации интересны тем, что они имеют двойственную природу. С одной стороны, они сохраняют свойства глаголов, с другой — подвергаются характерной для падежных и предложно-падежных форм существительных адвербиализации. Обстоятельственные выражения, которым посвящена данная работа, отражают один из промежуточных этапов процесса грамматикализации отглагольных имен, который (процесс) заключается в постепенном превращении «обычных» существительных — через стадию адвербиалов — в производные предлоги. Для обозначения интересующих нас обстоятельственных выражений мы будем использовать термин **адвербиальные дериваты** (сокращенно — АД); этот термин используется, в частности, в работах И. М. Богуславского [Богуславский 2003; 2008] и удобен тем, что акцентирует и обстоятельственную природу выражений типа в ожидании, по просьбе и под., и их производный характер (тем самым — связь с глагольной ситуацией).

Отглагольное существительное наследует многие важные свойства глагола и может выражать значения глагольных категорий — например, номинализация может отражать аспектуальные признаки глагола — длительность, завершенность и т. п.: *В ожидании поезда* — ждал (НСВ); *по просьбе X-а* — попросил (СВ). Этой проблематике посвящена большая литература, ср., например, [Падучева 1977; 2009; 2010]; [Пазельская 2008]; [Alexiadou 2001]; [Comrie 1980]; [Comrie & Thompson 1985]; [Filip 1999]; [Moltmann 2005]; [Nichols 1988]; [Rappaport 1992]; [Rozwadowska 1988; 1997; 2000]. В [Пазельская 2008], в частности, показано, что на объем и характер унаследованных от глагола свойств существительного влияет множество факторов — тип исходного глагола, количество и тип валентностей, суффикс, позиция в предложении и др.

Нас в данной работе будут интересовать валентные свойства номинализаций. Одной из главных проблем в этой области (ей посвящена целая серия работ Е. В. Падучевой) является проблема способов и возможностей выражения глагольных (унаследованных) валентностей.

В [Падучева 2010] исследуется, какими формами — родительным, творительным, притяжательным местоимением *(исполнение концерта Рахманиновым; исполнение Рахманинова; ваше исполнение концерта)* — выражаются валентности субъекта и объекта при отглагольном имени и каковы корреляции между этими способами выражения глагольных актантов. При этом Е. В. Падучева не выделяет адвербиальные дериваты как особую группу номинализаций, а рассматривает их вместе с другими номинализациями, т. е. случаи типа *охотник продолжал преследование волка* и *концерт в исполнении Рахманинова* рассматриваются в одном ряду.

И. М. Богуславский рассматривает выражения типа *в подарок*, *по привычке*, по приказу как адвербиальные дериваты исходных предикатных имен (и соответствующих глаголов), что является продолжением и развитием подхода, принятого в теории лингвистических моделей «Смысл ⇔ Текст» (см. [Мельчук 1974/1999]; [Апресян 1974]; согласно теории «Смысл ⇔ Текст», выражения типа *с победой* или *под арестом* должны включаться в словарную статью отглагольного существительного как разновидность лексической функции Adv, ср. [Мельчук 1974/1999: 115; 123]). В [Богуславский 2008] рассматривается вопрос, как заполняются пассивные валентности таких единиц, т. е. валентности, которые не выражены синтаксически зависимыми формами. Например, адресат адвербиала *по приказу (кто приказал **кому**)* в предложении может совпадать с подлежащим при активном главном глаголе: *По приказу командующего **войска** перешли в наступление*, — но может и не совпадать с подлежащим, если глагол употреблен в пассивной форме: *По приказу Шуйского была построена плотина* (кем-то, кому приказали); при этом «адвербиал *по привычке* жестко требует для своего актанта («чья привычка» — Г. К.) позиции подлежащего: ***Переселенцы** и на новом месте по привычке строили дома у реки* vs. ??*Дома и на новом месте по привычке строились (переселенцами) у реки*» (примеры из [Богуславский 2008: 125]). И. М. Богуславский приходит к выводу, что способы выражения пассивных валентностей адвербиалов нельзя «вычислить» по исходному глаголу, поэтому их надо указывать в словаре, в специальной словарной статье соответствующего адвербиала.

В данной работе мы рассмотрим другой аспект функционирования адвербиальных дериватов — а именно, системные требования к синтаксическому выражению их валентностей, т. е. попытаемся ответить на вопрос: когда синтаксическое выражение валентностей АД системно предусмотрено, а когда оно системно ограничено (запрещено)? При этом АД будут рассматриваться не просто как номинализации, но именно как редуцированные формы — и по отношению к глаголу, и по отношению к «обычной» номинализации, — формы, синтаксические свойства которых зависят не только от состава унаследованных от исходного глагола валентностей, но и от функции и значения самого адвербиала в предложении.

Мы будем исходить из того, что в парадигму слов некоторой части речи могут входить как ядерные, так и «редуцированные», периферийные формы. Ядерные формы выполняют основную синтаксическую функцию данной части речи, периферийные — функции, свойственные другим частям речи. Так,

ядерной частью глагольной парадигмы в русском языке являются предикативные (финитные) формы. В периферийную часть глагольной парадигмы входят нефинитные формы — деепричастия и причастия, — которые наследуют ряд важных признаков глагола (хотя при этом не могут выполнять основную его функцию — главного сказуемого; более сложная ситуация с инфинитивом, но его мы здесь не рассматриваем), но в то же время обнаруживают признаки других частей речи — прилагательных и наречий (авторы, которые считают эти неглагольные признаки доминирующими, даже рассматривают причастия и деепричастия не как формы глагола, а как особые части речи, ср. [Тихонов 1987]).

У существительных тоже бывают периферийные формы — адвербиальные дериваты (ср. *сбоку*, *на ходу*), — однако они рассматриваются не как часть именной парадигмы (подобно причастиям и деепричастиям в глагольной парадигме), а как утратившие связь с исходным существительным. Для «обычных» существительных такой подход можно считать правомерным, однако с отглагольными существительными ситуация другая. Поскольку номинализации коррелируют с глаголами, в их парадигме также можно усматривать два рода форм — ядерные и периферийные: к ядерным относятся управляемые падежные формы, которые в предложении являются подлежащими и дополнениями, к периферийным — адвербиальные дериваты. При этом парадигмы глаголов и отглагольных существительных обычно не сравниваются, т. к. их ядерная часть включает формы совершенно разной природы — предикативные (финитные) у глагола и управляемые у имени. Однако их вполне можно сравнить в периферийной части — на уровне адвербиальных дериватов глаголов и номинализаций.

Прежде чем перейти к рассмотрению АД, кратко остановимся на свойствах «обычных» — управляемых — номинализаций.

Примечание. Строго говоря, при обсуждении способов и возможностей выражения глагольных валентностей номинализации нужно делить не на два класса (управляемые формы и адвербиальные дериваты), а на три — необходимо учесть еще функционирование номинализаций в составе аналитических предикатов типа *оказать помощь*, *обратиться с просьбой* и под., т. е. в контексте лексико-функциональных глаголов (лексических функций в терминологии модели «Смысл ⇔ Текст»). Варианты поведения номинализаций в контексте полувспомогательных глаголов (ср. *оказывать влияние* vs. *находиться под влиянием / испытывать влияние / подвергаться влиянию* vs. *имеет место влияние*), в частности, правила выражения или не-выражения тех или иных валентностей (например, в составе глагольно-именного оборота ***X обратился с просьбой*** *о P* субъект *просьбы* X синтаксически не выражается, т. к. семантически совпадает с подлежащим лексико-функционального глагола *X обратился*) подробно описаны в работах представителей модели «Смысл ⇔ Текст» ([Мельчук 1974/1999]; [Апресян 1974: 45–48]). Мы присоединяем эти случаи к управляемым формам и будем исходить из противопоставления управляемых номинализаций и адвербиалов.

У управляемых номинализаций в принципе выразимы все глагольные валентности: *ожидание поезда встречающими; присутствие понятых при обыске; проверка документов аудиторами; участие прокуроров в рассмотрении дел судами; выбор товара покупателем; наблюдение инструктора за выполнением упражнений; охрана посольства автоматчиками; просьба студентов к преподавателю о переносе лекции* и т. п. Однако такая идеальная картина наблюдается только в словаре (в изолированном состоянии). В конкретных предложениях ситуация с выражением валентностей у номинализаций может быть разной.

Если у исходного глагола три валентности (как, например, у глаголов речи — субъект, адресат и содержание), у номинализаций обычно выражается две валентности, а не три — либо субъект и содержание *(Просьба* ***студентов о переносе*** *лекции осталась без ответа; Откликнулся на просьбу* ***студентов о переносе*** *лекции; Отказал студентам в* ***их*** *просьбе* ***о переносе*** *лекции)*, либо адресат и содержание (*Неоднократные просьбы к администрации Пушкинского района и милиции пресечь незаконную уличную торговлю остались неуслышанными* [«Неприкосновенный запас», 2010]; *Второе — просьба* ***к Президенту о*** *скорейшей* ***встрече*** *с представителями рыбаков* [«Рыбак Приморья», 2003.01.09]).

Но даже если у исходного глагола две валентности, они тоже не всегда выразимы в полном объеме: иногда нельзя выразить объект, иногда — субъект. В таких случаях Е. В. Падучева говорит о полной и редуцированной диатезе отглагольного имени. Под редуцированной диатезой понимается невозможность выражения при существительном какого-л. актанта в силу «позиции в предложении» [Падучева 2010: 25], например:

(а) *Картина изображает преследование* ***оленя волком*** (полная диатеза);

(б) *Волк продолжал преследование* ***оленя*** (*оленя* — объект; бессубъектная редуцированная диатеза, ср. **Волк продолжал преследование оленя волком*);

(в) *Олень спасался от преследования волка* (*волка* — субъект; безобъектная редуцированная диатеза) (мы используем примеры Е. В. Падучевой в несколько измененном виде).

Адвербиальный дериват включается в предложение не в качестве управляемой формы, а в качестве свободного обстоятельства (детерминанта): *приехал* ***по просьбе*** *друга;* ***по просьбе*** *друга он взял удочки и надувную лодку;* иногда адвербиальные дериваты могут выступать также в качестве присубстантивного члена: *пикет* ***в защиту*** *заключенных; курс* ***по выбору***.

Адвербиальные дериваты — это своего рода застывшие формы, устойчивые обороты, особые конструкции, которые имеют определенное (обычно обстоятельственное) значение (цель, причина, время, условие и под.), являющееся результатом взаимодействия значения предлога и собственной семантики отглагольного имени, ср. *при желании, по совету, в благодарность за, под влиянием* и под.

Адвербиальный дериват выражает редуцированную («свернутую») предикацию, которая является добавочным, дополнительным сообщением по отношению к основной предикации, выраженной главным глаголом, и вступает в определенные смысловые отношения с главной предикацией. В этом смысле АД аналогичны деепричастным и причастным оборотам, а также обстоятельственным придаточным, т. е. пропозициям в адвербиальной функции.

АД занимают промежуточное положение между «полноценными» номинализациями, у которых, в принципе, выразимы все актанты, и производными предлогами, у которых нередко происходит коренная перестройка актантной структуры и значения исходного существительного

Примечание. Ср., например, абстрактное существительное *причина* и производный предлог *по причине*. Выражение *причина пожара* предполагает, что ‘у пожара есть причина’, а выражение *по причине пожара* значит, что ‘пожар и есть причина’, т. е. у производного предлога по сравнению с исходным существительным семантика и синтактика существенно перестраивается, и родительный падеж — не та валентность, что у существительного *причина*, а само выражение *по причине пожара* превращается в аналог «обычной» предложной группы, например — *из-за пожара*. АД не настолько отличаются от управляемых отглагольных существительных, но тем не менее они теряют способность выражать полный набор актантов исходного существительного (ср. идущую от Л. В. Щербы идею разграничения собственно служебных и строевых слов, развиваемую в работах Г. Е. Крейдлина, ср. [Крейдлин 1982]).

Отличие АД от управляемых номинализаций с точки зрения возможности выражения актантов состоит в следующем. Поскольку неизвестно, с каким глаголом будет связано управляемое отглагольное имя, мы не можем заранее сказать, какие глагольные валентности сможет выразить номинализация в том или ином конкретном предложении. У адвербиальных дериватов, наоборот, заранее известно, какая валентность может быть выражена — или не может быть выражена — зависимыми падежными формами, т. к. у них СИСТЕМНО зафиксировано — или СИСТЕМНО заблокировано — выражение одного из актантов.

Принципы рассмотрения управляемых номинализаций и адвербиальных дериватов существенно различаются. В случае управляемой номинализации мы рассматриваем ОДНУ предикацию (и ее валентности), в случае адвербиального деривата — ДВЕ предикации: главную (сказуемое) и редуцированную (АД). В зависимости от соотношения этих двух предикаций (и отображаемых ими ситуаций) можно выделить две основные группы адвербиальных дериватов.

В ПЕРВУЮ группу входят АД, которые описывают действие ТОГО ЖЕ субъекта, что и главный предикат. В первой группе системно НЕ выражается субъект АД — он кореферентен субъекту главного глагола, т. е. совпадает с подлежащим, — поэтому такие АД условно можно назвать кореферентными. При этом вторая валентность АД (если она есть) — объектная или иная

- может выражаться: *Пассажиры ходили по перрону* ***в ожидании поезда*** = ожидая поезда (X ходил, X ждал поезда); *Пришел* ***в надежде поговорить*** = надеясь (X пришел, X надеется поговорить);
- может не выражаться: *Пассажиры* ***в волнении*** *ходили по перрону* = X ходил, X волновался (у предиката волноваться, вообще говоря, есть вторая валентность — причина: X волнуется из-за P, но она в данном случае не выражается);

- может вообще отсутствовать: ***В горячке*** *он не заметил, как выронил ключи*.

Кореферентный АД условно соответствует деепричастию: деепричастный оборот имеет такое же соотношение с главным глаголом. Аналогичное соотношение (совпадение) субъектов может встречаться в главном и придаточном предложении (ср. придаточные цели с инфинитивом: *Пришел, чтобы поговорить* — X пришел, X хочет поговорить).

Во ВТОРУЮ группу входят АД, которые описывают действие ДРУГОГО субъекта (в примерах — Y, или Сб-2), отличного от субъекта главного глагола (в примерах — X, или Сб-1). Поэтому их условно можно назвать некореферентными. Во второй группе субъект АД не совпадает с субъектом главного предиката, поэтому его можно выразить (но можно и не выражать):

Он [Сб-1] *явился на допрос в сопровождении* ***адвоката*** [Сб-2] = сопровождаемый адвокатом;

Спортсмен [Сб-1] *делал упражнения под наблюдением* ***тренера*** [Сб-2] = «наблюдаемый» тренером, ср. также без выраженного субъекта: *занимался под наблюдением*.

При этом не важно, имеет ли субъект адвербиального деривата реальное синтаксическое выражение, ср.: *закон был принят* [X-ом] *под давлением общественности* [Y-а] vs. *закон был принят* [X-ом] *под давлением* [→ Y-а]. Все равно АД выражает действие другого субъекта, и синтаксическое место при нем предназначено именно для этого субъекта. Такой оборот, вообще говоря, соответствует страдательному причастию (или «некореферентному» деепричастию в тех языках, где такие деепричастия допускаются).

Здесь мы должны сделать некоторые пояснения, касающиеся страдательных форм. В связи с нашим материалом необходимо различать страдательную форму в качестве главного предиката и страдательную конструкцию в качестве дополнительной предикации (причастного оборота). Страдательная форма главного предиката просто показывает, что подлежащим является не-первый актант глагола, — и как таковая ничего не дает для анализа АД. Страдательный оборот в качестве дополнительной предикации является показателем «разносубъектности» главной и зависимой предикаций (ср.: эксперт, приглашенный на заседание — ‘пришел эксперт; его пригласило другое лицо’ — иначе было бы эксперт, пригласивший…). В этом смысле страдательный оборот — лишь частный случай разносубъектности. Аналогично устроена разносубъектная каузативная конструкция, ср.: *Он поговорил с начальником* ***по просьбе*** *дяди* (X поговорил, т. к. Y попросил); *Он ушел пораньше* ***с разрешения*** *дежурного* (X ушел, т. к. Y разрешил).

Что касается выражения не-субъектных актантов — объекта или адресата, то у некореферентных АД они обычно не выражаются синтаксически зависимыми формами, но, разумеется, соответствующий семантический материал имеется в предложении (проблема семантического заполнения синтаксически не выраженных валентностей адвербиальных дериватов подробно рассматривается в уже упоминавшейся работе [Богуславский 2008]).

Таким образом, для управляемой номинализации главный вопрос — что из имеющихся актантов этой номинализации можно выразить синтаксически: субъект, объект или и то и другое. Для адвербиального деривата вопрос ставится

иначе: совпадает ли субъект АД с субъектом главной предикации; если да — его нельзя выразить; если нет — обычно можно (хотя это не всегда обязательно).

Таким образом, если возможности выражения валентностей у управляемой номинализации нужно анализировать в каждом конкретном предложении, то у АД эти возможности заранее известны. Легко убедиться, что если заданная схема нарушается, то мы имеем дело не с АД, а с управляемой номинализацией. Например, АД *в сопровождении кого* — некореферентный, т. е. описывает действие другого лица:

Подозреваемый (X=Сб-1) *явился на допрос* ***в сопровождении охранника*** (Y=Сб-2) = X пришел, Y сопровождал. Т.е. в таком обороте может быть выражен только субъект-2 адвербиального деривата, отличный от субъекта-1 главного глагола, но не объект сопровождения (который здесь совпадает с субъектом главной ситуации — подозреваемым). Следовательно, предложение *Охранник явился на допрос в сопровождении подозреваемого* не будет означать, что ‘охранник пришел, сопровождая подозреваемого’, а будет означать, что подозреваемый каким-то образом получил контроль над охранником (т. е. стал субъектом сопровождения) и привел охранника на допрос. Таким образом, у АД *в сопровождении кого* всегда выражается субъект сопровождения (‘кто сопровождает’). Однако на управляемую номинализацию — *сопровождение кого кем* — этот системный запрет ни в коей мере не распространяется, при управляемой номинализации в родительном падеже может стоять и объект (‘кого сопровождают’): *Истребители* (Сб) *участвуют в сопровождении штурмовиков* (Об) *и бомбардировщиков.* Возможно также одновременное выражение объекта и субъекта: *Нет ничего странного в сопровождении колонны* (Об) *военной техники автоматчиками* (Сб-1) — при этом субъекта-2 здесь вообще нет.

На свойства и поведение АД влияет семантика предлога и семантика отглагольного существительного. Обычно при определенном предлоге бывает определенный тип АД.

Так, с предлогом *В*, как правило, образуются некореферентные («разносубъектные») обороты, т. е. зависимая форма при них обозначает субъекта второстепенной предикации. В зависимости от семантического класса существительного можно выделить несколько групп таких оборотов: *(соната) в исполнении автора, (Хлестаков) в изображении Гоголя*, ср. также *в постановке / в переводе / в пересказе / в изложении / в переделке / в обработке* кого. К ним примыкает группа: *в оценке экспертов, в определении критиков, в трактовке Канта, в интерпретации Рихтера*, ср. также *в представлении обывателя, в понимании школьников* (ср. управляемую номинализацию: *эксперты были единодушны в своей оценке проекта*: проекта = объект). Другая группа — в *присутствии понятых, в сопровождении охраны, в окружении учеников*, к ним примыкают обороты с неотглагольными именами: *появляется на лыжных курортах в компании / в обществе девушек*.

Интересной особенностью некореферентных адвербиальных дериватов с предлогом *В* является невозможность выражения субъекта творительным падежом, ср. **Хлестаков в изображении Гоголем*; **Явился на допрос в сопровождении адвокатом*, при том что для управляемых номинализаций творительный

субъекта как раз весьма характерен, ср. *сопровождение колонны* ***автоматчиками***. Однако мы сейчас не можем углубляться в эту проблему.

С предлогом *В* бывают и кореферентные обороты — *в ожидании, в надежде, в стремлении* (см. выше), но в них участвует совершенно другой семантический класс существительных — внутренние (психологические) состояния человека.

Примечание. Обороты от предикатов с симметричными актантами имеют еще один предлог и присоединяют этот симметричный актант: *в союзе / в соавторстве / в согласии / в разладе / в дискуссии / в перепалке / в перебранке / в конкуренции / в сговоре с кем;* ср. также другие обороты с двумя личными актантами: *в тайне / в секрете от кого; в подчинении у кого; в оппозиции к кому.* Вообще, обороты с двумя предлогами — это особая тема, которую мы здесь не затрагиваем.

Большинство оборотов с **причинным** (в широком смысле) значением некореферентны, т. е. их зависимая форма выражает субъекта, отличного от субъекта главной предикации: [X] *под защитой, под охраной, под присмотром* [Y-а]; *под влиянием, под воздействием, под давлением; под угрозой, под страхом; с разрешения, с позволения, с согласия; по приказу, по совету, по просьбе, по настоянию, по рекомендации, по требованию, по поручению, по указу, по указанию, по указке, по приглашению, по назначению [врача], по предложению, по приговору, по решению, по доносу; по завещанию, по завету.*

Исключения здесь, как и в предыдущем классе, связаны с семантикой существительного: если существительное обозначает внутреннее состояние человека или неконтролируемую ситуацию с его участием, АД является кореферентным, т. е. относится к тому же субъекту, что и главная предикация: *по недомыслию, по неосторожности, по недосмотру, по ошибке, по незнанию* и т. п., ср.: *Он по незнанию нажал на кнопку* = ‘X не знал (чего-то) и поэтому X нажал на кнопку’. Аналогично (т. е. кореферентно) устроен АД с предлогом от и с названиями эмоций, ср. *от страха, от отчаяния*: *Он от страха стал нажимать на все кнопки подряд* = ‘X стал нажимать, потому что X испугался’; ср., однако: *От ветра открылась форточка* = ‘форточка (X) открылась, потому что ветер (Y) дул’ (ветер — не внутреннее состояние X-а, а внешняя причина главной ситуации).

Существенно иначе ведут себя адвербиальные дериваты с временным, условно-временным и условным значением: в их поведении обнаруживается меньше ограничений, чем у рассмотренных выше типов АД.

Большинство встретившихся в Корпусе АД с предлогом *ПРИ* во **временном** значении являются кореферентными, т. е. их невыраженный субъект совпадает с субъектом главного предиката (при АД может выражаться объект): *При проверке работы будьте внимательны* = ‘когда вы будете проверять, вы будьте внимательны’; *При проверке финансовых документов аудиторы обнаружили недостачу* = ‘когда X проверял, X обнаружил’. Речь идет о кореферентности именно с семантическим субъектом, а не формальным подлежащим главной предикации. При этом сам главный глагол может стоять в пассивной форме (т. е. его подлежащим будет объект):

При проверке была обнаружена недостача = ‘когда X проверял, X обнаружил’;

При подсчете голосов были допущены нарушения = ‘когда X подсчитывал голоса, X допустил нарушения’;

возможно также главное сказуемое в форме безличного глагола или безличного предикатива, когда нет материального подлежащего или оно неканоническое (в дательном падеже):

При совершении некоторых операций нужно быть особенно внимательным;

При одном взгляде на него [каждому] становится ясно…

Интересно, что если в главной предикации субъект не выражен, возможно выражение субъекта в АД: ***При проверке*** *документов* ***аудиторами*** *обнаружились нарушения*; ср. также пример из Корпуса: ***При проверке*** *работы поликлиники* ***комиссией*** *Московского областного отдела здравоохранения…*

Примечание. Не исключено, что возможно выражение субъекта при обеих предикациях — подчиненной (АД) и главной: ?При проверке документов аудиторами один из них обнаружил растрату — впрочем, такие примеры нам пока не встретились.

Возможно также некореферентное употребление АД с предлогом *ПРИ*:

Но при проверке документов Багдасарян и Степанян сбежали, оставив сумку с устройством в зале. [А. Д. Сахаров. Воспоминания (1983–1989)] = ‘когда Y проверял документы, X сбежал’.

Среди **условных** АД также встречаются как кореферентные *(При утере документа немедленно сообщайте дежурному)*, так и некореферентные *(При возгорании немедленно сообщайте дежурному*; *При попытке открыть замок включается сигнализация)* употребления.

Таким образом, временные и условные АД похожи скорее не на деепричастные и причастные обороты, у которых есть строгие (системные) ограничения на выражение того или иного (заранее определенного) актанта и на соотношение с главной предикацией, а на обстоятельственные придаточные, в которых, вообще говоря, могут выражаться разные актанты подчиненной предикации.

Разобранный материал свидетельствует о том, что разные АД находятся на разных ступенях грамматикализации и их поведение регулируется множеством факторов, которые нуждаются в подробном изучении. Так, группа АД с временным и условным значением находится на менее продвинутой ступени грамматикализации по сравнению с оборотами типа *в сопровождении* или под присмотром, а именно: (1) у таких АД ослаблено требование невыразимости субъекта АД при кореферентном употреблении (ср. *допустимое при проверке работы комиссией || были обнаружены нарушения* и недопустимое: **В ожидании поезда пассажирами бродили по перрону* [пассажиры]; ср. также при некореферентном употреблении: **В ожидании поезда пассажирами было объявлено о задержке прибытия)*; (2) у таких АД возможна кореферентная и некореферентная интерпретация не только при одном и том же предлоге, но и при одном и том же существительном: *при проверке* [X] *обнаружили* [X] vs. *при проверке* [Y] *сбежали* [X].

Примечание. В принципе, не исключено, чтобы одно и то же существительное могло входить в разные конструкции — кореферентную и некореферентную, — но обычно это конструкции с разными предлогами:

При выборе товара будьте внимательны = выбирая: вы выбираете, вы внимательны;

Фирма доставит любую модель по выбору клиента = выбранную: клиент выберет, фирма доставит (при этом надо иметь в виду, что совпадение ситуаций, обозначенных существительным выбор, не полное, а именно — различается их аспектуальная характеристика: при выборе = когда выбираете (НСВ), а *по выбору* = то, что выбрал (СВ)).

Итак, мы попытались показать, что адвербиальные дериваты имеют существенно иные свойства, чем управляемые номинализации, и их в первую очередь нужно сопоставлять не с управляемыми номинализациями, а с периферийными глагольными формами в составе зависимых (редуцированных) предикаций типа деепричастных и причастных оборотов, а также с обстоятельственными придаточными.

Адвербиальные дериваты представляют собой определенную ступень грамматикализации, промежуточную между «свободными» номинализациями и производными предлогами. АД стоят на более «высокой» ступени, чем производные предлоги, но по сравнению со управляемыми номинализациями свойства АД системно редуцированы, а именно:

(1) управляемая номинализация сохраняет — в принципе — полный набор возможностей выражения актантов (хотя не в каждом конкретном контексте выражаются они все); у АД один из актантов системно невыразим — или, наоборот, выразим (причем заранее известно, какой именно), — потому что АД — это редуцированная форма, с самого начала предназначенная для того, чтобы «сопровождать» другой глагол и обозначать ситуацию, определенным образом связанную с главной ситуацией (т. е. если для управляемой номинализации решается вопрос о выразимости ее субъекта и объекта, то для АД решается вопрос о соотношении субъектов главной и второстепенной предикации, и уже этим определяется возможность выражения субъекта АД как подчиненной предикации; что касается выражения объекта АД, то оно происходит «по остаточному принципу»);

(2) для АД, как и для других редуцированных предикаций (деепричастий, причастий) важной характеристикой является кореферентность / некореферентность с субъектом основной предикации.

Эти две важнейшие — и связанные между собой — характеристики (кореферентность / некореферентность с субъектом основной предикации и выразимость / невыразимость какого-л. актанта) сближают АД не только с деепричастиями и причастиями, но также с производными предлогами.

АД — это готовые формы, которые, вообще говоря, должны фиксироваться в словарях на том же основании, что и производные предлоги, однако большинство существующих словарей построены по «номинативному» принципу,

поэтому не могут иметь таких «входов», как *по просьбе* или *в защиту* (о необходимости создания неноминативных словарей мы подробно пишем в других работах, см. [Кустова 2008а; 2008б; 2011]). В то же время у адвербиальных дериватов есть и системные, «регулярные» свойства, которые могут описываться в грамматиках.

Литература

1. *Апресян Ю. Д.* Лексическая семантика. М., 1974.
2. *Богуславский И. М.* Замечания об актантной структуре адвербиальных дериватов // Die het kleine eert, is het grote weerd., Uitgeverij Pegasus, Amsterdam, 2003. P. 23–40. (Pegasus Oost-Europese Studies 1).
3. *Богуславский И. М.* Актантное поведение адвербиальных дериватов // Динамические модели: Слово. Предложение. Текст. Сб. статей в честь Е. В. Падучевой. — М.: ЯСК, 2008. С. 110–128.
4. *Крейдлин Г. Е.* Служебные и строевые слова // Семантика служебных слов. Межвузовский сборник научных трудов. — Пермь, 1982, изд-во ПермГУ.
5. *Кустова Г. И.* О «неноминативных» электронных словарях // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2008». — М, 2008а. С. 297–302.
6. *Кустова Г. И.* Обстоятельственные группы типа во всяком случае в современном русском языке // Инструментарий русистики: корпусные подходы. Slavica Helsingiensia 34. — Хельсинки, 2008б. С. 126–139.
7. *Кустова Г. И.* Конструкции с абстрактными существительными и их отражение в электронном словаре // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2011». — М, 2011. С. 379–390.
8. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл ⇔ Текст». — М., 1974 (2-е изд. — 1999).
9. *Падучева Е. В.* О производных диатезах отпредикатных имен в русском языке // Проблемы лингвистической типологии и структуры языка. — М., 1977. С. 84–107.
10. *Падучева Е. В.* Посессивы и имена способа действия // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2009». — М., 2009. С. 365–372.
11. *Падучева Е. В.* Актантная структура и диатеза отглагольного имени // Научно-техническая информация. Сер. 2. Информационные процессы и системы. — 2010, № 6. С. 24–30.
12. *Пазельская А. Г.* Образование отглагольных существительных и актантные преобразования в русском языке // Динамические модели: Слово. Предложение. Текст. Сб. научных трудов в честь Е. В. Падучевой. — М.: ЯСК, 2008. С. 634–645.
13. *Тихонов А. Н.* Морфология // Современный русский язык. В 3 ч. Ч. 2. Н. М. Шанский, А. Н. Тихонов. Словообразование. Морфология. — М., 1987.

14. *Alexiadou, Artemis.* Functional Structure in Nominals. — Amsterdam-Philadelphia: John Benjamins, 2001.
15. *Comrie, Bernard.* Nominalization in Russian: lexical noun phrases or transformed sentences? // C. V. Chvany & R. D. Brecht (eds.). Morphosyntax in Slavic. — Columbus, 1980. P. 212–220. — Рус. пер.: Комри Б. Номинализация в русском языке: словарно-задаваемые именные группы или трансформированные предложения? // Новое в зарубежной лингвистике. Вып. XV. — М., С. 42–49.
16. *Comrie, Bernard & Sandra A. Thompson.* Lexical nominalization. // Shopen, Timothy. (ed.). Language typology and syntactic description. Vol. 2–3. — Cambridge, 1985, vol. 3. P. 349–398.
17. *Filip, Hana.* Aspect, eventuality types and noun phrase semantics. — New York, London: Garland Publishing, 1999.
18. *Moltmann, Frederike.* Nominalizations, Events, and Other Concrete Objects. — Ms., University of Stirling, 2005.
19. *Nichols, Johanna.* Nominalization and assertion in scientific Russian prose. // John Haiman & Sandra A. Thompson (eds.). Clause Combining in Grammar and Discourse. — Amsterdam; Philadelphia, 1988. P. 349–428.
20. *Rappaport, Gilbert C.* On the adnominal genitive and the structure of noun phrases in Russian and Polish // Linguistique et Slavistique. Melanges offerts a Paul Garde. Eds. M. Guiraud-Weber, C. Zaremba. — Paris, 1992. P. 241–262.
21. *Rozwadowska B.* Thematic Restrictions on Derived Nominals // Syntax and Semantics 21: Thematic Relations. Ed. W. Wilkins. — San Diego, CA: Academic Press, 1988. P. 147–66.
22. *Rozwadowska B.* Towards a unified theory of nominalizations. External and internal eventualities. — Wroclaw, 1997.
23. *Rozwadowska B.* Event Structure, Argument Structure and the by-phrase in Polish Nominalizations // Lexical Specification and Insertion. Eds. P. Coopmans, M. Everaert, J. Grimshaw. — Amsterdam; Philadelphia, 2000. P. 329–347.

References

1. *Alexiadou, Artemis* (2001), Functional Structure in Nominals, John Benjamins, Amsterdam-Philadelphia.
2. *Apresjan Ju. D.* (1974), Lexical Semantics [Leksicheskaja semantika], Nauka, Moscow.
3. *Boguslavskij I. M.* (2003), On the argument structure of adverbial derivatives [Zamechanija ob aktantnoj structure adverbial'nyh derivatov] // Die het kleine eert, is het grote weerd., Uitgeverij Pegasus (Pegasus Oost-Europese Studies 1), Amsterdam, pp. 23–40.
4. *Boguslavskij I. M.* (2008), Argument behavior of adverbial derivatives [Aktantnoe povedenie adverbial'nyh derivatov], in: Dynamic models: Word. Sentence. Text. [Dinamicheskie modeli: Slovo. Predlozhenie. Tekst], JaSK, Moscow, pp. 110–128.

5. *Comrie, Bernard & Sandra A.* Thompson (1985), Lexical nominalization, in: Shopen, Timothy. (ed.). Language typology and syntactic description. Vol. 2–3, Cambridge, vol. 3, pp. 349–398.
6. *Comrie, Bernard* (1980), Nominalization in Russian: lexical noun phrases or transformed sentences?, in: C. V. Chvany & R. D. Brecht (eds.). Morphosyntax in Slavic, Columbus, pp. 212–220.
7. *Filip, Hana* (1999), Aspect, eventuality types and noun phrase semantics, Garland Publishing, New York, London.
8. *Krejdlin G. E.* (1982), Auxiliary and structural words [Sluzhebnye i stroevye slova], in: Semantics of auxiliary words [Semantika sluzhebnyh slov], Perm' Univ. Press, Perm'.
9. *Kustova G. I.* (2008a), About «non-nominative» dictionaries (lexical databases) [O nenominativnyh electronnyh slovarjah], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2008) [Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2008"], Issue 7 (14), Moscow, pp. 297–302.
10. *Kustova G. I.* (2008b), Adverbial phrases like in any case in modern Russian [Obstojatel'stvennye gruppy tipa vo vsjakom sluchae v sovremennom russkom jazyke], Slavica Helsingiensia, 34, Helsinki, pp. 163–175.
11. *Kustova G. I.* (2011), Constructions with abstract nouns in an electronic database [Konstruktsii s abstraktnymi sushchestvitel'nymi i ih otrazhenie v electronnom slovare], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2011) [Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Po Materialam Ezhegodnoj Mezhdunarodnoj Konferentsii "Dialog 2011"], Issue 10 (17), Moscow, pp. 379–390.
12. *Mel'chuk I. A.* (1974/1999), An experience of linguistic theory «Meaning ⇔ Text» [Opyt teorii lingvisticheskih modelej «Smysl ⇔ Tekst»], Moscow (2-nd edition — 1999).
13. *Moltmann, Frederike* (2005), Nominalizations, Events, and Other Concrete Objects, University of Stirling, Ms.
14. *Nichols, Johanna* (1988), Nominalization and assertion in scientific Russian prose, in: John Haiman & Sandra A. Thompson (eds.). Clause Combining in Grammar and Discourse, Amsterdam; Philadelphia, pp. 349–428.
15. *Paducheva E. V.* (1977), On derived diathesis of nominalizations in Russian [O proizvodnyh diatezah otpredikatnyh imen v russkom jazyke], in: Problems of linguistic typology and language structure [Problemy lingvisticheskoj tipologii I struktury jazyka], Moscow, pp. 84–107.
16. *Paducheva E. V.* (2009), Possessives and manner of action nouns: corpus based exploration [Posessivy i imena sposoba dejstvija], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2009". Issue 8 (15), Moscow, pp. 365–372.
17. *Paducheva E. V.* (2010), Argument structure and diathesis of nominalizations [Aktantnaja struktura i diateza otglagol'nogo imeni], Научно-техническая информация. Scientific and Technical Information. Series 2. Information processes and systems [Nauchno-tehnicheskaja informatsija. Ser. 2. Informatsionnye protsessy i sistemy], № 6, pp. 24–30.

18. *Pazel'skaja A. G.* (2008), Derivation of nominalizations and argument transformations in Russian [Obrazovanie otglagol'nyh sushchestvitel'nyh i aktantnye preobrazovanija v russkom jazyke], in: Dynamic models: Word. Sentence. Text [Dinamicheskie modeli: Slovo. Predlozhenie. Tekst], JaSK, Moscow, pp. 634–645.
19. *Rappaport, Gilbert C.* (1992), On the adnominal genitive and the structure of noun phrases in Russian and Polish, in: Linguistique et Slavistique. Melanges offerts a Paul Garde. Eds. M. Guiraud-Weber, C. Zaremba, Paris, pp. 241–262.
20. *Rozwadowska B.* (2000), Event Structure, Argument Structure and the by-phrase in Polish Nominalizations, in: Lexical Specification and Insertion. Eds. P. Coopmans, M. Everaert, J. Grimshaw. Amsterdam; Philadelphia, pp. 329–347.
21. *Rozwadowska B.* (1988), Thematic Restrictions on Derived Nominals, Syntax and Semantics 21: Thematic Relations. Ed. W. Wilkins. Academic Press, San Diego, CA, pp. 147–66.
22. *Rozwadowska B.* (1997), Towards a unified theory of nominalizations. External and internal eventualities. Wroclaw.
23. *Tikhonov A. N.* (1987), Morphology [Morfologija], in: Modern Russian, In 3 p. P. 2. N. M. Shanskij, A. N. Tikhonov. Word-formation. Morphology [Sovremennyj russkij jazyk. V 3 ch. Ch. 2. N. M. Shanskij, A. N. Tikhonov. Slovoobrazovanie. Morfologija], Moscow.

ТИПОЛОГИЧЕСКАЯ БАЗА ДАННЫХ АДЪЕКТИВНОЙ ЛЕКСИКИ[1]

Кюсева М. В. (mkyuseva@gmail.com),
Резникова Т. И. (tanja.reznikova@gmail.com),
Рыжова Д. А. (daria.ryzhova@mail.ru)

НИУ ВШЭ, Москва, Россия

Ключевые слова: база данных, лексическая типология, признаковая лексика, словарь

A TYPOLOGICALLY ORIENTED DATABASE OF QUALITATIVE FEATURES

Kyuseva M. V. (mkyuseva@gmail.com),
Reznikova T. I. (tanja.reznikova@gmail.com),
Ryzhova D. A. (daria.ryzhova@mail.ru)

HSE, Moscow, Russian Federation

The article presents the Typological Database of Qualities, which aims at providing a new tool for research in lexical typology. The database contains information on the lexicalization of several semantic fields of adjectives in different languages (like ‘sharp’ — ‘blunt’, ‘empty’ — ‘full’, ‘solid’ — ‘soft’, ‘thick’ — ‘thin’, ‘smooth’ — ‘rough’, etc.). We discuss issues concerning database structure (in particular, the choice of information units that would make the meanings from different languages comparable to each other). Special attention is devoted to the representation of figurative meanings in the Database which allows to investigate the models of their derivation from the literal meanings. The developed database can be used for solving both theoretical and practical tasks. On the practical level, the Database may serve as a multilingual dictionary which accounts for fine-grained differences in meaning between individual words. On the theoretical side, the Database allows for various generalizations on cross-linguistic patterns of polysemy and semantic change.

Key words: database, lexical typology, adjectives, dictionary

[1] Исследование выполнено при поддержке гранта РФФИ №11-06-00385-а

1. Введение

Настоящее исследование посвящено семантическому анализу признаковой лексики в типологической перспективе. В работах по лексической типологии — направления, переживающего бурное развитие в последние годы, — большее внимание пока уделяется глаголам, см. широкомасштабные проекты по изучению предикатов позиции [Newman (ed.) 2002], движения в воде [Майсак, Рахилина (ред.) 2007], разрушения [Majid et al. 2007], еды и питья [Newman (ed.) 2009], перемещения объекта [Kopecka, Narasimhan (eds.) 2012]. В отличие от глагольной, признаковая лексика, за исключением цветообозначений (с изучения которых традиционно ведется отсчет лексико-типологических исследований в целом, см. [Berlin, Kay 1969]), до недавнего времени редко становилась объектом межъязыкового анализа. Среди работ в этой области можно выделить только описания перцептивных прилагательных, выполненные в рамках психолингвистического направления в Институте им. Макса Планка в Неймегене [Majid, Levinson (eds.) 2011].

Эта исследовательская лакуна стала предпосылкой для запуска силами Московской лексико-типологической группы[2] большого проекта по изучению признаковой лексики. Объектом исследования стали прилагательные со значением базовых физических признаков. Так, с разной степенью подробности были проанализированы семантические поля ‘острый’ — ‘тупой’ (Кюсева 2012), ‘мягкий’ — ‘твердый’ (Павлова 2012), ‘легкий’ — ‘тяжелый’ (Кюсева, Рыжова, Холкина 2012), ‘полный’ — ‘пустой’ (Тагабилева 2011), свойства поверхностей (‘ровный’ — ‘гладкий’ — ‘шершавый’) (Кашкин 2012) и др. С накоплением значительного числа языковых данных стало очевидно, что для обобщения материала и выявления типологически релевантных характеристик лексических систем необходимо свести всю полученную информацию и представить ее в виде базы данных. О структуре и пользовательских возможностях такой базы и пойдет речь в этой статье.

2. Структура Базы данных

У нас уже был опыт работы с лексическими базами данных: в 2010 году завершился проект по созданию Базы данных по многозначным качественным прилагательным и наречиям русского языка. Призванная отразить модели полисемии, характерные для русской признаковой лексики, она включает в себя 300 частотных прилагательных и наречий, обозначающих физические свойства объектов. Основным входом в Базе является значение слова, для которого

---

[2] Московская лексико-типологическая группа (MLexT) была создана для реализации крупномасштабных проектов, которые требовали привлечения значительного числа специалистов по разным языкам. Среди результатов работы группы можно отметить исследования глаголов движения в воде [Майсак, Рахилина (ред.) 2007], предикатов боли [Брицын и др. (ред.) 2009], вращения [Круглякова, Рахилина 2010], звука и др.

указывается следующая информация: краткое словарное толкование, таксономический класс прилагательного в этом значении, семантические свойства контекста, в котором реализуется данное значение (прежде всего таксономические признаки определяемого существительного), значение, от которого производно данное, тип перехода и некоторая другая релевантная информация (подробнее см. Карпова и др. 2010[3]). Для слов строились семантические сети, в которых каждое последующее значение образовано от предыдущего посредством семантической деривации определенного типа. Анализ материала, собранного в единую Базу данных, показал, что семантические переходы в лексике не случайны: они строятся по определенным моделям, и количество этих моделей в языке ограниченно.

Полученный каталог переходов в русской признаковой лексике естественным образом заставил нас задуматься о том, насколько универсальны выявленные модели. Однако выход в типологическую перспективу в первую очередь потребовал от нас решения качественно новых задач. Дело в том, что сопоставительному изучению моделей полисемии должно предшествовать сравнение исходной семантики соответствующих лексем: нужно понять не только, как соотносятся между собой производные значения, но и насколько близки друг к другу их источники. Ведь общеизвестно, что и в зоне прямых, физических значений между переводными эквивалентами практически никогда нет полного соответствия.

Так, русскому прилагательному *острый* (ср. *острый нож, игла, подбородок*) в коми-зырянском языке соответствуют две лексемы — *лэчыд* и *ёсь*. Первое слово характеризует остроту режущих инструментов — ножей, пил, кос (*лэчыд пурт* — ‘острый нож’), а второе — колющих инструментов или объектов с зауженным кончиком — стрел, копий, колов, а также носов, подбородков (*ёсь пу* — ‘острая палка’). Во французском языке идею остроты передают три прилагательных: *tranchant*, описывающее объекты с острым краем (*un couteau tranchant* — ‘острый нож’), *aigu*, характеризующее острые колющие инструменты (*une flèche aiguë* — ‘острая стрела’), и *pointu*, передающее идею острой формы (*un nez pointu* — ‘острый нос’).

Нередки в нашей выборке и случаи, когда русское прилагательное оказывается у́же своего переводного эквивалента. Так, французская лексема *dur*, соответствующая русскому слову *твердый*, характеризует не только объективное свойство предметов ‘иметь плотную консистенцию, сохранять свою форму’ (*une pierre dure* — твердый камень), но и субъективные ощущения экспериенцера при соприкосновении с такими предметами, которые в русском языке описывает слово *жесткий* (*un matelas dur* — жесткий матрас).

Ситуация частичного пересечения значений прилагательных представлена, например, в зоне признаков размера. Русскому прилагательному *тонкий* в хантыйском языке соответствует две лексемы — *иохәł* и *vas’*. *Иохәł* описывает только плоские тонкие объекты — «слои»: книги, матрасы, стены. *Vas’* характеризует вытянутые объекты цилиндрической формы: стебли,

[3] База доступна по ссылке: http://rakhilina.ru/adjectives_query.html

веревки, столбы. Однако вместе с тем лексема *vas'* покрывает и значения, которые в русском языке передаются с помощью прилагательного *узкий*: например, в словосочетании со значением ‘узкая дорога’ смысл ‘узкий’ будет выражен прилагательным *vas'*. Таким образом, лексема *vas'* пересекается частью своих контекстов с прилагательным *тонкий*, а частью — с прилагательным *узкий*.

Наличие таких примеров ведет к двум важным следствиям. Во-первых, с выходом в типологическую перспективу мы переходим от изучения значений отдельных слов к исследованию семантики полей: в признаковое поле ‘острый’ войдет русское прилагательное *острый*, коми-зырянские лексемы *лэчыд* и *ёсь*, французские слова *tranchant*, *aigu* и *pointu*; в поле ‘твердый’ наряду с французским прилагательным *dur* и русским *твердый* попадет лексема *жесткий*; в поле ‘тонкий’ из русского материала будет включено не только само слово *тонкий*, но и *узкий*. Такой подход необходим для того, чтобы обеспечить корректное межъязыковое сопоставление значений признаковых лексем.

Во-вторых, типологическое описание лексики требует изменений и в самой структуре Базы. Теперь отдельным входом должна быть единица меньшая, чем значение. Одна из возможностей — принять в качестве такой единицы минимальную ситуацию, которая может в каком-либо языке обслуживаться отдельной лексемой. В нашей терминологии за этими ситуациями закреплено наименование «фрейм» (см. подробнее Рахилина, Резникова 2013). Так, в случае с полем ‘острый’ входами в Базе будут следующие фреймы:

- ‘инструмент с заточенным краем’ (объекты только такого типа характеризуют в прямом значении коми-зырянское прилагательное *лэчыд* и французское *tranchant*);
- ‘инструмент с колющим концом’ (класс объектов, который является единственной сферой действия для французской лексемы *aigu* в прямом значении);
- ‘объект острой формы’ (только такие объекты описывает французское прилагательное *pointu*).

Многие прилагательные, однако, охватывают в прямом значении несколько фреймов: так, коми-зырянская лексема *ёсь* обозначает и инструменты с колющим концом, и объекты острой формы, а русское прилагательное *острый* покрывает все три ситуации. В таких случаях одному значению лексемы в Базе должно соответствовать несколько строк — по одной для каждого фрейма (для прилагательного *ёсь* в исходном значении нужно иметь две строки, для *острый* — три).

В поле ‘твердый’ исходные значения прилагательных покрывают фреймы ‘объекты, способные сохранять неизменными свою форму’ (= *твердые*) и ‘объекты, плотные на ощупь’ (= *жесткие*)..В поле ‘тонкий’ мы выделяем фреймы ‘тонкие плоские объекты (=«слои»)’, ‘тонкие вытянутые предметы цилиндрической формы’, ‘полосы’ (= *узкие* объекты) и др.

Принятие в качестве входа в Базу фрейма позволило бы отслеживать проиллюстрированные выше случаи расхождения исходных значений прилагательных. Однако с исследовательской точки зрения интересны и более тонкие различия в употреблении переводных эквивалентов: например, случаи, когда признак того или иного объекта, никогда не лексикализуясь в отдельном прилагательном, в одних языках соотносится с одним фреймом, а в других — с другим. Примером такой ситуации может вновь послужить поле ‘острый’. Так, ‘острые когти/ногти’ ни в одном из пятнадцати исследованных языков не описываются самостоятельной признаковой лексемой. Однако в языках, разделяющих режущие и колющие инструменты, этот тип объекта ведет себя по-разному: иногда он «присоединяется» к фрейму инструментов с зауженным концом (как во французском, коми-зырянском, китайском языках), как обладающий внешним сходством с ними, а иногда — с острым краем (как в итальянском языке), как предмет не укалывающий, а царапающий, режущий. Чтобы отмечать такие случаи, мы выбрали в качестве входа в базу единицу, меньшую, чем фрейм — элемент этого уровня мы называем «микрофреймом».

Список микрофреймов составляется в результате обобщения материалов типологической анкеты. Анкеты представляют собой таблицы, строки которых заполняются существительными (отражающими те или иные ситуации, релевантные для данного признакового поля), а столбцы — прилагательные данного поля. В ячейках, находящихся на пересечении прилагательного и существительного, отмечается, может ли данная ситуация описываться данной лексемой (см. фрагмент анкеты для поля ‘тонкий’ в русском языке в таблице 1).

Первоначально списки существительных составляются на основе анализа русского материала, который проводится с помощью толковых словарей и НКРЯ, потом они дополняются материалами словарей и корпусов других языков. В итоге получается набор из 200–300 единиц, соответствующих очень подробной классификации ситуаций, которые гипотетически могут описываться прилагательными данного поля. Далее при помощи корпусов и опроса информантов каждое прилагательное тестируется по всем строчкам анкеты. В результате проверки большого числа прилагательных различных языков выясняется, что в анкете имеются строчки, которые никогда не разводятся по разным лексемам, иными словами, если лексема может использоваться для одной из этих строк, то все остальные строки этого «комплекса» также могут описываться этой лексемой. Так, если в данном языке признаковое слово со значением ‘тонкий’ сочетается с существительным ‘веревка’, то оно же будет сочетаться со словами ‘канат’, ‘трос’, ‘шнур’. Поэтому для Базы данных все эти строки анкеты «склеиваются» в один микрофрейм, а в качестве названия для него выбирается наиболее нейтральный и культурно независимый тип объектов данного ряда. Таким образом, заносимые в Базу уже после подробного исследования поля микрофреймы включают только значимые строки — те, для которых потенциально возможна разная сочетаемость признаковых слов.

Таблица 1. Фрагмент типологической анкеты для признакового поля ‘тонкий’ в русском языке

	Тонкий	Узкий
Стебель	+	–
Ткань	+	–
Дорога	–	+
…		

Итак, типологически ориентированная База данных адъективной лексики, в которую планируется занести информацию о 30 признаковых полях по 5–15 языкам, устроена следующим образом. Единица входа — это микрофрейм, т. е. релевантная для признакового поля ситуация. Для нее указано прилагательное, которое этот микрофрейм покрывает, язык, из которого взято данное прилагательное, фрейм, который оно реализует в сочетании с данным существительным, а также признаковое поле и таксономический класс прилагательного в этом употреблении. Каждый микрофрейм сопровождается примером. Так, одна из строк Базы для русского прилагательного *тонкий* будет иметь следующий вид:

Таблица 2. Пример заполнения типологически ориентированной базы данных признаковой лексики: русская лексема *тонкий*

Микрофрейм	Лексема	Язык	Фрейм	Поле	Таксономический класс
‘тонкий матрас’	Тонкий	Русский	‘тонкие вытянутые плоские предметы’	‘тонкий’	‘размер’

А одна из строк для слова *тяжёлый* будет выглядеть так:

Таблица 3. Пример заполнения типологически ориентированной базы данных признаковой лексики: русская лексема *тяжёлый*

Микрофрейм	Лексема	Язык	Фрейм	Поле	Таксономический класс
‘тяжёлая сумка’	Тяжёлый	Русский	‘предмет, который тяжело нести’	‘тяжёлый’	‘вес’

Таким образом, в Базе указывается информация разной степени обобщенности: от самой конкретной (микрофрейм) до самой общей (таксономический класс). Такое представление информации позволяет проводить самые разные наблюдения над семантикой лексики в разных языках (подробнее см. в разделе «Примеры запросов»).

3. Переносные значения

Особой проблемой является представление в Базе переносных значений слов. В Базе данных по многозначным качественным прилагательным и наречиям русского языка, созданной специально для изучения моделей полисемии, для каждого переносного значения указывалось значение, от которого образовано данное, и тип семантического перехода. Сплошное исследование полисемии русской признаковой лексики позволило выделить распространенные, часто повторяющиеся модели переносов. К ним относится, например, метафора, связывающая физическое свойство предмета и нефизическое свойство лица (*гнилое яблоко — гнилой интеллигент, вялый цветок — вялый студент*); или же физическое свойство объекта и нефизическое свойство абстрактной сущности (*пустая коробка — пустое замечание, прохладная поверхность — прохладные отношения*). С помощью этой базы стало возможным выделить и распространенные типы метонимических переносов, а также исследовать третий, отличный от метафоры и метонимии переход, который мы называем ребрендингом (подробнее о переходах в русской атрибутивной лексике см. Рахилина и др. 2010).

Получившиеся результаты естественным образом повлекли за собой вопрос: можно ли считать эти переносы и их пропорциональное соотношение универсальными? Иными словами, повторятся ли найденные метафоры, метонимии и ребрендинги в других языках? Обнаружатся ли те, которые отсутствуют в русском материале? И верно ли, что частотные переносы в русском языке окажутся таковыми в лексике других языков?

Типологическая база данных адъективной лексики должна уметь отвечать и на такие вопросы. Однако представлять в ней переносные значения так же, как и в русской Базе данных, не представляется возможным, так как определение типа перехода в лексике неродного для исследователей языка оказывается очень субъективным решением, которое не хотелось бы навязывать будущим пользователям Базы. Поэтому было решено характеризовать каждый микрофрейм Базы только по типу значения: прямое или переносное. Такой подход позволяет прослеживать семантические связи между отдельными значениями многозначных прилагательных и сравнивать их с моделями полисемии в русской признаковой лексике (ср. в этой связи проект под руководством Анны А. Зализняк, также ориентированный на установление подобных семантических отношений: Зализняк 2009).

4. Примеры запросов

Особенности структуры разработанной нами Базы позволяют извлекать из неё данные различных типов. Приведём несколько примеров возможных запросов.

А) Поиск по лемме

Поиск по столбцу, в котором указывается конкретная лемма, позволяет получить полный список фреймов (как прямых, так и переносных значений), покрываемых данной лексемой. В этой ситуации База данных выступает в роли качественного типологически ориентированного одноязычного словаря, в котором представлены все значения лексемы и особенности её сочетаемости.

Б) Поиск по фрейму

Этот тип запросов аналогичен предыдущему: он также подразумевает поиск по одному столбцу таблицы. В этом случае База данных выступает в качестве переводного словаря принципиально нового типа, выгодно отличающегося от традиционных переводных словарей, во-первых, мультиязычностью, а во-вторых, наличием понятных и исчерпывающих правил употребления переводных эквивалентов.

И действительно, сведений, указанных в традиционных переводных словарях, обычно недостаточно для правильного употребления лексем в речи. Так, например, во всех сербско-русских словарях в качестве переводного эквивалента лексемы *тежак* указано русское прилагательное *тяжёлый*, однако нигде нет сведений о том, что вполне приемлемое словосочетание *тешке паре* (букв.: «тяжёлые деньги») в современном сербском языке означает не ‘деньги, которые много весят’ и даже не ‘деньги, которые добываются с трудом’, а ‘много денег’, т. е. на русский язык это прилагательное должно переводиться не лексемой *тяжёлый*, а лексемой *большой* или *бешеный*.

Наша База данных позволяет избежать таких ошибок, так как в ней чётко прописано, какие фреймы какими лексемами покрываются, а особенности сочетаемости каждого слова легко выводятся из его набора фреймов. Иными словами, если мы осуществим в Базе поиск по фрейму ‘большое количество’ (именно к этому фрейму относится микрофрейм ‘много денег’), то мы найдём строки с прилагательными *тежак, big, большой,* но не *тяжёлый*.

В) Поиск случаев совмещения фреймов

Этот тип запросов является комбинацией первых двух. Он позволяет искать в Базе информацию о том, с какими фреймами в одной лексеме может совмещаться данный фрейм и насколько такое объединение частотно (= типологически релевантно). Запросы такого вида позволяют, например, определить, различаются ли в данном языке в поле ‘тонкий’ класс вытянутых цилиндрических и класс вытянутых плоских объектов. Эта информация может оказаться полезной для лексического типолога: она позволяет делать обобщения по поводу различных моделей заполнения поля. Например, такие данные

позволяют выявить тенденцию к более слабой разработке зон малых размеров по сравнению с большими в разных языках. Так, в хантыйском языке в зоне именно малых размеров не проводится в целом обычного для европейских языков разграничения между вертикально ориентированными объектами жёсткой конфигурации (такими, как столбы или стены) и ёмкостями (такими, как водоём или миска), ср. английские *low* vs. *shallow*, немецкие *niedrig* vs. *seicht*, русские *низкий* vs. *мелкий*: в хантыйском языке и низкое дерево, и неглубокая река описываются одним прилагательным *łeł*, в то время как в зоне больших размеров эти фреймы не объединяются. Тенденции к меньшей разработке зоны малых размеров наблюдаются и во французском и латинском языках.

Г) Поиск переносных фреймов для того или иного исходного фрейма

Типологическая база данных адъективной лексики позволяет прослеживать закономерности и в зоне метафорических переносов. Как мы уже сказали, в ней не отмечается тип перехода, которым один фрейм прилагательного связан с другим. Однако классификация каждого входа по типу значения (прямое/производное) позволяет искать информацию о том, какие исходные фреймы есть у слова с данным переносным. Таким образом можно проверять типологическую релевантность различных моделей семантических сдвигов.

Оказывается, что для любого признакового поля можно выделить 20–30 переносных значений, которые регулярно повторяются из языка в язык. Так, для поля ‘острый’ такими значениями будут ‘сильная боль’, ‘хороший ум’, ‘высокий/неприятный звук’, ‘обидные слова’ и другие. Несмотря на то, что всегда остается шанс встретить новую метафору, ядро переносов определяется уже после 10–12 языков и обычно в ходе дальнейшего исследования существенно не изменяется.

Более того, при расширении языковой выборки появляется возможность определить в исходной физической зоне конкретный фрейм, к которому восходит тот или иной переносный. Так, в поле ‘острый’ фрейм ‘резкий, порывистый ветер’, по всей вероятности, связан с исходным ‘инструмент с режущим краем’: в нашей выборке он регулярно встречается у лексем, которые в прямом значении описывают только режущие инструменты, и ни разу не встречается у «колющих» прилагательных.

5. Выводы

Таким образом, мы разработали новый инструмент, который может быть полезен при решении разного рода прикладных и теоретических задач. К первым относится применение Базы в качестве мультиязычного словаря (как для ручного, так и, в перспективе, для машинного перевода), ко вторым — разнообразные типологические исследования в области семантики признаковой лексики, в том числе изучение моделей полисемии.

Литература

1. *Брицын В. М., Рахилина Е. В., Резникова Т. И., Яворская Г. М.* (ред.). Концепт БОЛЬ в типологическом освещении. Киев: Видавничий Дім Дмитра Бураго, 2009
2. *Зализняк, Анна А.* О понятии семантического перехода // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: Изд-во РГГУ, 2009, с. 107–112.
3. *Карпова О. С., Архангельский Т. А., Кюсева М. В., Рахилина Е. В., Резникова Т. И., Рыжова Д. А., Тагабилева М. Г.* База данных по многозначным качественным прилагательным и наречиям русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2010» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). — М.: РГГУ, 2010.
4. *Кашкин Е. В.* 2012. Свойства поверхностей как объект лексической типологии: направления и перспективы исследования (на материале некоторых уральских языков) // Вестник Томского государственного педагогического университета, 1.
5. *Круглякова В. А., Рахилина Е. В.* Глаголы вращения: лексическая типология // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2010» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16).– М.: РГГУ, 2010.
6. *Кюсева М. В.* (2012) Лексическая типология семантических сдвигов названий качественных признаков ‘острый’ и ‘тупой’. Дипломная работа. М., МГУ.
7. *Кюсева М. В., Рыжова Д. А., Холкина Л. С.* Прилагательные *тяжёлый* и *лёгкий* в типологической перспективе // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2012» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). — М.: РГГУ, 2012.
8. *Майсак Т. А., Рахилина Е. В.* (ред.) 2007. Глаголы движения в воде: лексическая типология. М.: Индрик.
9. *Павлова Е. К.* 2012. Качественные признаки ‘мягкий’ и ‘твердый’ в типологической перспективе. Курсовая работа. М., МГУ.
10. *Рахилина Е. В., Резникова Т. И., Карпова О. С.* 2010. Семантические переходы в атрибутивных конструкциях: метафора, метонимия и ребрендинг // Е. В. Рахилина (отв. ред.) Лингвистика конструкций. М.: Азбуковник.
11. *Рахилина Е. В., Резникова Т. И.* 2013. Фреймовый подход к лексической типологии // Вопросы языкознания, № 2, с. 3–31.
12. *Тагабилева М. Г.* 2011. Качественные признаки ‘пустой’, ‘полный’: к построению семантической типологии. Курсовая работа. М., МГУ.
13. *Berlin, B.; Kay, P. 1969. Basic color terms: Their universality and evolution. Berkeley: Univ. of California press.*

14. *Haspelmath, M.* The geometry of grammatical meaning: Semantic maps and crosslinguistic comparison // M. Tomasello (ed.). The new psychology of language. V. 2. Mahwah (NJ), 2003.
15. *Kopecka, A., & Narasimhan, B.* (Eds.) 2012. Events of putting and taking: A crosslinguistic perspective. Amsterdam: Benjamins.
16. *Majid, A., Bowerman, M., Van Staden, M., Boster, J. S.* The semantic categories of cutting and breaking events: A crosslinguistic perspective, (2007) *Cognitive Linguistics*, 18 (2), pp. 133–152.
17. *Majid A., Levinson S. C.* (eds.) 2011. The senses in language and culture. The Senses & Society [Special Issue], 6(1).
18. *Newman, J.* (ed.) 2002. The Linguistics of Sitting, Standing, and Lying. [Studies in Typological Linguistics 51]. Amsterdam/Philadelphia: John Benjamins.
19. *Newman, J.* (ed.) 2009. The Linguistics of Eating and Drinking. [Studies in Typological Linguistics 84]. Amsterdam/Philadelphia: John Benjamins.

References

1. *Berlin, B.; Kay, P.* 1969. Basic color terms: Their universality and evolution. Berkeley: Univ. of California press.
2. *Britsyn V. M., Rahilina E. V., Reznikova T. I., Javorskaja G. M.* (eds.) 2009. Kontsept BOL' v tipologicheskom osveshchenii [Concept of PAIN: towards a Typology]. Kiev: Vidavnichij Dim Dmitra Burago.
3. *Haspelmath, M.* The geometry of grammatical meaning: Semantic maps and crosslinguistic comparison // M. Tomasello (ed.). The new psychology of language. V. 2. Mahwah (NJ), 2003.
4. *Karpova O. S., Arhangel'skij T. A., Kjuseva M. V., Reznikova T. I., Rahilina E. V., Ryzhova D. A., Tagabileva M. G.* 2010. The Database on Russian Polysemous Adjectives and Adverbs [Baza dannyh po mnogoznachnym kachestvennym prilagatel'nym i narechijam russkogo jazyka]. *Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010"* [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Moscow, pp.163–168.
5. *Kashkin E. V.* (2012) Qualities of Surface as the Object of Lexical Typology: Some Perspectives of the Research (Based on the Materials of the Uralic languages) [Svojstva poverhnostej kak ob"ekt leksicheskoj tipoogii: napravlenija i perspektivy issledovanija (na materiale nekotoryh ural'skih jazykov)], Vestnik Tomskogo gosudarstvennogo pedagogicheskogo universiteta [Bulletin of the Tomsk State Pedagogical University], 1.
6. *Kjuseva M. V.* (2012) Lexical Typology: Semantic Shifts in the Words Designating Qualities 'sharp' and 'blunt' [Leksicheskaja tipologija semanticheskih sdvigov nazvanij kachestvennyh priznakov 'ostryj' i 'tupoj']. Master's these, M., MGU (MSU).

7. *Kjuseva M. V., Ryzhova D. A., Holkina L. S.* 2012. Adjectives 'heavy' and 'light' on the typological background [Prilagatel'nye tjazhelyj i legkij v tipologicheskoj perspektive]. *Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"* [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"]. Moscow, pp. 247–255.
8. *Kopecka, A., & Narasimhan, B. (Eds.) 2012. Events of putting and taking: A crosslinguistic perspective. Amsterdam: Benjamins.*
9. *Krugljakova V. A., Rahilina E. V.* 2010. Verbs of Rotation: Lexical Typology [Glagoly vrashchenija: leksicheskaja tipologija]. *Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010"* [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Moscow, pp.241–248.
10. *Majid, A., Bowerman, M., Van Staden, M., Boster, J. S.* 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, 18 (2), pp. 133–152.
11. *Majid A., Levinson S. C.* (eds.) 2011. The senses in language and culture. The Senses & Society [Special Issue], 6(1).
12. *Majsak T. A., Rahilina E. V.* (eds.) 2007. Verbs of Aquamotion: Lexical Typology [Glagoly dvizhenija v vode: leksicheskaja tipologija]. M.: Indrik.
13. *Newman, J.* (ed.) 2002. The Linguistics of Sitting, Standing, and Lying. [Studies in Typological Linguistics 51]. Amsterdam/Philadelphia: John Benjamins.
14. *Newman, J.* (ed.) 2009. The Linguistics of Eating and Drinking. [Studies in Typological Linguistics 84]. Amsterdam/Philadelphia: John Benjamins.
15. *Pavlova E. K.* 2012. Qualities 'soft' and 'solid': towards a Typology [Kachestvennye priznaki 'mjagkij' i 'tverdyj' v tipologicheskoj perspektive]. Term paper. M., MGU (MSU).
16. *Rahilina E. V., Reznikova T. I., Karpova O. S.* 2010. Semantic Shifts in Adjectives: Metaphor, Metonymy and Rebranding [Semanticheskie perehody v atributivnyh konstruktsijah: metafora, metonimija i rebrending] in Lingvistika konstruktsij [Construction Linguistics]. M.: Azbukovnik. pp. 396–455.
17. *Rahilina E. V., Reznikova T. I.* 2013. Frame approach to lexical typology [Frejmovyj podxod k leksicheskoj tipologii]. *Voprosy jazykoznanija*, 2, 3–31.
18. *Tagabileva M. G.* 2011. Qualities 'full' and 'empty': towards a Lexical Typology [Kachestvennye priznaki 'pustoj', 'polnyj': k postroeniju semanticheskoj tipologii]. Term paper. M., MGU (MSU).
19. *Zaliznjak Anna A.* On the Notion of Semantic Shift [O ponjatii semanticheskogo perehoda]. *Komp'juternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010"* [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009"]. Bekasovo, 2009, pp. 107–112.

СВОЙСТВА НУЛЕВОЙ СВЯЗКИ В РУССКОМ ЯЗЫКЕ В СОПОСТАВЛЕНИИ СО СВОЙСТВАМИ ВЫРАЖЕННОГО ГЛАГОЛА[1]

Летучий А. Б. (alexander.letuchiy@gmail.com)

НИУ Высшая школа экономики, Москва, Россия

В статье рассматриваются свойства нулевой связки в русском языке. Целью исследования — выяснить, совпадают ли ограничения на употребление нулевой связки в функции глагола настоящего времени с теми, которые наблюдаются для выраженных глаголов. Для анализа привлекаются сложные предложения, ряд конструкций малого синтаксиса со временным значением, а также предложения с предикативами (категорией состояния). Как оказывается, связка в ряде случаев отличается по употреблению от выраженных глаголов, причём как в одну, так и в другую сторону: с одной стороны, есть контексты (например, адвербиальные придаточные или конструкция *не то чтобы X*), где может использоваться нулевая связка, хотя прочие (выраженные) формы настоящего времени запрещены или менее частотны. С другой стороны, конструкции со словосочетаниями типа *так и* или *на каждом шагу* сочетаются только с выраженными предикатами, не допуская нулевой связки. В результате мы приходим к выводу, что предложения со связкой — не просто вариант «полных» конструкций. Нулевая связка способна употребляться в контекстах, где употребление выраженных форм настоящего времени приводит к явному нарушению ограничений.

Ключевые слова: нулевая связка, полипредикативные конструкции, фраземы со значением времени, предикативы, время, наклонение

[1] В данной научной работе использованы результаты проекта «Корпусные технологии в лингвистических и междисциплинарных исследованиях», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2013 году.

PROPERTIES OF ZERO COPULA IN RUSSIAN IN COMPARISON WITH PROPERTIES OF NON-ZERO VERBS

Letuchiy A. B. (alexander.letuchiy@gmail.com)

National Research University Higher School of Economics, Moscow, Russia

The article is focused on the properties of the zero copula used as a present tense form in Russian. The principal aim is to check whether the zero copula can be used in the same contexts as non-zero verbs or if it has particular features. I find out that there are contexts where the zero copula is allowed while non-zero verbs in the present tense are prohibited; conversely, there are constructions which require a non-zero verb and prohibit the zero copulas. The former contexts include mainly biclausal constructions. The reason is that the zero copula lacks morphological tense and mood markers and does not apparently contradict any syntactic restrictions. The latter contexts, where the zero copula is prohibited belong to constructions with temporal meanings and constructions with predicatives. In the end I draw attention to the fact that constructions with the zero copula are not simply a reduced variant of some full structures, they have some particular rules of use which differ in some respects from those of non-zero verbs.

Keywords: zero copula, complex sentences, constructions with temporal meanings, predicatives, tense, mood

1. Введение

Среди наиболее общих свойств русского языка, которые проявляются в самых разных синтаксических контекстах и конструкциях, можно назвать возможность нулевого выражения предикатов и их актантов. Д. Вайс (1993) говорит даже об «очаровании нуля» для русского языка.

Часть из этих нулей широко распространены в языках мира. Например, к ним относится нуль, возникающий в результате сочинительного сокращения предиката (1), предиката и его подлежащего (2) или предиката и его дополнения (3):

(1) *Вася съел яблоко, а Петя — грушу.*

(2) *Сегодня Вася съел яблоко, а вчера — грушу.*

(3) *Сегодня яблоко съел Вася, а вчера — его брат.*

Однако есть другие типы нулей, которые значительно менее типологически распространены. К ним, в частности, относится нулевая связка при существительных в примерах типа (4), предложных группах (5) или наречиях (6):

(4) *Иван Иваныч — прекрасный человек* (ср. *Он* ***был*** *прекрасный человек*).

(5) *Я сейчас в Москве* (ср. *Я* ***был*** *в Москве).*

(6) *Автобус уже близко* (ср. *Автобус* ***был*** *уже близко*).

К таким же случаям мы относим нулевую форму глагола *быть* при предикативах, как в примере (7):

(7) *Приятно пройтись по улице* (ср. *Приятно* ***было*** *пройтись*).

Не все исследователи придерживаются такой точки зрения. Некоторые, например, А. П. Гвоздев (1973), говорят о «спряжении предикатива», считая, что по временам изменяется не связка, а сам предикатив. В этом случае в примере (7) усматривается не нулевая связка, а форма настоящего времени «*надо* + нулевой элемент». Однако мы считаем, что вполне допустимо трактовать отсутствующий в (7) глагол как нулевую связку. Как бы ни осмыслялось теоретически сочетание предикатива и глагола, всё же устройство словосочетания (неизменяемый по временам элемент + глагол с бытийной или близкой к ней семантикой) роднит сочетания с предикативами с представленными в (4)–(6).

Синтаксические условия возникновения нулей, эллипсиса и смежных процессов подробно рассмотрены (см., среди многих других, работу [Казенин 2009]). Наша задача будет несколько иной. Мы бы хотели понять, отличаются ли нулевые связки по сочетаемости с различными синтаксическими контекстами от выраженных форм настоящего времени. Как мы покажем, ответ на этот вопрос должен быть отрицательным.

2. Методологическое замечание

Необходимо сделать одно методологическое замечание. При анализе материала мы руководствуемся довольно простым определением контекста с нулевой связкой: нулевая связка усматривается там, где невыраженному глаголу в настоящем времени в прошедшем и будущем соответствуют выраженные формы прошедшего и будущего времени глагола *быть*, которые соединяют между собой участника (в т. ч. место, например, *там холодно*) и некоторое предицируемое на него свойство или динамическую ситуацию. В этом смысле мы не различаем контекстов типа *Васе холодно* (ср. *было, будет холодно*) *Вася дома* (ср. *был, будет дома*), *Вася — молод(ой)* (ср. *был, будет молод(ой)*)

и *Вася — учитель* (ср. *был, будет учитель / учителем*). Ключевым для нас является противопоставленность по временам нулевой и лексически выраженных форм. При этом мы не настаиваем ни на какой конкретной теоретической трактовке данного типа нулей.

В то же время мы не рассматриваем случаи типа *Я в Москву* или *На пол!* В первом случае аналогов данного примера с выраженным глаголом *быть* (**Я был / буду в Москву*). Во втором случае предложение выражает побудительное значение и семантически аналогично предложениям с повелительным наклонением (*Ложись на пол!*), а значит, говорить об аналогичных примерах с будущим или прошедшим временем здесь и вовсе не приходится.

3. Снятие для нулей ограничений, действующих для выраженных предикатов

В первой группе случаев для нулей не действуют некоторые ограничения, имеющие силу для выраженных предикатов. Прежде всего это касается полипредикативных конструкций и выбора в их составе временных форм глагола.

3.1. Нули в обстоятельственных придаточных

В русском языке выбор временно́й формы глагола в придаточном предложении не регулируется каким-либо единым правилом. Так, в актантных предложениях, как правило, временная форма подчинённого глагола, как правило, выражает относительное время (время по отношению к ситуации, выраженному в главном предложении):

(8) *Петя сказал, что хочет немного прогуляться.*

В предложении (8) форма настоящего времени *хочет* обозначает, что желание имеет место одновременно с актом речи (*сказал*) или, во всяком случае, отчасти совпадает с ним во времени.

Напротив, в обстоятельственных предложениях, как правило, мы имеем дело с абсолютной интерпретацией формы глагола:

(9) *Пока комнату убирали, я гулял по городу.*

(10) *Когда комната освободилась, я тут же заселился.*

В (9) ситуация в придаточном *комнату убирали* одновременна с ситуацией в главном предложении *я гулял по городу*, а в (10) ситуация в придаточном *комната освободилась* предшествует ситуации в главном *я заселился*. Однако в обоих случаях в придаточном используется прошедшее время: эта форма

обозначает предшествование моменту речи, не соотнося ситуацию с обозначенной в главном предложении.[2]

Однако для нуля данное правило смягчается. Возьмём для примера союз *пока*, который, как правило, вводит временные формы с абсолютной интерпретацией (ср. (10)). В придаточном предложении с *пока* (особенно в современной разговорной речи) в качестве предиката способна выступать нулевая связка:

(11) *Щенок,* ***пока дома****, писал около миски.* [система Google]

(12) *Утром,* ***пока муж дома*** *я обычно быстренько купалась и себя в порядок приводила.* [www.babyblog.ru/community/post/.../1724858]

(13) *Работал,* ***пока молодой*** *на шахте в забое, а как на пенсию вышел, осел на поверхности и трудовой стаж прерывать не собирался.* [www.donjetsk.com/.../28-glava-13.-tri-obraza-zhizni.htm]

(14) *??Работал,* ***пока может****, в шахте при заводе.*

Если нулевая связка, как считается в большинстве работ, обладает грамматическим значением настоящего времени, это означает, что настоящее время в примерах типа (11)–(13) интерпретируется как относительно (так, в (12) ситуация *муж дома* полностью или частично совпадает с ситуацией *я купалась*).

Информанты подтверждают данные примеров. Хотя мы не преследовали цель получить большую выборку ответов, уже опрошенные носители оценивают предложения типа *Пока муж дома, она готовила еду* лучше, чем *Пока муж играет в компьютер, она готовила еду*. Эти два предложения отличаются только тем, что в первом место предиката занимает нулевая связка, а во втором — полнозначный выраженный предикат, однако относительная интерпретация времени лучше для первого предложения.

Статистически были обработаны оценки предложений *Она пекла пироги, пока муж дома* vs. *Она пекла пироги, пока муж сидит дома*. Первое предложение оценивают как приемлемое 40 % информантов (131 человек из 327 опрошенных), второе — 13,5 % информантов (44 из 327 опрошенных). Хотя даже предложение с нулевой связкой принимается менее чем половиной носителей, ясно, что разница примерно в три раза между нулевой связкой и выраженной формой настоящего времени симптоматична (особенно в сочетании с тем, что в поисковых системах обнаруживаются примеры типа (11)–(13). Сочетание нулевой связки в придаточном предложении и формы прошедшего времени в главном явно не безоговорочно запрещено в современном русском.

---

[2] Разумеется, положение ситуации, обозначенной в придаточном, по отношению к ситуации в главном может фиксироваться с помощью вида глагола, но функция видовых форм выходит далеко за пределы настоящей работы.

Аналогичным образом ведёт себя причинный союз *поскольку*. Относительная интерпретация времени скорее возможна для нулевой связки (15), чем для выраженного глагола (16) (впрочем, для данного союза реальных примеров такого рода не найдено):

(15) *Поскольку муж дома, я решила приготовить котлеты.*

(16) *# Поскольку муж сидит дома, я решила приготовить котлеты*
(возможно только, если муж сидит дома в момент речи).

Наконец, самый частотный временной союз *когда* также допускает несовпадение времён, если в придаточном предложении — нулевая связка (ср. пример *Машина чуть хуже заводилась когда холодно* [forum.onliner.by/viewtopic.php?t=191193&p…] и недопустимое *Машина плохо ехала, когда идёт дождь*). Если в главном предложении выступает форма прошедшего времени, в придаточном не может быть лексически выраженной формы настоящего, но может быть нулевая связка.

Как следует интерпретировать данные примеры? Возможна точка зрения, при которой мы имеем дело с двумя разными правилами: для полнозначного глагола требуется абсолютная интерпретация времени, а для глагола *быть* в связочном употреблении возможна и абсолютная, и относительная, и при относительной интерпретации используется нулевая связка.

Однако эта интерпретация обладает избыточной объяснительной силой. На сегодняшний день не удалось найти контекстов, где ненулевые формы глагола *быть* вели бы себя особым образом. Скорее стоит считать, что особое поведение демонстрирует именно нулевая связка. По всей вероятности, это связано с тем, что нулевая связка, хотя и употребляется обычно в функции формы настоящего времени, не имеет временных показателей. Следовательно, в отличие от примера (14), в (13) не возникает конфликта между различными временными формами в главном предложении и обстоятельственном придаточном.

3.2. Нули в конструкции *чем …, тем …*

Конструкция *чем …, тем …*, где в обеих частях выступают прилагательные, наречия или предикативы в сравнительной степени, чаще всего требует параллелизма форм глаголов, как в примере (17). Если глаголы в двух частях выступают в разных временах, предложение зачастую становится неграмматичным (18)–(19):

(17) *Чем больше у него* ***становилось*** *денег, тем труднее ему* ***жилось****.*

(18) **Чем больше у него* ***становится*** *денег, тем труднее ему* ***жилось****.*

(19) **Чем больше у него* ***становилось*** *денег, тем труднее ему* ***живётся****.*

Безусловно, существуют случаи вроде *Чем больше он (сейчас) зарабатывает, тем лучше будет жить (в старости),* где сочетание настоящего времени в первой части и будущего — во второй не делает предложение неправильным (мы благодарны анонимному рецензенту за это замечание). Вероятно, точным будет следующее утверждение: **если обе части обозначают значение некоторого признака или степень проявления свойства участника** (например, в (17)–(19) — 'количество денег' и 'трудность жизни'), **возрастающую или убывающую со временем в зависимости друг от друга, то требуется совпадение времён**. В примере *Чем больше он (сейчас) зарабатывает, тем лучше будет жить (в старости)* дело обстоит по-другому: речь идёт не о том, что участник живёт или будет жить всё лучше и лучше, а о возможных вариантах его жизни. Признак 'качество жизни' не меняет своего значения в ходе рассматриваемой ситуации, поэтому совпадения времён не требуется. В то же время могут существовать и примеры, где совпадение времён не строго обязательно даже при семантике изменения свойства. В настоящее время мы не можем их объяснить.

Однако нулевая связка способна обойти и это ограничение. Предложения (20)–(21) вполне допустимы, несмотря на то, что нулевая связка обычно употребляется в функции настоящего времени, а в другой части предложения выраженный глагол стоит в прошедшем:

(20) *Потому что чем больше народу — тем страшнее всем становилось от обилия бродящих вокруг палестинских террористов.* [skoblov.livejournal.com/160528.html]

(21) *Норма, конечно, давалась в соответствии с возрастом, чем старше, тем она была больше.* [library.kat.kg/?author=1&paged=9]

Предельный случай представляет конструкция *чем дальше, тем …*, способная сочетаться с глаголом в любой форме:

(22) *Чем дальше, тем лучше он* ***понимал / понимает*** *происходящее.*

В данном случае объяснение должно быть таким же, как в предыдущем. Нулевая связка не вызывает конфликта, который возникает при выраженных различных формах времени, ср. (18)–(19). Впрочем, вполне естественно и то, что такие примеры редки: требование параллелизма двух частей, действующее для конструкции *чем …, тем …*, в таких примерах не соблюдается.

3.3. Нули в конструкции с союзом *чтобы*

Наконец, ещё один синтаксический контекст, где для нулевой связки смягчаются синтаксические ограничения, — это полипредикативная конструкция с союзом *чтобы*.

Союз *чтобы*, как известно, стоит особняком на фоне прочих русских союзов, поскольку требует прошедшего времени или инфинитива глагола в своей клаузе (настоящее время невозможно).

(23) *Я взял с собой несколько комплектов одежды, чтобы всегда* ***было*** *что надеть.*

(24) *Чтобы* ***понять*** *происходящее, нужно почитать что-нибудь по истории страны.*

На самом деле в примерах типа (23) стоит говорить не о формах прошедшего времени, а о формах сослагательного наклонения, поскольку союз *чтобы* включает в себя показатель *бы*. Вопрос о том, можно ли рассматривать *чтобы* на современном этапе как сочетание *что* + *бы*, широко обсуждался в лингвистической литературе (см. Brecht 1985, Dobrushina 2012). Мы считаем, что в определённой степени это возможно — в частности, как показывает Н. Р. Добрушина (2012), *бы* в составе *чтобы* может дублироваться другим, уже автономным показателем *бы*:

(25) *Я хочу, чтобы у него не было сомнений и было бы что сказать.*

Основных употреблений у данного союза два: целевое, обстоятельственное (*Я позвонил, чтобы мне объяснили, что происходит*) и актантное (*Я хочу, чтобы мне объяснили, что происходит*). Грамматические различия между ними касаются употребления инфинитива: в актантном употреблении при совпадении подлежащих сослагательное наклонение регулярно заменяется на инфинитив (*Я хочу тебе всё объяснить* и редко или невозможно *Я хочу, чтобы я тебе всё объяснил*), исключения носят весьма локальный характер. В целевом употреблении замены в большинстве случаев не происходит, ср. неграмматичное *Я позвонил объяснить ему, что происходит*. Однако в обоих основных употреблениях *чтобы* не сочетается с формами настоящего времени.

Помимо этого, *чтобы* употребляется и в других функциях и значениях — например, как частица с повелительным значением (26) или дискурсивный показатель несогласия с предыдущей репликой (27):

(26) *Чтобы быстро всё убрал со стола!*

(27) *Чтобы я — и поехал к чёрту на рога встречаться с этим Васей?!*

Однако и в этих функциях обязательно употребление форм сослагательного наклонения / прошедшего времени, а формы настоящего и будущего делают предложения неграмматичным:

(28) **Чтобы быстро всё уберёшь со стола!*

В ряде переносных употреблений *чтобы* сочетается с нулевым предикатом. Правда, нуль, возникающий в таких примерах, не относится к рассматриваемым нами нулевым связкам:

(29) *Чтобы без глупостей!*

Впрочем, суть данного явления — во многом та же, что и в «наших» случаях, анализируемых выше. Исходно существует конструкция *Без глупостей!* — если возводить её к какомулибо «полному» варианту с выраженным глаголом, то ближе всего находится императивная или инфинитивная конструкция (*Веди себя без глупостей!* или *Вести себя без глупостей!*)[3]. Если бы глагол был выражен, подчинить его *чтобы* было бы невозможно (повелительное *чтобы* с инфинитивом не сочетается). Однако при опущении глагольной формы остаток — предложная группа *без глупостей* — может быть подчинена *чтобы*, поскольку не имеет конфликтующих с *чтобы* показателей времени.

Требование инфинитива или сослагательного наклонения снимается в конструкции с показателем *не то чтобы*. Однако вариант (30) с формой прошедшего времени всё равно преобладает над (31), где время настоящее:

(30) *Я не то чтобы чего-то боялся.*

(31) *Я не то чтобы чего-то боюсь.*

Отметим, что здесь мы не касаемся вопроса об интерпретации формы прошедшего времени / сослагательного наклонения в примерах типа (30). В действительности возможны две интерпретации: (i) ‘Неверно, что я чего-либо боялся в прошлом’ и (ii) ‘Неверно, что я чего-либо боюсь сейчас’ (во втором случае сослагательное наклонение используется просто в силу ограничений, налагаемых *чтобы*, а не в связи с временем развёртывания ситуации).

Поиск в Национальном корпусе русского языка даёт 101 пример в ответ на запрос «*не то чтобы* + форма настоящего времени» и 476 — на запрос «*не то чтобы* + форма прошедшего времени». Аналогичные результаты мы получаем и при поиске в системе Яндекс на отдельные глаголы.

Картина и здесь меняется при подсчёте соотношения между формой прошедшего времени *был* и нулевой связкой. Запрос «*не то чтобы* + глагол *быть*» даёт 21 пример, а «*не то чтобы* + форма именительного падежа» — 992. Хотя последний запрос даёт много примеров, которые должны быть отсеяны (например, *Этот дом не то чтобы плохие строители строили*) вариант с нулевой связкой всё равно явно не уступает варианту с *быть*, а скорее даже частотнее.

Анализ затрудняется тем, что в примерах типа *Оно не то чтобы какая-нибудь завидная материя, — старина, из моды вышла; дорогого купить не могу, а жене нужно гостинца купить*. [И. И. Панаев. Раздел имения

[3] Впрочем, сопоставление с полным вариантом в данном случае сомнительно, поскольку зачастую такой вариант неграмматичен или неупотребителен.

(1850–1860)] *не то чтобы* можно рассматривать как модификатор не предложения или глагольной группы, а именной группы или, при другой трактовке, предикативной группы (именной или любой группы в позиции сказуемого[4]).

Казалось бы, если считать, что *не то чтобы* в подобных примерах модифицирует ИГ *какая-нибудь завидная материя*, а не сочетание ИГ и нулевой связки, вопрос о налагаемых *чтобы* требованиях снимается. Именные группы не имеют формы прошедшего времени, а значит, ИГ никак не может удовлетворить требования *чтобы*. Однако в действительности это вовсе не выводит из нашего рассмотрения примеры такого вида.

Действительно, если бы конструкция со словосчетанием *не то чтобы* налагала на зависимое *чтобы* жёсткое требование: содержать форму прошедшего времени, то с именными группами *не то чтобы* должно было бы быть попросту невозможно. Однако это не так. Следовательно, при отсутствии выраженного предиката требование смягчается. Впрочем, однозначно решить, что именно — именная группа или клауза, содержащая нулевую связку, — зависит от *не то чтобы* в примерах типа *не то чтобы какая-нибудь завидная материя*, по всей видимости, невозможно.

Если проанализировать статистику встречаемости примеров, где *не то чтобы* стоит в абсолютном начале предложения, перед целой клаузой, то статистическое различие между *быть* и нулевой связкой не столь показательно.

4. Дополнительные ограничения на нулевую связку

Выше мы рассмотрели случаи, когда нулевая связка не подпадает под действие синтаксических ограничений, релевантных для выраженных глаголов. Однако существуют и обратные случаи, где нулевая связка либо запрещена (синтаксический контекст требует выраженного глагола), либо малочастотна, в то время как лексически выраженные формы глагола допустимы.

4.1. Фраземы с временным значением

Первая группа таких случаев состоит из единиц так называемого малого синтаксиса, синтаксических фразем. Как оказывается, в некоторых группах (квази)синонимичных конструкций одни члены сочетаются с нулевой связкой (и вообще с невыраженным предикатом), другие же её запрещают. В наибольшей мере это характерно для конструкций, несущих значение времени, повторяемости и близкие к ним.

---

[4] Мы благодарны анонимному рецензенту за справедливое замечание о том, что наиболее характерна для сочетаний с *не то чтобы* позиция сказуемого. Каноническая актантная позиция выглядит сомнительно (ср. *Он купил не то чтобы дорогую материю*), однако позиция зависимого при глаголах типа *считать*, предполагающих свёрнутую предикацию, возможна (*Он считал его не то чтобы какой-то сволочью, но не очень хорошим человеком* = ‘Он считал, что он не то чтобы сволочь, но не очень хороший человек’).

Приведём в качестве примера единиц такого рода *всё ещё*, *до сих пор* и *так и*. Все три словосочетания очень близки по значению: их семантику можно приближённо представить следующим образом:

до сих пор / всё ещё / так и P = ‘Ситуация P имеет место в момент речи или в точке наблюдения. Говорящий или Наблюдатель считает, что более вероятным было, что к этому времени ситуация P закончится. В этой связи ему кажется неожиданным, что она продолжается’.[5]

Однако, несмотря на всю семантическую близость, именно по интересующему нас признаку три квазисинонима различаются. *До сих пор* и *всё ещё* сочетаются со структурами с нулевой связкой, тогда как *так и* требует выраженного глагола.

(32) *Он до сих пор / всё ещё / *так и в Москве?*

Очевидным образом, причина неграмматичности варианта с *так и* именно в отсутствии глагола, а не в семантике предложения. При переносе ситуации в прошлое и употреблении формы *был* грамматичны все три варианта:

(33) *Он до сих пор / всё ещё / так и был в Москве.*

Анонимный рецензент предложил нам следующее объяснение этого факта: «Адвербиал *так и* не сочетается с нулевой связкой, так как фонологически представляет собой клитику и присоединяется только к выраженной глагольной вершине». Однако, на наш взгляд, это объяснение имеет свои недостатки. Во-первых, ряд других единиц, близких по свойствам к *так и* — например, *едва* или *еле* — допускают невыраженную глагольную вершину (обнаружены примеры типа *Он едва в сознании, еле на ногах*), хотя также являются клитиками. Во-вторых, сама формулировка «присоединяется только к выраженной глагольной вершине» в нашем случае ведёт к логическому кругу. Действительно очевидно, что *так и* требует выраженного глагола, однако это не вытекает напрямую из клитического статуса (клитика не обязательно образует одно фонетическое слово именно со своим синтаксическим хозяином).

Зато другое различие между, казалось бы, семантически близкими словосочетаниями объяснить легче. Речь о единицах *на каждом шагу*, *то и дело* и *сплошь и рядом*. В этой группе, наоборот, два квазисинонима (*на каждом шагу* и *то и дело*) не сочетаются с нулевой связкой, а третий (*сплошь и рядом*) её допускает. Ср. следующие примеры, в которых *сплошь и рядом* не может быть заменено на *то и дело* или *на каждом шагу*:

(34) *Но, скажет демократическая интеллигенция, олигархи с высокой вероятностью сплошь и рядом— подонки, конечно.* [Сергей Доренко. Левые силы — перезагрузка (2003) // «Завтра», 2003.08.13]

---

[5] Компонент неожиданности может сниматься при определённых контекстных условиях.

(35) *Посмотрите на актеров крупных габаритов — это сплошь и рядом комедийные персонажи.* [Елена Светлова. Поколение XXL (2003) // «Совершенно секретно», 2003.09.01]

Все три единицы могут обозначать высокую частотность ситуации:

(36) *Если делалось это в спешке или не в подобающем рабочем настроении (а настроение в таких случаях вряд ли может быть хорошим) — на место сочного слова приходило случайное, бледное, а главное — сплошь и рядом ломался ритм.* [Аркадий Мильчин. В лаборатории редактора Лидии Чуковской // «Октябрь», № 8, 2001] (ср. возможное *то и дело ломался ритм*).

(37) *Ровно полжизни прожил во Франции, но то и дело сравнивал с Россией — единственной точкой отсчёта.* [Вадим Крейд. Георгий Иванов в Йере // «Звезда», № 6, 2003]

(38) *Столько человек за мной ухаживали, пытались добиться взаимности, и лишь один заботился, по-настоящему понимал, как я одинока, беззащитна, что родные далеко, муж погиб на фронте и на каждом шагу меня могут обидеть.* [Лидия Смирнова. Моя любовь (1997)][6]

Однако оказывается, что в примерах типа (34)–(35), где *сплошь и рядом* сочетается с нулевой связкой, оно просто имеет значение, которого нет у фразем *то и дело* и *на каждом шагу*. Это значение квантификатора, выбора подмножества из множества. В (34), например, речь идёт не о высокой частотности ситуации ‘олигархи — подонки’, а о том, что многие олигархи — подонки. Тем самым, *сплошь и рядом* выступает здесь как квантификатор со значением ‘многие X’.

4.2. Конструкции с предикативами

Другой блок синтаксических контекстов, где на нулевую связку могут налагаться строгие ограничения, составляют некоторые конструкции с предикативами. Напомним, что предикативами называются единицы, подобные *приятно* в *здесь приятно гулять* или *необходимо* в *необходимо, чтобы была разработана инструкция*. Стандартно предикативы могут иметь два зависимых: дативную именную группу с семантической ролью Экспериенцера или Субъекта оценки и сентенциальный актант.

Рассмотрим вначале уже знакомую нам конструкцию *чем …, тем …*, только на этот раз тот её вариант, где обе части содержат сравнительную степень предикативов. Предикатив *ясно* с большим трудом выступает в этой

[6] Здесь мы не рассматриваем тонкие семантические различия между данными конструкциями (отсылаем читателя к статье Рахилина, Летучий в печати).

конструкции, если при нём нет выраженного глагола (или, иначе говоря, есть нулевая связка), но гораздо легче — в сочетании с глаголом *быть* или *становиться:*

(39) ?*Чем яснее, что наш эксперимент провалился, тем хуже я себя чувствую.*

(40) *Чем яснее было, что наш эксперимент провалился, тем хуже я себя чувствовал.*

Поисковая система «Яндекс» находит 8 результатов на контекст *чем яснее, что* и около 35 — на *чем яснее становилось, что*. Естественно, вне конструкции *чем …, тем …* статистическое соотношение обратное: *стало / становилось ясно / ясно стало / становилось, что* встречается примерно в 2 млн. 100 тыс. примеров, а *ясно, что* — в 5 млн. (из них, тем самым, 2 млн. 900 тыс. без *стать* или *становиться*).

К сожалению, данный тест не даёт одинаковых результатов для всех предикативов. Например, *очевидно*, близкий по значению к *ясно*, чаще выступает в конструкции *чем …, тем …*с нулевой связкой, нежели с выраженным глаголом. Объяснение такому различию между близкими единицами на сегодняшний день не найдено.

Случай, содержательно близкий к *ясно*, можно наблюдать на примере предикатива *хорошо*. При данной лексеме сравнительно редко выражаются в одном предложении и сентенциальный актант (инфинитивный оборот), и Экспериенцер. В Корпусе найдено всего 10 таких примеров с нулевой связкой:

(41) — *И мне хорошо смотреть на вас и на свечи и вспоминать то, что можете вспоминать и вы…* [Юрий Бондарев. Берег (1975)]

С другой стороны, Корпус содержит 7 примеров такого рода с формой *было*:

(42) *Мне было хорошо стоять, глядеть на огонь, слушать его, подчиняясь его руке — и, может быть, от тьмы и внезапных высоких взлетов огня — немного страшно.* [Л. К. Чуковская. Спуск под воду (1949–1957)]

При этом для конструкции *мне хорошо* без сентенциального актанта вариант *мне было хорошо* встречается 128 раз, а *мне хорошо* с нулевой связкой — 735. Даже без учёта паразитических примеров, где *хорошо* не является предикативом (*мне хорошо известно …*) оказывается, что конструкция с нулевой связкой более чем в три раза превосходит конструкцию с *было*. Для контекстов типа *мне хорошо стоять* разница гораздо меньше — следовательно, предикатив *хорошо* с обоими выраженными актантами встречается скорее с выраженной, чем с невыраженной связкой.

Хотя предикативы по данным критериям ведут себя крайне индивидуально, некоторую интерпретацию предложить можно. Как показано в диссертации

А. А. Бонч-Осмоловской (2003), а также в работах [Zimmerling 2009], [Сай 2011], [Летучий 2012] предикативы зачастую демонстрируют непредсказуемые запреты на выражение актантов или затруднения в образовании сравнительной степени. Однако эти запреты наблюдаются именно в сочетании с нулевой связкой, которая для предикативов является основной. Как только предикатив присоединяет связку *быть* или полувспомогательный глагол, его идиосинкратические свойства проявляются слабее — например, для *ясно* облегчается участие в конструкции *чем …, тем …*, а для *хорошо* — употребляться с одновременно выраженными экспериенцером в дативе и инфинитивным оборотом (*Мне было хорошо там сидеть*).

5. Заключение

В нашей статье мы проанализировали некоторые свойства предложений с нулевой связкой в русском языке. Нашей целью было понять, проявляет ли нулевая связка все те же свойства, что выраженный глагол или демонстрирует особый набор свойств.

Хотя во многом свойства нулевой связки и выраженного глагола совпадают, наблюдаются и некоторые различия. С одной стороны, в ряде синтаксических контекстов ограничения на связку слабее, чем на формы настоящего времени выраженных глаголов. Это некоторые полипредикативные конструкции: с обстоятельственными придаточными, с союзом *чтобы*, а также конструкция *чем …, тем …* со сравнительной степенью. Обобщая, можно сказать, что ограничения на связку смягчаются в полипредикативных конструкциях. Это связано с тем, что связка не несёт морфологических показателей времени и наклонения, а значит, её характеристики не конфликтуют напрямую с требованиями *чтобы*, адвербиальных придаточных или конструкции *чем …, тем …*

Контексты, где ограничения на нулевую связку строже, чем на выраженный глагол, хуже поддаются объединению в естественные классы. Так, нулевой связки не допускает ряд конструкций со временным значением, причём не для всех их ясна причина такого запрета. Кроме того, для некоторых предикативов сочетаемость с нулевой связкой (при определённых синтаксических условиях или в сравнительной степенью) затруднена — а при добавлении выраженного глагола эти ограничения ослабевают.

Наш материал важен, прежде всего, потому, что позволяет сделать существенный общий вывод. Предложения с нулевой связкой — не просто неполный вариант аналогов с лексически выраженным глаголом *быть* и другими глаголами. Отсутствие морфологически выраженных категорий оказывает влияние на сочетаемость нулевой связки. В ряде случаев ограничения становятся более строгими, чем для выраженных глаголов, в других — смягчаются.

Литература

1. *Bonch-Osmolovskaja, Anastasia A.* Konstrukcii s dativnym subjektom v russkom jazyke: opyt korpusnogo issledovanija, Moscow, Moscow State University, 2003.
2. *Brecht, Ričard D.* (1985). O vzaimosvjazi meždu nakloneniem i vremenem: sintaksis časticy by v russkom jazyke. Novoe v zarubežnoj lingvistike. Vypusk XV. Sovremennaja zarubežnaja rusistika.
3. *Dobrushina, Nina R.* Subjunctive complement clauses in Russian. Russian Linguistics 36.2. 2012.
4. *Gvozdev, Alexander N.* The modern Russian literary language [Sovremennyj russkij literaturnyj jazyk], Moscow, Prosveshchenie, 1973.
5. *Kazenin K. I.* On some constraints on ellipsis in Russian [O nekotorykh ogranichenjiakh na Ellipsis v Russkom Iazyke]. Voprosy Jazykoznaniia, 3: 92–107, 2009
6. *Letuchiy A. B.* On some properties of sentential arguments in Russian [O nekotorykh svojstvakh sentencial'nykh aktantov v russkom jazyke]. Vorposy Jazkoznanija, 5: 57–87. 2012.
7. *Rakhilina Ekaterina V., Letuchiy Alexander B.* The initial stage of grammaticalization of constructions with the meaning of predicate plurality [Načal'naja stadija grammatikalizacii konstrukcij so značenijem glagol'noj množestvennosti]. In press.
8. *Saj, Sergei S.* "Dative subject" in constructions with predicatives ending on -o: a lexical dependency or a construction element [«Dativnyj subjekt» v konstrukcijax s predikativami na *-o*: prislovnaja zavisimost' ili component konstrukcii], Talk at the conference dedicated to the 50th anniversary of Saint-Petersburg typological school, Saint-Petersburg, 2011.
9. *Weiss, Daniel.* Die Faszination der Leere. Die moderne russische Umgangssprache und ihre Liebe zur Null. In: Zeitschrift für Slavische Philologie LIII/1993 (Beiträge zum XI. Internationalen Slavisten-Kongreß, Bratislava 1993), 48–82.
10. *Zimmerling, Anton.* Dative Subjects and Semi-Expletive pronouns // G. Zybatow, U. Junghanns, D. Lenertová, P. Biskup (eds.). Studies in Formal Slavic Phonology, Syntax, Semantics and Information Structure. Frankfurt am Main; Berlin; Bern; Bruxelles; New York; Oxford; Wien, 2009.

О ПРИЧИННОМ ЗНАЧЕНИИ СОЮЗА *А ТО*[1]

Левонтина И. Б. (irina.levontina@mail.ru)

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

Союзы *а то* и *а не то* неоднократно становились объектами лингвистического описания. В первую очередь исследователей интересовали смысловые различия между данными союзами и условия их взаимозаменимости. Кроме того, большое внимание уделено структуре многозначности этих союзов, особенно *а то*. При этом, как нам кажется, одно из интересных значений союза *а то* до сих пор не получило адекватного описания. Речь идет о значении, которое обычно описывается как причинное: *Сходи в булочную, а то хлеба нет; Пойдем домой, а то завтра рано вставать; Нет ли у тебя соли, а то у меня кончилась?* Оказывается, однако, что одной идеи причины здесь совершенно недостаточно. В докладе союз *а то* в причинном значении рассматривается на фоне, с одной стороны, различных значений данного союза, а с другой — иных причинных слов.

Ключевые слова: семантика, союзы, полисемия, причинность, импликатура, русский язык

ON THE CAUSATIVE MEANING OF THE RUSSIAN CONJUNCTION *A TO*

Levontina I. B. (irina.levontina@mail.ru)

Russian Language Institute (Vinogradov Institute), Russian Academy of Sciences, Moscow, Russia

The Russian conjunctions *a to* as well as *a ne to* '≈ or else' have repeatedly become objects of linguistic studies. First of all researchers were interested in semantic distinctions between these conjunctions and conditions of their interchangeability. Besides, much attention has been paid to the structure

[1] Работа выполнена при финансовой поддержке РФФИ, грант № 13-06-00403 «Контрастивное корпусное исследование специфических черт семантической системы русского языка», РГНФ, грант № 13-04-00307а «Подготовка второго выпуска Активного словаря русского языка» (2013–2015) и гранта НШ-6577.2012.6 для поддержки научных исследований, проводимых ведущими научными школами РФ «Разработка материалов для Активного словаря русского языка» (2012–2013).

of polysemy of these items, especially *a to*. Yet one of the interesting meanings of the conjunction *a to* seems not to have received an adequate description. It is the meaning which is usually described as causal: *Sxodi v bulochnuju, a to xleba net* 'Go to the baker's since we are out of bread';. *Pojdem domoj, a to zactra rano vstavat'* 'Let's go home because tomorrow we have to get up early tomorrow'; *Net li u tebja soli, a to u menja konchilas'* 'Do you have some salt, since mine is over?' Apparently, the idea of cause alone is absolutely insufficient. The paper addresses this causative meaning of *a to* contrasting it with other senses of the conjunction and other words of causation'

Key words: semantics, conjunctions, polysemy, causation, implicature, Russian

Союз *а то*, как и *а не то,* неоднократно становился объектом лингвистического описания [Санников 1989; Белошапкова 1970; Подлесская 2000; Собинникова 1969; Колосова 1980; Инькова-Манзотти 2000; Израэли 2000, Урысон 2008, 2010]. В первую очередь исследователей интересовали смысловые различия между данными союзами и условия их взаимозаменимости. Кроме того, большое внимание уделялось структуре многозначности этих союзов, преимущественно *а то*. Особое место занимают работы Е. В. Урысон, в которых сделана попытка выявить семантический вклад элементов *а*, *то, не* в семантику *а то* и *а не то*.

В настоящей работе ставится более скромная задача. Как нам кажется, одно из интересных значений союза *а то* до сих пор не получило адекватного описания. Речь идет о значении, которое обычно описывается как причинное:

Сходи в булочную, а то хлеба нет.

Пойдем домой, а то завтра рано вставать.

Нет ли у тебя соли, а то у меня кончилась?

Действительно, можно сказать, что отсутствие хлеба представлено здесь как причина того, что необходимо идти в булочную, а завтрашнее раннее вставание как причина решения идти домой (пример с солью, правда, несколько сложнее). В работе [Урысон 2008] читаем:

Союз *а то* имеет еще одну (разговорную) лексему. Она вводит указание причины [Белошапкова 1970; Санников 1989]. Ср.

(33) *Пойду, а то уже поздно* ['пойду, потому что уже поздно'].

(34) *Шестьдесят копеек оставила себе, шестьдесят отложила Лидии Афанасьевне, а то ей не на что было возвращаться* (Ю. Трифонов, пример В. А. Белошапковой).

Элемент *то* вводит здесь обозначение ситуации-причины. Вопрос о причине в разговорной речи вводится словом *что*, ср. *Что он такой расстроенный?, Что ты так торопишься?* и т. п. [Арутюнова 1980]. Если в вопросе причина обозначается словом *что*, то в утверждении она естественно маркируется словом *то*. Элемент *то* данной лексемы *а то* отсылает к последующему фрагменту (катафора). Союз *а не то* в таких случаях невозможен, ср. **Пойду, а не то уже поздно*. Причина в том, что в семантической структуре подобного высказывания нет компонента, к которому мог бы отсылать элемент *то* с отрицанием.

Понятно, что здесь не ставилась задача дать исчерпывающую экспликацию союза *а то*. Тем не менее, нельзя не отметить, что одной идеи причины совершенно недостаточно. Никак нельзя сказать:

**На улице лужи, а то шел дождь.*

**У него кашель, а то он промочил ноги.*

Между тем, если бы значение союза сводилось здесь к указанию на причину, такого запрета не было бы.

По мнению А. Израэли, *а то* здесь имеет следующее значение:

p *a to* q: q explains why the speaker thinks p or states p; q is the reason for p

Израэли отмечает, что «это исключительно иллокутивное *а то*, в отличие от других».

В самом деле, в большинстве примеров с причинным *а то* во второй части говорящий обосновывает то речевое действие, которое совершал в первой части, или ту мысль, которая была в ней высказана.

— Дядь, — сказал Генка, — мне домой надо. А то поздно. Отец будет ругать. [Юрий Коваль. Гроза над картофельным полем (1974)]

Говорящий поясняет, почему просится домой.

— Ты на ночь окно не запирай, а то душно.
[Ю. О. Домбровский. Леди Макбет (1970)]

Здесь говорящий также обосновывает свою просьбу.

То, что союз *а то* часто употребляется в контекстах, где он вступает в связь с иллокутивной компонентой первой клаузы, было отмечено еще в [Падучева 1985: 46]; рассматривался, в частности, пример *Где Иван, а то им начальство интересовалось*. Там же указывалось, что первая клауза поэтому часто является вопросом или императивом.

Особенно характерен приведенный выше пример: *Нет ли у тебя соли, а то у меня кончилась?* В этом случае ясно видно, что *а то* вводит именно пояснение, почему говорящий спрашивает. Ср. также следующий пример, где говорящий также поясняет, почему задает именно такой вопрос:

Квартира своя? Ванна есть? Гут. А то тут общая только. Будете возить её к себе мыться. Она мыться любит.
[Татьяна Толстая. Река Оккервиль (1983)]

Как будто, есть, впрочем, и такие примеры, которые такой интерпретации противоречат.

Возле её кровати лежали томики Ахматовой, Пастернака, Баратынского… Когда тётка умерла, библиотеку сразу же распродали. Предварительно брат и его жена вырвали листы с автографами. А то неудобно. [Сергей Довлатов. Наши (1983)]

Говорят, что хозяева «Пивоварен Ивана Таранова» ищут площадку поближе к центру — а то далеко возить «ПИТ» в Москву из Оренбурга. [Евгений Толстых. Пивка для рывка (2003) // «Совершенно секретно», 2003.09.01].

Один из дилеров BMW дал такое объяснение: «Мы просто решили давать» честные «цены, а то всё равно приходит человек в салон и видит совсем не то, что в рекламе». [Хасан Ганиев. Новости (2002) // «Автопилот», 2002.09.15]

Однако и в этих примерах нетривиальный говорящий не исчезает полностью. Можно сказать, что здесь представлена несобственно-прямая речь. В приведенном выше примере с автографами это хорошо видно. Если бы было сказано *Предварительно брат и его жена вырвали листы с автографами. Потому что <поскольку, так как> иначе неудобно продавать книги* оценка *неудобно* воспринималась бы как разделяемая самим говорящим. У Довлатова же очевидным образом имеется в виду, что родственники умершей, с точки зрения говорящего, совершенно ее не понимают, в частности это проявляется в том, что они уничтожают в книгах самое ценное, что в них есть — автографы авторов. *Неудобно* — это именно оценка родственников: для них автограф Ахматовой — не культурная ценность, а просто свидетельство того, что книга раньше принадлежала другому человеку. Подчеркнем еще раз: уже по приведенном примерам видно, что речь часто идет не собственно о мотивировке речевого акта, а об основаниях для предпринятых человеком действий, о которых он сообщает.

Здесь уместно провести некоторую аналогию.

Слово *почему* в качестве отдельной реплики в русском языке свободно употребляется, если спрашивающий интересуется причиной обсуждаемой ситуации, но не причиной самого речевого акта. Так, нормально:

— *Соль отсырела.*

— *Почему?*

Однако странно:

— *У тебя соль есть?*

— **Почему?* [в смысле, «Почему ты спрашиваешь?»]

Заметим, что формулировка *Почему ты спрашиваешь?* в данной ситуации возможна, неправильно именно изолированное *Почему?* При этом, скажем, английское *Why?* или немецкое *Warum?* свободно используются в подобных контекстах. По-русски же здесь надо использовать другую единицу: *А что?*

— *У тебя соль есть?*

— *А что?*

А что, в свою очередь не используется в вопросе об объективной причине[2]. Таким образом, как мы видим, и другие русские средства выражения причинного значения могут быть чувствительны к различию между причиной ситуации и мотивировкой речевого акта.

Тем не менее, толкование Алины Израэли также не во всех случаях объясняет ограничения на употребление обсуждаемого *а то*. Рассмотрим следующий пример:

Хорошо, что ты приехал, а то я так скучала.
[Алексей Варламов. Купавна // «Новый Мир», 2000]

Это абсолютно естественное и очень типичное предложение. Говорящий делает некоторое оценочное высказывание и затем обосновывает свою оценку. Теперь попробуем это высказывание немного изменить:

**Жаль, что он уехал, а то с ним было так весело.*

Очевидно, что эта фраза неудачна. А ведь на первый взгляд смысловое соотношение между частями совершенно такое же, как и в рассмотренных выше естественных примерах: во второй части говорящий обосновывает, почему делает высказывание, содержащиеся в первой части.

Аналогичным образом сомнительны или невозможны следующие фразы:

---

[2] О типах причинных значений и средствах их выражения в русском языке см. [Богуславская, Левонтина 2004].

**Не надо вырывать листы с автографами. А то с ними книги еще дороже.*
**Ищут площадку поближе к центру, а то это гораздо удобнее.*
**Не закрывай окно, а то воздух такой чудесный.*

Между тем, исходные фразы, рассмотренные выше, были совершенно нормальными. Очевидно, что причина здесь семантическая, и мы не можем считать, что адекватно описали значение *а то*, пока наша экспликация не будет объяснять, почему неудачны последние примеры.

Рассмотрим подробнее пару предложений:

Хорошо, что ты приехал, а то я так скучала.
[Алексей Варламов. Купавна // «Новый Мир», 2000]

**Жаль, что он уехал, а то с ним было так весело.*

Можно заметить, что в правильном примере во второй части речь идет, грубо говоря, о чем-то плохом, во всяком случае, о ситуации, которая сулит что-то нежелательное. Стремлением избежать нежелательного говорящий и объясняет высказывание, которое содержится в первой части. То же и в других правильных примерах (*а то опоздаем, а то душно, а то завтра рано вставать*). Это, так сказать, «негативная» мотивировка. Если же мы имеем дело с «позитивной» мотивировкой, *а то* оказывается неуместным (**а то так удобнее <выгоднее, проще, веселее>*).

Рассмотрим еще две аналогичных пары примеров:

Можете перезвонить? — а то мне очень дорого. [Андрей Волос. Недвижимость (2000) // «Новый Мир», 2001]

**Давайте я перезвоню. А то мне это бесплатно.*

Слушай, ты не можешь найти мне работу? Убираться в квартире. У каких-нибудь новых русских, только не опасных. А то мне в театрах ничего почти не платят. [Катя Метелица. Фруска (1997) // «Столица», 1997.06.17]

*Слушай, ты не можешь найти мне работу? Убираться в квартире. У каких-нибудь новых русских, только не опасных. *А то за уборку очень хорошо платят.*

Как мы видим, и здесь в правильных фразах с *а то* говорящий ссылается на негативный аргумент, а в неудачных — на позитивный.

Необходимо пояснить, что в самом по себе обстоятельстве, на которое ссылается говорящий, может не быть ничего плохого, однако в контексте конкретной ситуации оно сулит определенные осложнения. Так, вполне возможно: *Дай мне путеводитель, а то мне на следующей неделе в Венецию ехать*. В этом примере, предложенном рецензентом «Диалога», можно также представить себе такие «мотивировочные части»: *а то мой очень тяжелый <бестолковый>*, но не такую: **а то он у тебя очень толковый*. В удачных фразах здесь

подразумевается, что без путеводителя или с другим путеводителем поездка будет хуже. В неудачной же фразе мотивировка чисто позитивная.

Ср. также следующий пример:

Квитанция, товарищ директор, в моём паспорте под обложкой, на ремонт велосипеда, уж пять дней пропущено, а то завтра опять выходной.
[Ю. О. Домбровский. Факультет ненужных вещей, часть 1 (1978)]

В принципе, если завтра выходной, то это скорее хорошо. Однако здесь имеется в виду, что в выходной мастерская будет закрыта и снова не удастся забрать велосипед из ремонта. Это нежелательное обстоятельство, на которое говорящий ссылается, мотивируя необходимость отправиться в мастерскую срочно.

Стоит также отметить, что а то проспективно: этот союз предполагает не просто некую неприятную ситуацию, а именно возможность появления или продолжения такой ситуации в будущем. Поэтому хорошо: *Долго нам еще ждать? А то я уже начинаю сомневаться, что хочу покупать это платье*, но плохо: **Идем отсюда. А то я передумала покупать это платье.*

Идея чего-то нежелательного в перспективе сближает «причинное» *а то* с одним из других значений этого союза. Обратимся еще раз к классификации Е. В. Урысон:

2.3. Контексты угрозы (Отдай машинку, а то <а не то> маме скажу)

Данный тип контекстов иллюстрируется примерами типа

(30) *Придется дать кролику капусты, а то <а не то> он убежит.*

(31) *Отдай машинку, а то <а не то> маме пожалуюсь.*

(32) *Держу вас только из уважения к вашему почтенному батюшке, а то бы вы у меня давно полетели со службы* (А. П. Чехов, [Санников 1989]) [здесь возможен и союз а не то].

<…> Действительно, императив Р выражает желание говорящего, чтобы имела место ситуация P^1 ‘ты отдаешь мне машинку’. Семантическая структура высказывания (31) содержит следующий фрагмент:

(31а) (i) [цель речевого акта] *‘ты отдаешь мне машинку’.*
(ii) *‘если ты не отдашь мне машинку, я пожалуюсь маме’.*

В более эксплицитном виде этот смысл выражен в примерах:

(31б) *Отдай машинку. Если не отдашь* (Р)*, пожалуюсь маме.*

(31в) *Отдай машинку. Не отдашь* (Р)*, пожалуюсь маме.* [Урысон 2008]

Очевидно, что термин «угроза» используется здесь в качестве ярлыка и понимается широко. Например, в случае с кроликом и капустой речь не идет о речевом акте угрозы. Во многих случаях имеет место не угроза, а предупреждение: если не сделать чего-то, последствия могут быть неприятными.

В этом случае существенно, что могут использоваться и союз *а то*, и союз *а не то*. Их синонимами здесь будут также такие единицы, как *иначе, или, в противном случае* (разумеется, между ними есть и свои тонкие различия).

Приведем еще несколько примеров, в которых союз *а то* фигурирует в контексте угрозы/предупреждения:

«Давай чертежи, а то я никуда не поеду!» — к этому сводилось требование Андрея. [Анатолий Азольский. Лопушок // «Новый Мир», 1998]

Гениев много не надо. А то их будут сажать и расстреливать.
[Людмила Улицкая. Казус Кукоцкого [Путешествие в седьмую сторону света] *// «Новый Мир», 2000]*

Только не давай мамке. А то она всё порвёт. [Борис Екимов. Пиночет (1999)]

Сейчас Алексей Петрович умоется, приведёт себя в порядок; Мамочка сходит проверить, не напачкал ли там, а то опять соседи заругают; а потом и кушинькать! [Татьяна Толстая. Ночь (1983)]

смотрите потом не перекормите его, а то он уснёт на уроке: [Наши дети: Подростки (2004)]

Мама сказала — поднимись к ним, отнеси ему что-нибудь. А то они там опять все напьются и забудут про него. [Андрей Геласимов. Жанна (2001)]

А можно, я чуть-чуть гостинцев попробую? — Можно, — усмехнулся Дед Мороз. — Только не увлекайся, а то живот заболит.
[Юрий Макаров. Про зайца // «Мурзилка», 2001]

Попрощайтесь и идите к выходу. А ещё лучше — бегите, а то, не ровен час, заболеете. На вас смотреть холодно.
[Вера Белоусова. Второй выстрел (2000)]

Как мы видим, во всех этих случаях *а то* можно заменить не *а не то*. Теперь рассмотрим такой пример:

Иди в дом, а то холодно, замерзнешь.

Это абсолютно естественный пример, и а то употреблено в нем очень уместно. Попробуем, однако, разделить эту фразу на две:

*Иди в дом, а то <*а не то> холодно.*

Иди в дом, а то <а не то> замерзнешь.

Теперь видно, что в первом случае представлено наше «причинное» значение, а во втором — значение предупреждения, и, как это и должно быть, во втором случае возможна замена на *а не то*, а в первом невозможна. Тем не менее, при таком сопоставлении хорошо заметно, что *а то* в двух случаях выражает похожую идею. В первой части говорится о том, что надо совершить определенное действие, во второй — указывается на что-то плохое, что будет иметь место, если этого не сделать. Только в случае *а то замерзнешь* неприятное последствие, которого желательно избежать, указано прямо, а во втором мысль о его возможности является прагматической импликатурой (сообщается, что холодно, а слушающий, естественно, понимает, что он может замерзнуть).

Аналогично:

— Надо остановку делать, — сказал он. — Чай надо пить, а то голова чего-то болит… [Юрий Коваль. Лабаз (1972)]

Если бы было сказано *а то голова заболит или* а то *голова не пройдет*, это было бы значение предупреждения, и *а то* можно было бы заменить на *а не то* или, скажем, *иначе*. В данном же случае это невозможно. Потому что здесь идея, что будет иметь место неприятное последствие в виде продолжения головной боли, не выражена прямо, а представлена в качестве импликатуры.

— Прочти кусочек, а? А то у меня руки в краске. — Какой кусочек? — спросил Андрей. — Любой. — Тогда, — сказал Андрей, — я с того места начну, где сам читаю. [Виктор Пелевин. Желтая стрела (1993)]

Здесь импликатура также очевидна: книга может быть испорчена (поскольку у говорящего руки в краске, так что если он возьмет ее в руки, то останутся пятна). Естественно, **А не то у меня руки в краске* невозможно.

Таким образом, логически, а вероятно, и исторически, «причинное» значение *а то* является развитием «угрозительно-предупредительного». В данной работе мы не ставим задачи проследить историю формирования рассматриваемого значения, отметим только, что новым оно не является. Приведем всего один пример:

Было уже двадцать минут третьего, а учителя истории не было ещё ни слышно, ни видно даже на улице, <…> — Кажется, Лебедев

нынче не придёт, — сказал Володя, отрываясь на минутку от книги Смарагдова, по которой он готовил урок. — Дай бог, дай бог… а то я ровно ничего не знаю…: однако, кажется, вон он идёт, — прибавил я печальным голосом. [Л. Н. Толстой. Отрочество (1854)]

Теперь обратимся к вопросу о том, каково место причинного *а то* среди других средств выражения причинного значения. Если посмотреть на приведенные выше примеры, можно заметить, что в большинстве случаев «причинное» *а то* не очень свободно заменяется на классические причинные союзы *потому что, так как, поскольку, ибо*.

Сходи в булочную, а то <поскольку, потому что> хлеба нет.
Пойдем домой, а то <так как> завтра рано вставать.
Нет ли у тебя соли, а то <?поскольку> у меня кончилась?

Даже там, где причинные союзы возможны, полученные фразы не тождественны фразам с *а то*. В них причинно-следственные отношения выражаются гораздо более эксплицитно, более форсированно.

Конечно, в *а то* есть идея причины-основания, но место ее иное, чем в причинных союзах.

Как кажется, в определенном отношении *а то* напоминает другое интересное слово — *ведь*. Обычно считается, что *ведь* выражает апелляцию к общему фонду знаний говорящего и слушающего. Это не объясняет, однако, многие контексты *ведь*, прежде всего те, что в работах С. В. Кодзасова и К. Бонно называются *ведь* прозрения (*А ведь это Петя!*) [Бонно, Кодзасов 1998]. Кроме того, рассмотрим один из самых типичных контекстов пояснения. Человек рассказывает, например, о встрече с немцем и поясняет: *А я ведь в школе немецкий учил*. Адресат раньше мог этого и не знать. Более того, такое высказывание может быть сделано и в беседе со случайным собеседником. Скорее в основе *ведь* лежит другая идея: ‘Я считаю, что нужно это принять во внимание для правильного понимания ситуации’ [см. Левонтина 2005].

Как кажется, в этом смысле *а то* похоже на *ведь*: оба слова вводят какое-то сообщение, которое должно, с точки зрения говорящего, помочь адресату понять, почему он сделал свое высказывание или почему сделал то, о чем говорит в своем высказывании. Не случайно во многих наших примерах *а то* можно заменить на *ведь*:

Пойдем домой, а то <ведь> завтра рано вставать.
Иди в дом, а то замерзнешь <замерзнешь ведь>.

Однако *ведь*, естественно, свободно употребляется и в тех случаях, где говорящий приводит «положительный» аргумент и *а то* поэтому невозможно:

**Давайте я перезвоню. А то мне это бесплатно.*
Давайте я перезвоню. Мне ведь это бесплатно.

Итак, как мы старались показать, в значении причинного в широком смысле *а то* есть следующие идеи: (1) *а то* присоединяет часть высказывания, в которой говорящий поясняет, зачем он делает свое высказывание или зачем он сделал то, о чем сообщил в первой части; причем поясняет следующим образом: (2) он указывает на некоторое обстоятельство, которое обусловливает неприятные последствия, которые могли бы иметь место, если бы говорящий этого не сделал.

Литература

1. *Белошапкова В. А.* Предложения альтернативной мотивации в современном русском языке // Исследования по современному русскому языку. М.: Издательство Московского университета, 1970. С. 13–29.
2. *Богуславская О. Ю., Левонтина И. Б.* Смыслы ‘причина’ и ‘цель’ в естественном языке // ВЯ 2004, № 2.
3. *Бонно К., Кодзасов С. В.* Семантическое варьирование дискурсивных слов и его влияние на линеаризацию и интонирование (на примере частиц *же* и *ведь*) // Дискурсивные слова русского языка: опыт контекстно-семантического описания. М., 1998.
4. *Israeli A.* The meaning and polysemy of the alternative conjunction *a to*. Manuscript.
5. *Inkova-Manzotti O.* Encore sur la conjonction Russe *A TO*. http://www.academia.edu/1381390/Encore_sur_la_conjonction_russe_a_to
6. *Колосова Т. А.* О сигналах неразвернутости некоторых имплицитных сложных предложений // Синтаксис предложения. Калинин. 1980.
7. *Левонтина И. Б.* Частица *ведь*: загадки сочетаемости (On the co-occurrence of one Russian particle) // Труды международного семинара Диалог 2005 по компьютерной лингвистике и ее приложениям. Том 1. М., «Наука», 2005.
8. *Падучева Е. В.* Высказывание и его соотнесенность с действительностью. М., Наука, 1985.
9. *Подлесская В. И.* ИНАЧЕ, А ТО, А НЕ ТО: резумптивные союзы как способ выражения отрицательного условия // Сложное предложение: традиционные вопросы теории и описания и новые аспекты его изучения. Вып. 1. М., 2000.
10. *Санников В. З.* Русские сочинительные конструкции. М.: «Наука»,1989.
11. *Собинникова В. И.* Сложные предложения с союзом *а то* в русских и украинских говорах // Материалы по русско-славянскому языкознанию. Т. 4. Воронеж. 1969.
12. *Урысон Е. В.* Союзы *а то* и *а не то*: почему в некоторых контекстах они синонимичны. http://www.dialog-21.ru/digests/dialog2008/materials/html/82.htm
13. *Урысон Е. В.* Составные союзы А ТО и А НЕ ТО: возможности семантического композиционального анализа// ВЯ, 2010, No. 1. С. 61–73.

References

1. *Beloshapkova V. A.* Predlozheniya al'ternativnoi motivacii v sovremennom russkom yazyke // Issledovaniya po sovremennomu russkomu yazyku. M.: Izdatel'stvo Moskovskogo universiteta, 1970. pp. 13–29.
2. *Boguslavskaya O. Yu., Levontina I. B.* Smysly prichina i cel' v estestvennom yazyke // VYa 2004, 12.
3. *Bonno K., Kodzasov S. V.* Semanticheskoe var'irovanie diskursivnyh slov i ego vliyanie na linearizaciyu i intonirovanie (na primere chastic *zhe* i *ved'*) // Diskursivnye slova russkogo yazyka: opyt kontekstno-semanticheskogo opisaniya. M., 1998.
4. *Israeli A.* The meaning and polysemy of the alternative conjunction *a to*. Manuscript.
5. *Inkova-Manzotti O.* Encore sur la conjonction Russe *A TO*. http://www.academia.edu/1381390/Encore_sur_la_conjonction_russe_a_to
6. *Kolosova T. A.* O signalah nerazvernutosti nekotoryh implicitnyh slozhnyh predlozhenii // Sintaksis predlozheniya. Kalinin. 1980.
7. *Levontina I. B.* Chastica*ved'*: zagadkisochetaemosti (On the co-occurrence of one Russian particle) // TrudymezhdunarodnogoseminaraDialog 2005 pokomp'yut ernoilingvistikeieeprilozheniyam. Tom 1. M., «Nauka», 2005.
8. *Paducheva E. V.* Vyskazyvanie i ego sootnesennost› s deistvitel›nost›yu. M., Nauka, 1985.
9. *Podlesskaya V. I.* INACHE, A TO, A NE TO: rezumptivnye soyuzy kak sposob vyrazheniya otricatel›nogo usloviya // Slozhnoe predlozhenie: tradicionnye voprosy teorii i opisaniya i novye aspekty ego izucheniya. Vyp. 1. M., 2000.
10. *Sannikov V. Z.* Russkie sochinitel›nye konstrukcii. M.: «Nauka»,1989.
11. *Sobinnikova V. I.* Slozhnye predlozheniya s soyuzom *a to* v russkih i ukrainskih govorah // Materialy po russko-slavyanskomu yazykoznaniyu. T. 4. Voronezh. 1969.
12. *Uryson E. V.* Soyuzy *a to* i *a ne to*: pochemu v nekotoryh kontekstah oni sinonimichny http://www.dialog-21.ru/digests/dialog2008/materials/html/82.htm
13. *Uryson E. V.* Sostavnye soyuzy A TO i A NE TO: vozmozhnosti semanticheskogo kompozicional'nogo analiza// VYa, 2010, No. 1. pp. 61–73.

ЦИТИРОВАНИЕ В УСТНОМ ДИСКУРСЕ: ИНТОНАЦИЯ КАК СРЕДСТВО ИНТЕГРАЦИИ

Литвиненко А. О. (alla1978@gmail.com)

Московский государственный университет
имени М. В. Ломоносова, Москва, Россия

Ключевые слова: цитирование, чужая речь, прямая речь, косвенная речь, полупрямая речь, интонация, просодия

REPORTED SPEECH IN SPOKEN DISCOURSE: INTONATION AS A MEANS OF INTEGRATION[1]

Litvinenko A. O. (alla1978@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

I discuss typical intonation patterns in Russian reported speech constructions, based on the data from the Prosodically Annotated Corpus of Spoken Russian which consists of 4 experimental subcorpora of Russian spoken discourse (the current version of the corpus is available on the website http://spokencorpora.ru/). More than 400 occurrences of reported speech of different types (direct speech, indirect speech, semi-direct speech) have been analyzed. I have attempted to show that (i) intonation patterns in preceding framing clauses (falling tone in main phrasal accent, rising tone in main phrasal accent and absence of main phrasal accent) correspond to the type of the reported speech (direct, indirect and semi-direct, accordingly); (ii) however, this correspondence is more a tendency than a cause-and-effect relationship; (iii) there are some typical patterns in semi-direct speech that use 'mixed' intonation in order to keep the 'original' illocutionary meanings and to integrate the reported speech into the following context as much as possible: the *list pattern* and the *head-tail-pattern*.

Keywords: reported speech, direct speech, indirect speech, semi-direct speech, intonation, prosody

[1] This research is supported by Russian Foundation for Humanities, grant #12-04-00258.

1. Introduction. Reported speech as an ambiguous phenomenon

Reported speech in spoken discourse presents most interesting object for linguistic research as it appears on the border between two different discourses, the main one and the one that is being reported, or rather, re-created (see Tannen 1989; Holt 1996; Klewitz, Couper-Kuhlen 1999; Litvinenko et. all 2009; Litvinenko 2011). The speaker needs both to recreate the 'other' discourse as close to verbatim as possible and to integrate it into current discourse as well as possible, at the same time; we consider this ambiguity to be the main reason for the reported speech types' variety. Thus, *direct speech* and *indirect speech* are the prototypical models that speaker uses to prioritize one or the other of these two opposite objectives; for anything that lies in-between, we will use the term *semi-direct speech* (please refer to Litvinenko et. al 2009 and Litvinenko 2011 for detailed analysis of reported speech types). Typical reported speech construction consists of two parts; one is the reported speech itself, the other is the inquit of some kind (that can be a proper framing clause, a discourse marker, or both); the inquit can also be absent with some types of reported speech.

As opposed to written text, speakers use intonation as the most complex and the most powerful means to both 'recreate' the original's tone and attitude, and to smoothen the transition between discourses, to express the speaker's own feelings and evaluations in relation to the discourse being reported. Speakers change voice quality, tempo, pitch and loudness to express several things simultaneously:

- the very fact that some part of what they say does not belong to them, but is being reported;
- the intentions and emotions, as well as the illocutionary intent of the 'original' speaker;
- the speakers' attitude to what they report;
- the connection between what is being reported and the current discourse itself.

There are two main points that are most interesting from this point of view; these are the points where the reported speech 'meets' the main (current) discourse 'first-hand'. One of these points of interest is the framing clause; the other is the reported speech itself, but especially its right border (the last elementary discourse unit, or EDU[2]). Both possess some very interesting prosodic characteristics.

This paper is based on the data from Prosodically Annotated Corpus of Spoken Russian, that consists now of 4 experimental subcorpora of Russian spoken discourse, more than 4 hours of sound in total (children's Night Dream Stories collection (NDS), and 3 adults subcorpora: Stories about Presents and Skiing (SPS), Siberian Lifestories collection (SLS) and Funny Lifestories collection (FLS)). Night Dream Stories collection is published; the other 3 subcorpora are in various states of development. The current version of the corpus is available on our website http://spokencorpora.ru/. On the site, you can also find basic information on the general principles our transcription is based

[2] The elementary discourse unit (EDU) is the key unit of our discourse transcription system. The prototypical EDU is a single clause that is also a single intonation unit. It is pronounced in one breath and expresses one situation. See also Kibrik, Podlesskaja, Korotaev 2009 for details.

on, corpora descriptions, etc. This paper uses simplified transcription (only main phrasal accents; pauses without precise length; no marking for sound prolongation, except in the places where it is needed specifically; no aspiration and various stops marked). There are 472 occurrences of reported speech in the corpus that were analyzed for this research.

2. Common intonation patterns in framing clause

Preceding framing clause is the main means to introduce reported speech into current discourse. In our corpora it is used in 88% of the occurrences, both in children's and adults' narrative. Postpositional or incorporated framing clause is very rare; in most of the other cases, the framing clause is not used at all, and the reported speech is marked either by heavy intonation changes (so-called 'doing voices') or by special discourse markers (or both). For our research, we used only the occurrences with the preceding framing clause, which make more than 400 examples (150 in children's stories and about 260 in adults' ones).

As far as the framing clause is concerned, there are three main intonation patterns that are used to introduce reported speech of different types: ***falling tone*** in the main accent, ***rising tone*** in the main accent, and ***no accent*** at all in the framing clause (basically, in this case the framing clause and the first EDU of the reported speech form one intonation unit). There are also some cases of ***level tone***, which show more or less the same tendencies as the falling tone. Probably that means that level tone should be considered an allophone of the latter; however, the total quantity of such occurrences is too small to make any positive conclusions.

In the *Table 1* we can see the total statistics for using different tones in the framing clause before different types of reported speech. As the frequency of usage for these types depends heavily on narrative genre and speakers' age, and some types are used more often that the others, we give here the statistics per 100 occurrences in each type.

Table 1. Tones in framing clause before reported speech (per 100 occurrences in type)

	Direct	%	Semi-direct	%	Indirect	%	Total	%
–	12	49.0%	8	31.2%	5	19.8%	25	8.3%
\	51	**42.1%**	37	30.5%	33	27.3%	122	40.6%
/	4	7.2%	12	19.0%	45	**73.8%**	61	20.4%
No accent	32	34.8%	43	**47.1%**	17	18.1%	92	30.7%
Total	100		100		100		300	

The ***falling tone*** in the framing clause corresponds mostly with the direct speech (42%), but is also quite frequent with the semi-direct speech (30.5%). Its frequency decreases slightly more with the indirect speech (27%). We can also see that it is the most frequent tone used to introduce reported speech.

(1) FLS, #40
···· *A Elizaveta Petrovna \skazala:*
And Elizaveta Petrovna said
··· *«Eta malenkaja zapytaja ··· ochen' mnogo \znachit.*
This small comma very much means
·· *Ochen' \mnogo.»*
Very much.
(And Elizaveta Petrovna said, "This small comma is very important. Very important.")

(2) SLS, #5
···· *Potom nam /pozvonili,*
Then us.Dat they.called
·· *\skazali,*
they.told
chto «V gorode idyot sil'nyj /dozhd',
that In city.Loc goes strong rain
··· *vam luchshe /\svalivat' ottuda¡!»*
you.Dat better to.get.away from.there
(Then [they] called us, said that "It's raining heavily in the city, you better get away from there!")

(3) SPS, #R1-8
··· *muzhik \reshil,*
the.guy decided
chto ne \stoit pokupat' etu mashiny,
that not worths to.buy this car
slishkom uzh \dorogo.
too emph.part expensive
(The guy decided that he should not buy this car, it cost too much.)

However, the reasons for using falling tone in such contexts can be different. In some cases, its purpose is indeed to mark the border between two different discourses, like in the example (1) with direct speech. Sometimes it used in the same way before semi-direct speech and even before indirect one too, as the speaker most probably rearranges discourse on the fly, changing strategy. Such obvious cases are marked in our transcription system with a colon, and they are indeed most frequent before the direct speech. In other cases, falling tone is obviously used automatically as an adaptive tone before a rising one, like in the example (2). The most obvious reason for the falling tone on *skazali* is the necessity to make a fall before the rising tone in the next EDU, on the word *dozhd'*. There are contexts, though, where the purpose of falling tone usage is unclear, as in the ex. (3), where the speaker uses a series of falling tones, which can be interpreted either as a case of slight emphasis or as some personal preference of the speaker.

The ***rising tone*** corresponds mostly with indirect speech (almost 74%); take the example (4) where we can see a typical case of it, with rising tone in the framing clause and falling one in the reported speech itself.

(4) SPS, #R1-3
····· *i /–skazal,*
and he.said
·· *chto stoit ona ··· \dorogo.*
that costs it.Fem expensive
(... and [he] said that it cost much.)

The rising tone is almost never used with direct speech, and the cases of semi-direct speech with a rising tone in the framing clause usually follow indirect intonation pattern in the reported speech itself (see section 3 below).

The ***absence of any accent*** in the framing clause corresponds more with the semi-direct speech (47%), but is also frequent before the direct speech (almost 35%), as in the examples (5) and (6) below.

(5) FLS, #5
On govorit
he says
«/\Zdravstvujte,
Hello
ja redactor /gasety.»,
I editor newspaper.Gen
·· *kakoj-to voobshche \neponyatnoj,*
some.Gen totally unknown.Gen
(He says, “Hello, I am an editor in a newspaper”, some totally unknown one)

(6) FLS, #7
·· *i kto-to menya \oklikajet,*
and somebody me.Acc calls
kakoj-to \paren’,
some lad
on krichit
he shouts
·· *«/\Devushka@!*
young.girl
·· */ Devushka@,*
young.girl
chto vy /delaete?!»
what you.Pl do.Pres.Pl
(... and someone calls out to me, some young man, he shouts, “Miss! Miss, what are you doing?!”)

This is a common occurrence in the contexts where the first EDU of the reported speech is short, and the phrasal accent falls on the first content word (not counting conjunctions, particles and other discourse markers). That is the case for 71% of such contexts. One of the typical examples of such short EDUs in the beginning of the reported speech is a vocative expression that forms a single EDU, like in the example (6). It also often occurs with exclamations and emphatic questions.

3. Common intonation patterns in reported speech. Mixed intonation in semi-direct speech

Of all the prosodical variety that can be found in reported speech depending on various illocutionary meanings, we are interested in those patterns that contribute to reported speech being properly integrated into context.

In this regard, 'classic' direct and indirect speech occurrences are simple cases. Indirect speech is not prosodically and illocutionary independent, and therefore is usually pronounced as a part of a typical polypredicative construction, as in the example (4). Direct speech, on the opposite, is prosodically and illocutionary independent, and as such uses typical intonation patterns (e.g. for a statement, question, exclamation, etc.), as in examples (1) and (6).

However, semi-direct speech provides a broad range of different 'mixed' intonation patterns, where the 'original' intonation combines with the one that the speaker uses to integrate the reported speech into the context and/or to express his/her attitude and emotions concerning the text that is being reported. In this paper, we would like to demonstrate two typical patterns, or strategies, that are often used with semi-direct speech in our corpus.

The first one is the ***list pattern***. This scheme uses a series of identical or similar accents to convey the idea of retelling/reporting someone else's words. Rhythmical organization of reported speech has been often described as typical (e.g. Couper-Kuhlen 1999; Levontina 2010); however, here we can observe not only a specific temporal structure, but a series of similar accents. Such series of accents can be combined with additional meanings, e.g. exclamations, emphasis, surprise, etc. In the example (7) below, the speaker retells her own admonishing of her friend who was drunk and tried to steal a road sign.

(7) FLS, #18
ja govoryu
I say
«/\Brenton@
Brenton
Kakoj /\koshma-ar!,,,
What nightmare
nelzya takije /\ve-eshchi!,,,
must.not such.Pl things
eto zhe /\u-uzhas prosto!,,,

this emph.part terror simply
·· *nas posadyat v /\tyur'mu nepremenno!».*
us.Acc will.lock.up.Pl in prison for.sure
(I say, "Brenton! It's terrible! One must not do such things! It's simply dreadful! They will lock us up in prison for sure!")

Here we have typical persuasive intonation on one hand (rising-falling tone marked as '/\' and slight emphasis), but at the same time, we have series of prolongations, that adds the idea of 'open list' to the pattern. The result is marked with a combination of '!' and ',,,' in our transcription.

This is a very common scheme. It can be used with simple rising tone (/, marked as ','); with rising to high-level tone (/–, marked as ',,,'), and sometimes with persuasive rising-falling tone, like in the example we have just discussed. In some cases this scheme is used for the whole reported speech construction, and sometimes the list starts somewhere in the middle, like above, where the first EDU with the vocative phrase is pronounced normally.

The second typical pattern is the ***head-tail pattern***. This scheme uses more or less 'direct' intonation in the most part of the reported speech construction, and then on the last EDU switches to normal narrative intonation, either with 'comma intonation' (rising tone or not-deep falling tone that ends up in a medium pitch range) or with a 'period' one (falling into low pitch). As a result, the last EDU of the reported speech also serves as a means to make the whole construction a part of the discourse.

In the example (8) below, the speaker recreates her mother's exclamations in the first two EDUs of the reported speech, but the last one has weakened accents and typical narrative intonation, needed to incorporate the reported speech into the story line.

(8) SLS, #12
·· *Ona –govorila:*
She said
·· *«/\Vot!,*
Here
·· */\posmotrite¡!*
look.Imp
·· *Kakije –doma.»,*
what buildings
no my s Galechkoj ne smotreli ni na \Kreml',
but we with Galechka.Instr not looked neither at Kremlin
·· *ni na \kakije doma,*
nor at any buildings
(She said, "There! Look! What [beautiful] buildings!", but Galechka and me did not look either at Kremlin, or at any buildings, ...)

In the similar way, the speaker in the example (8) retells the policeman's exclamations and prompting with correct Russian intonation for such illocutionary acts,

but in the end of the last EDU she switches to the 'uncertain open list' intonation, instead of making it a proper question. It is more important to her to convey the idea that the policeman said 'many things of the same type' than to recreate his intonation precisely.

(9) FLS, #2
V itoge ona real'no /zazhuzhzhala,
In result.Dat it.she really buzzed
i militsioner skazal
and the.policeman said
«/\Devchonki@
Girls
U vas chto-to s /\dvigatelem!
of you.Gen something with the.engine.Instr
Ezzhajte v avtoservis s avarijkoj /\bystree¡
Drive.Imp to the.service with the.alarm.light quick
Ili mozhet byt' vam vyzvat' etogo kak ego tipa /"Angela"?»...
Or may be you.Dat to call this how it.Acc like 'Angel'
(In the end, the policeman said, "Girls! There is something [wrong] with your engine! Turn on the alarm lights and go to the service station quickly! Or maybe you should call for that – what's-its-name – 'Angel'?" [*Angel* is a name for vehicle recovery service])

This is also one of the most common patterns for the semi-direct speech. Basically, it is a compromise between needing to express several illocutionary meanings at the same time and to make the reported speech a part of the discourse as a whole (in the case of a narrative speech, a part of the storyline).

4. Concluding remarks

In this short study, I have examined the basic intonation patterns in Russian reported speech constructions, based on the data from the Prosodically Annotated Corpus of Spoken Russian. I have attempted to show that (i) intonation patterns in preceding framing clauses correspond with the type of the reported speech; (ii) however, this correspondence is more a tendency than a cause-and-effect relationship; (iii) there are some typical patterns in semi-direct speech that use 'mixed' intonation in order to keep the 'original' illocutionary meanings and to integrate the reported speech into following context as much as possible: the ***list pattern*** and the ***head-tail-pattern***.

This lays foundation for the future research that will include working with a fully annotated corpus, where in addition to main and secondary accents, and punctuation marks for illocutionary meaning we plan to use markings for specific intonation schemes. This will allow for a full analysis of intonation patterns in reported speech and other polypredicative constructions in Russian spoken discourse.

References

1. *Couper-Kuhlen E.* (1999), Coherent voicing: On prosody in conversational reported speech, in Wolfram Bublitz & Uta Lenk, eds., Coherence in Spoken and Written Discourse: How to create it and how to describe it, Benjamins Amsterdam, pp. 11–32.
2. *Holt E.* (1996), Reporting on talk: The use of direct reported speech in conversation, Research on language and social interaction, 29 (3), pp. 219–245.
3. *Kibrik A. A., Podlesskaja V. I., Korotaev N. A.,* (2009), Spoken discourse structure: main elements and iconic features [Struktura ustnogo diskursa: osnovnye èlementy i kanonicheskie javlenija], in Kibrik A. A., Podlesskaja V. I. (eds.) (2009) Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse], Languages of Slavonic Culture, Moscow, pp. 55–101.
4. *Klewitz G., Couper-Kuhlen E.* (1999). Quote — Unquote? The Role of Prosody in the Contextualization of Reported Speech Sequences. Pragmatics, Vol. 9, No. 4, 1999, pp. 459–485.
5. *Levontina I. B.,* Pereskazyvatel'nost' v russkom jazyke [Quotation and rendering markers in Russian]. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, 2011, pp. 284–288.
6. *Litvinenko A. O., Korotaev N. A., Kibrik A. A., Podlesskaja V. I.* (2009), Reported speech constuctions [Construktsii s tsitatsiej, ili "chuzhoj rech'ju"], in Kibrik A. A., Podlesskaja V. I. (eds.) (2009) Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse], Languages of Slavonic Culture, Moscow, pp. 288–308.
7. *Litvinenko A. O.,* Strategii peredachi 'chuzhoj rechi' v rasskazah po kartinkam (na materiale russkogo jazyka) [Speech reporting strategies in Russian comics-based stories]. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 425–433.
8. *Shvedova N. Ju.* (ed.) (1980) Russkaja grammatika [Russian grammar]. Moscow: Nauka.
9. *Tannen D.* (1989) "Oh talking that is so sweet": Constructing dialogue in conversation, in Talking voices, Cambridge University Press, Cambridge, pp. 98–133.

АВТОМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ СЛОЖНЫХ СЛОВ ПУТЕМ КОМБИНИРОВАНИЯ ЯЗЫКОЗАВИСИМЫХ И ЯЗЫКОНЕЗАВИСИМЫХ ПРИЗНАКОВ

Логинова-Клуэ Е. А. (elizaveta.loginova@univ-nantes.fr),
Daille B. (beatrice.daille@univ-nantes.fr)

Университет г. Нант, Нант, Франция

Ключевые слова: сегментация сложных слов, меры близости, правила трансформации компонентов, специализированный корпус

MULTILINGUAL COMPOUND SPLITTING COMBINING LANGUAGE DEPENDENT AND INDEPENDENT FEATURES

Loginova-Clouet E. A. (elizaveta.loginova@univ-nantes.fr),
Daille B. (beatrice.daille@univ-nantes.fr)

Nantes University, Nantes, France

Compounding is a common phenomenon for many languages, especially those with rich morphology. Dealing with compounds is a challenge for NLP systems since compounds are not often included in the dictionaries and other lexical sources. We present a compound splitting method combining language independent features (similarity measure, corpus data) and language specific component transformation rules. Due to the usage of language independent features, the method can be applied to different languages. We report on our experiments in splitting of German and Russian compound words, giving positive results compared to matching of compound parts in a lexicon. To the best of our knowledge. elaborated compound splitting is a rare component of NLP systems for Russian, yet our experiments show that it could be beneficial to use a specialized vocabulary.

Key-words: compound splitting, multilingual tool, similarity measure, component transformation rules, specialized corpora

1. Introduction

Compounding is a method of word formation consisting in a combination of two (or more) autonomous lexical elements that form a unit of meaning. This phenomenon is common in German, Dutch, Greek, Swedish, Danish, Finnish and many other languages. In Russian compounding is less regular, but also present, especially in specialized fields. Compound treatment is a problem for the automatic NLP systems because most of compounds are not listed in lexical sources, and not so frequent to be observed in training data. However, their recognition and splitting could be of benefit for various NLP tasks (machine translation, information retrieval, terminology extraction, etc.).

Compounding mechanisms are more or less complex depending on language. In highly analytical languages such as English or French, compound parts are just concatenated: EN[1] *parrotfish*, FR *kilowatt-heure*, kilowatt-hour. In languages with a rich morphology, some transformations are possible at the boundary of the compound parts. The word ending can be omitted, and/or boundary morphemes can be added, for example in DE:

(1) *Staat**s**feind* (state enemy) = *Staat* (state) + *Feind* (enemy)*;*

(2) *Muse**en**verwaltung (museum administration) = Museum (museum) + Verwaltung (administration);*

For some languages the list of such rules is rather short and exhaustive, for others it is more complicated. Sometimes a modification of the stem is possible, as in Russian:

(3) *ве**тро**генератор (wind generator) = ве**тер** (wind) + генератор (generator);*

A special case appears with the "neoclassical compounds", i.e. compounds which one element or more has Latin or Greek etymological origin [Namer, 2009]. For example EN *multimedia*, DE *turbomaschine* (turbomachine), etc. Usually these elements are not autonomous, but represent the units of meaning. Sometimes neoclassical elements are included in dictionaries or lexical databases.

In this paper, we examine some existing methods of compound splitting and then we propose our method combining language dependent and independent features. We report on our experiments in splitting German and Russian compounds. We conclude with some remarks on compound splitting in general and on its particularity in Russian language.

[1] EN — English language, FR — French language, DE — German language, RU — Russian language

2. Compound Splitting Methods

Compound splitting methods can be divided into supervised (generally rule-based) and unsupervised (fully statistical) methods. Let us illustrate the first type of methods on the example of German compound splitters. These are often based on the study of [Langer, 1998], describing the transformation rules for compound formation in this language. Systems check whether a word's component matching with a dictionary [Ott, 2005] or with a monolingual corpus [Koehn and Knight, 2003], [Weller and Heid, 2012]. Corpus-based approaches give also a probability for each segmentation, estimated from the components frequencies in the corpus. A parallel English corpus could be involved to check correspondences of decomposed parts [Koehn and Knight, 2003]. These methods are robust and give high results for the languages they were designed for.

The second group of approaches are language independent. [Macherey et al., 2011] propose to automatically extract morphological operations at the components boundary. The training of the model for a new language requires a parallel English corpus. It allows the authors to test their method for several languages: Danish, German, Norwegian, Swedish, Greek, Estonian, Finnish. [Hewlett and Cohen, 2011] detect automatically the places of components boundaries. The algorithm is based on the probability of the character's sequences in a language. The measure of probability is entropy: the entropy inside the word is relatively low, whereas the entropy on the word/component boundaries is much higher. Currently fully statistical models are not as precise as rule-based methods, but their advantage is their reusability for any language.

3. Multilingual Compound Splitting Algorithm

Our goal was to design a compound splitting tool that could be applied to different languages through language independent features, but also able to integrate linguistic rules if they are available for a given language. As a language independent feature, we chose monolingual corpus data and similarity measure between a word subsequence and the candidate lemmas.

To split a compound, we start with forming all its possible two-part segmentations beginning with the components of minimum permitted length (which is a parameter):

(4) *DE Magnetisierungszustand (magnetisation state)*
magnetisierungszustand -> ma + gnetisierungszustand
magnetisierungszustand -> mag + netisierungszustand
…
magnetisierungszustand -> magnetisierungszusta + nd

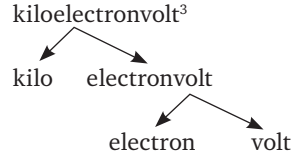
If the specific rules for component transformation into independent lexemes are available for this language, we apply them to the candidate components before matching with the dictionary/corpus. These are the rules of type: "s" → " " cf. (1), "en" → "um" cf. (2), etc.

For each candidate segmentation, both parts are matched with a monolingual dictionary, and optionally with a monolingual corpus. The corpus serves to calculate words frequency, which enables the tool to choose more plausible component candidates if several variants are possible. The corpus should be of reasonable size to obtain realistic distribution of words. The corpus is particularly useful if we deal with specialized vocabulary, containing many highly specialized terms not described in general language dictionaries.

If we do not have transformation rules or if they do not let finding a lemma, we exploit similarity measure. When searching in the dictionary/corpus, we calculate similarity between the segmentation part and candidate lemmas to choose the "closest" lemmas[2]. Various similarity measures could be used. So far we tried "normalized edit distance", based on Levenshtein distance, and "longest common prefix" measures (for detailed outlook of existing measures see [Frunza and Inkpen, 2009]).

If some acceptable lemmas are found for the left part of the current segmentation, but not for the right part, we try to split further the right part in a recursive manner, and so on up to a certain level. This level is a parameter corresponding to the maximum number of components.

RU килоэлектронвольт: kiloelectronvolt[3]



If we have acceptable candidate lemmas for all components, we calculate the score for this segmentation based on obtained similarity value, existence in the dictionary and (optionally) frequency in the corpus for each component. Finally, the tool returns a top N of the best segmentations ordered by their score. For example, for DE Magnetisierungszustand (magnetisation state) the output is:

magnetisierung + zustand	2.00
magnete + sie + erregungszustand	0.75
magnete + sicherungskasten[4]	0.69

The correct split is Magnetisierung (magnetisation) + Zustand (state), and it has the best score given by the program. The algorithm enables to set various parameters depending on our heuristics for a given language and on the application aimed. The user chooses either to split all given words (it supposes that all given words are compounds), either to first match each word with a dictionary and not to split the words found (the case of application to machine translation). Source code with detailed algorithm description, as well as test data, are available online[5].

---

[2] The threshold for lemma acceptance is a parameter

[3] Russian word is given in transliterated form

[4] The word "sicherungskasten" has similarity of 0,6 with the substring "sierungszustand", that is why it was found by the program

[5] http://www.lina.univ-nantes.fr/?Compound-Splitting-Tool.html

4. Experiments

In this section we report on our experiments using the algorithm described above. So far we applied it for compound splitting in two languages, German and Russian. We chose German because in this language compounding is very productive and well-described. Compounding in Russian is less frequent, so the question can be asked: does an NLP system for Russian really need a splitting mechanism, or is it sufficient just to add all known components in the system lexicon? Our experiments were guided by this question.

For both languages, we analyzed compound words from the domain of wind energy. We varied some parameters to observe the impact of corpus usage, of boundary transformation rules and similarity measure on the quality of splitting. As a baseline we performed splitting only with a dictionary, as if we were simply searching for the word components in the lexicon (that is applied in some NLP systems).

To evaluate the results, we calculated precision at rank 1 (top 1) and precision at rank 5 (top 5). Precision is calculated as the number of correct splits divided by the total number of compounds. We did not calculate recall in these experiments because we only analyzed compound words. A procedure deciding to split a token or not can constitute a topic for future researches.

4.1. Experiments with German compounds

For German language, three experiments were done: baseline splitting (only with a dictionary); splitting with dictionary and boundary transformation rules; and splitting with dictionary, rules and corpus filtering. The rules used in the second and third experiments are based on [Langer, 1998] work. Similarity is based on Levenshtein distance measure.

We used a German part of free German-English dictionary Dict. cc[6] (800,000 word entries); a specialized corpus related to wind energy domain crawled from the web[7] (300,000 words); and a test set of 446 compounds for splitting[8] consisting of two, three or four components. The results are presented in Tab. 1.

Table 1. Splitting Precision for German Language

	Baseline	Rules, no corpus	Rules, corpus
Top 1	66.59%	**93.04%**	87.44%
Top 5	66.59%	95.06%	**95.51%**

[6] http://www.dict.cc

[7] http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html

[8] [Weller and Heid, 2012], data available at: http://www.ims.uni-stuttgart.de/~wellermn/tools.html

The results with addition of transformation rules and similarity measure are clearly better than those of baseline-experiment. The results are comparable with those given by methods designed for German language: thus, [Koehn and Knight, 2003] report on accuracy of 95,7%[9] for their monolingual frequency based method.

The usage of corpus slightly increases precision for top 5. It allows a correct splitting of some additional words, whose components are not present in the dictionary (Netzanschluß, "network connection"). In some cases, it also improves the ranking: Traktionsbatterie without corpus returns two equal-ranked splits traktion + batterie 1.0 and trakt + ion + batterie 1.0. The usage of corpus raises the correct split: traktion + batterie 1.50, trakt + ion + batterie 1.25. Though, in other cases corpus affects the ranking because it promotes the splits consisting in shorter and more frequent components: Aberrationswinkel, "aberration angle", without corpus is correctly split in aberration + winkel, and with corpus the best-ranked split is aber + ration + winkel. That is why the precision in top 1 with corpus is lower than without. This problem may be resolved in replacing simple corpus frequency by specificity, i.e. comparing special corpus frequency to the frequency in a general corpus (cf. "weirdness ratio" [Ahmad et al., 1992]).

4.2. Experiments with Russian compounds

For Russian language, in addition to baseline experiment, we varied three parameters: the usage or not of the corpus; similarity measure (Levenshtein distance vs. The longest common prefix, later "Prefix"); and small rules-set or large rules-set. So we did 9 experiments with different combinations of these parameters.

The transformation rules for Russian were formulated on the base of description of Russian morphology in [Zaliznjak, 1977]. The small rules-set consists in two simple rules expressing the common knowledge that linking morphemes "о" and "е" operate as boundary morphemes in Russian. For the full rules-set see Table 2.

Table 2. Transformation Rules for Russian compounds

N	Left context	Transformation	Example
Small rules-set			
1	-	"о" → " "	
2	-	"е" → " "	
Large rules-set			
3	-	"о" → "а"	водо- / вода
4	-	"е" → "я"	земле- / земля
5	"ж"|"ш"|"щ"|"ч"|"ц"	"е" → "а"	тысяче- / тысяча
6	-	"е" → "ь"	жизне- / жизнь
7	-	"о" → "ый"	крупно- / крупный

[9] Accuracy is calculated here in the same way we calculate precision.

N	Left context	Transformation	Example
8	-	“о” → “ой”	криво- / кривой
9	-	“е” → “ий”	обще- / общий
10	“к”|“г”	“о” → “ий”	высоко- / высокий
Inflexion rules			
11	-	“ый” → “ ”	
12	-	“ий” → “ ”	
13	-	“ой” → “ ”	

We used electronic version of [Ozhegov, 1991] dictionary (nearly 61,000 words), completed by a list of neoclassical elements taken from [Béchade, 1992] and translated into Russian. Neoclassical elements are very frequent in Russian compounds and necessary for correct splitting, but only few of them were already included in the dictionary. We used a Russian monolingual wind energy corpus of 300,000 words crawled from the web[10].

Test data are issued from the wind energy corpus. Among 7,000 most frequent lexemes in this corpus, 348 are compounds. It confirms that compounding in Russian, even if it is not as productive as in some Germanic languages, needs to be taken into account, at least for specialized domains. The results for all compounds are presented in Table 3.

Table 3. Splitting Precision for Russian Language

	Base-line	Levenshtein, no corpus		Levenshtein, corpus		Prefix, no corpus		Prefix, corpus	
		Small rules	Large rules	Small rules	Large rules	Small rules	Large rules	Small rules	Large rules
Top 1	35.06%	62.64%	75.57%	76.44%	84.77%	58.05%	68.97%	72.99%	78.74%
Top 5	35.06%	71.84%	81.32%	86.78%	92.82%	69.83%	80.17%	90.52%	92.24%

We noted a significant difference between baseline and other results. The usage of corpus was definitely beneficial in all analyzed cases. Some component lemmas were not present in the dictionary (*дизель*, diesel, *интернет*, internet, etc.). Compounds containing these components were correctly split through the corpus. In small rules set experiments, the Prefix similarity measure was a bit better for top 5. In some cases this measure compensates for the absence of inflexion treatment because it compares the common beginning of strings. However, the addition of large rules allowed inflexion treatment, and the measure based on Levenshtein distance became more efficient for top 1 and 5.

The impact of large rules set was not spectacular for top 5 with the usage of corpus, since for certain compounds the corpus compensates for the lack of rules. For example, the adjective *электромагнитный* (electromagnetic) could not be correctly split just with a baseline-method because its right component *магнитный* (magnetic)

[10] http://www.lina.univ-nantes.fr/?Linguistic-Resources-from-the.html

is not in the dictionary. It could be correctly split either with the usage of corpus (where *magnetic* is present and has a relatively high frequency), either with the transformation rules which enable to find the noun *магнит,* magnet:

(5) *электромагнитный → электро + магнитный*
rule 11 (магнитный) = магнитн
similarity (магнитн, магнит) = 0,86
result : электромагнитный = электро + магнит

By contrast, we noted a good improvement through the large rules for top 1 with the corpus (6–8 % increasing of precision), and also for all experiments without corpus (10–13 % increasing of precision).

5. Conclusion

We have presented a compound splitting algorithm combining language independent features (similarity measures, word frequencies in a corpus) with language dependent features (component boundary transformation rules). For the two analyzed languages, this mechanism outperforms a baseline method, consisting in a matching of the word components in a dictionary. The usage of a specialized corpus allows us to correctly split some additional compounds including components unknown in a dictionary, and enables to a certain extent to compensate the lack of transformation rules. Using more rules enables although to achieve better ranking of splits. The algorithm can be applied to other languages by changing the lexical sources and, optionally, editing transformation rules.

Concerning compound splitting in Russian, it seems to deserve a special treatment in the NLP systems, at least for the systems dealing with specialized texts. Another solution, currently used in some systems, is to keep all compound components in the lexicon, which largely increases lexicon size. This solution does not seem satisfactory for multilingual systems since it requires to complete the lexicon for each new language.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 248005.

References

1. *Ahmad K., Davies A., Fulford H. and Rogers M.* (1992), "What is a term? The semi-automatic extraction of terms from text". Translation Studies: An Interdiscipline, John Benjamins, Amsterdam/Philadelphia, pp. 267–278.
2. *Béchade H.-D.* (1992), Phonétique et morphologie du français moderne et contemporain, Presses Universitaires de France, Paris.
3. *Frunza O., Inkpen D.* (2009), "Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques". International Journal of Linguistics, Vol. 1, No. 1, available at: http://www.macrothink.org/journal/index.php/ijl/article/view/309/193
4. *Hewlett D., Cohen P.* Fully Unsupervised Word Segmentation with BVE and MDL. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 540–545, Portland, Oregon, 2011.
5. *Koehn, P., Knight, K.* Empirical methods for compound splitting. Proceedings of EAC-2003, Budapest, Hungary.
6. *Langer, S.* Zur Morphologie und Semantik von Nominalkomposita. Proceedings of 4th Conference Computers, linguistics and phonetics between language and speech (KONVENS) , Bonn, 1998, pp. 83–97.
7. *Macherey K., Dai A. M., Talbot D., Popat A. C., Och F.* Language-independent Compound Splitting with Morphological Operations. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL). Portland, Oregon, 2011, pp. 1395–1404.
8. *Namer Fiammetta* (2009), Morphologie, lexique et traitement automatique des langues, Lavoisier, Paris.
9. *Ozhegov S. I.* (1991), Tolkovyj slovar' russkogo jazyka [Russian Language Dictionary], web version available at: http://speakrus.ru/dict/ozhegovw.zip.
10. *Ott, N.* (2005), "Measuring Semantic Relatedness of German Compounds using GermaNet", available at: http://niels.drni.de/n3files/bananasplit/Compound-GermaNet-Slides.pdf
11. *Weller M., Heid U.* Analyzing and Aligning German Compound Nouns. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, 2012.
12. *Zaliznjak, A. A.* (1977), Grammaticheskij Slovar' Russkogo Jazyka [Grammatical Dictionary of the Russian Language], Russkij jazyk, Moscow.

ВИЗУАЛИЗАЦИЯ ДАННЫХ ДЛЯ КАТАЛОГА РУССКИХ ЛЕКСИЧЕСКИХ КОНСТРУКЦИЙ (НА МАТЕРИАЛЕ НКРЯ)

Ляшевская О. Н. (olesar@gmail.com)

НИУ Высшая школа экономики, Москва, Россия; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Митрофанова О. А. (alkonost-om@yandex.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Паничева П. В. (ppolin86@gmail.com)

EPAM Systems, Россия

Доклад отражает новые результаты, полученные в ходе совместного проекта кафедры математической лингвистики СПбГУ с разработчиками Национального корпуса русского языка (НКРЯ, http://ruscorpora.ru). Цель исследования — разработка технологии автоматического распознавания в тексте конструкций, связанных с той или иной лексической единицей, и применение этой технологии в создании каталога русских лексических конструкций. Выделение конструкций предполагает использование потенциала многоплановой лингвистической разметки НКРЯ (прежде всего, лексико-семантической). В докладе обсуждается использование модуля визуализации данных для уточнения информации о конструкциях, реализующих значения исследуемых слов. Модуль предназначен для лексикографов и исследователей лексики и грамматики русского языка.

Ключевые слова: именные конструкции, сочетаемость, НКРЯ, лексико-семантическая разметка, лексико-грамматическая разметка, визуализация данных

DATA VISUALIZATION FOR BUILDING THE CATALOGUE OF RUSSIAN LEXICAL CONSTRUCTIONS (BASED ON RNC)

Lyashevskaya O. N. (olesar@gmail.com)
NRU Higher School of Economics, Moscow, Russia;
Vinogradov Institute of Russian Language RAS, Moscow, Russia

Mitrofanova O. A. (alkonost-om@yandex.ru)
Saint-Petersburg State University, St. Petersburg Russia

Panicheva P. V. (ppolin86@gmail.com)
EPAM Systems, Russia

Our research aims at automatic identification of constructions associated with particular lexical items and its subsequent use in building the catalogue of Russian lexical constructions. The study is based on the data extracted from the Russian National Corpus (RNC, http://ruscorpora.ru). The main accent is made on extensive use of morphological and lexico-semantic data drawn from the multi-level corpus annotation. Lexical constructions are regarded as the most frequent combinations of a target word and corpus tags which regularly occur within a certain left and/or right context and mark a given meaning of a target word. We focus on nominal constructions with target lexemes that refer to speech acts, emotions, and instruments. The toolkit that processes corpus samples and learns up the constructions is described. We provide analysis for the structure and content of extracted constructions (e. g. r:ord der:num t:ord r:qual|*pervyj 'first'* + LJUBOV' 'love'; LJUBOV' 'love' + PR|*s 'from'* + ANUM m sg gen|*pervyj 'first'* + S f inan sg gen|*vzgljad 'sight' = love at first sight*). As regards their structure, constructions may be considered as n-grams (n is 2 to 5). The representation of constructions is bipartite as they may combine either morphological and lemma tags or lexical-semantic and lemma tags. We discuss the use of visualization module PATTERN.GRAPH that represents the inner structure of extracted constructions.

Key words: nominal constructions, word co-occurrence, patterns, Russian National Corpus, lexico-semantic annotation, lexico-grammatic annotation, data vizualization

1. Введение

Данная статья продолжает цикл публикаций, посвященных автоматическому выделению лексических конструкций в контекстах НКРЯ (см., например, [Lashevskaya et al. 2012, Mitrofanova et al. 2012]). Цель проекта — предложить основанную на статистических методах технологию автоматического распознавания типичных конструкций, связанных с той или иной лексической единицей.

Современные корпуса и веб-архивы дают возможность собрать статистику о поведении лексической единицы в контексте, составить портрет контекстного окружения слова (Behavioural Profile, [Divjak, Gries 2009]). Дистрибутивная гипотеза [Firth 1957/1968, Sahlgren 2008] предполагает, что различающиеся группы контекстов отражают употребления слова в разных значениях. Например, резкая мена контекстного окружения в корпусе нового времени может обозначать, что у слова появилось новое значение.

Традиционно контекстный профиль представляют как наборы n-грамов (обычно 1…4 словоформ, примыкающих к ключевому слову справа и слева), которые сгруппированы по определенным признакам. На *n*-грамах строятся векторные модели, с которыми удобно работать машине, однако человеческому глазу как векторные таблицы, так и сами списки *n*-грамов не слишком удобны. Секрет в том, что человеку свойственно видеть вместо множеств — структуру: вместо наборов *n*-грамов в словарях и грамматиках содержатся указания на типичные признаки контекста: например, что лексема употребляется в переходной конструкции, управляет творительным падежом, присоединяет тот или иной предлог; приводится наиболее показательная лексическая сочетаемость. Тем самым, какие-то повторяющиеся в *n*-грамах признаки признаются важными, а другие элементы, случайные, отметаются.

Задача автоматического распознавания структуры контекстов призвана перебросить мост между корпусной выборкой, на которой строятся *n*-грамы, и словарем/грамматикой. Адресат модуля визуализации корпусных данных, о котором пойдет речь в статье — лингвист-лексикограф или исследователь лексики, морфологии и синтаксиса. Идея состоит в том, чтобы электронный помощник кластеризовал корпусные примеры на употребление той или иной лексемы и выделял повторяющиеся в контекстном окружении паттерны.

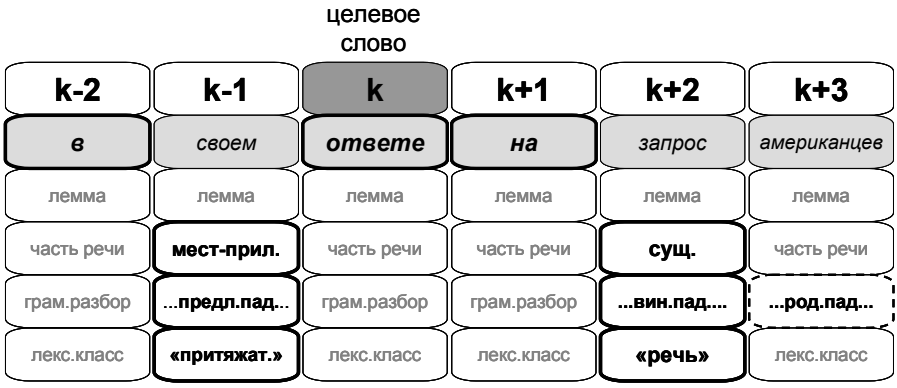


Рис. 1. *N*-грам *в своем ответе на запрос американцев* с разметкой разных уровней

Для выделения повторяющихся признаков можно воспользоваться корпусным арсеналом, чтобы за каждым элементом n-грама стояла разметка разных уровней (часть речи, синтаксическая группа и т. п., для русского языка — лемма

и словоизменительные грамматические характеристики). Тогда *n*-грам *в своем ответе на запрос американцев* (см. Рис. 1) попадет в кластер с повторяющимися элементами, среди которых будут предлоги *в* и *на,* притяжательное прилагательное в предл. падеже, имя из класса «речь» в вин. падеже и слово в род. падеже (элементы расположены в определенном порядке). Более крупный кластер будет включать цепочку *в* + *ответе* + *на* + «речь». Использование разметки разного уровня даст более мощный инструмент, нежели обычные кластеры *n*-грамов словоформ (как в корпусах М. Дейвиса, http://corpus.byu.edu). Он также будет более гибким, нежели Sketch Engine (http://www.sketchengine.co.uk), т. к. набор грамматических паттернов в нем не будет задан заранее.

Наша гипотеза состоит в том, что выделяемые последовательности должны интерпретироваться как лингвистически значимые законченные лексические конструкции (центр конструкции — искомое целевое слово). Вместе с тем, следует иметь в виду, что пользователи могут быть заинтересованы в получении разных конструкций — разной длины, разной степени абстрактности и т. п. Инструмент должен показывать, как изменится паттерн при переходе от более дробных к более крупным кластерам, можно ли изменить расстояние между элементами и что ожидается во «вставке» и т. п. Нужно также предусмотреть взаимодействие конструкции с элементами, традиционно в нее не включаемыми — например, как изменится паттерн, если за целевым словом *ответ* будет следовать частица *же*.

Тема динамической визуализации данных, к сожалению, пока еще редко поднимается в корпусной лингвистике — особенно если речь идет о пользователях, не искушенных в квантитативной лингвистике и в работе со статистическими программами. В этой статье описаны пилотные эксперименты по визуальному представлению именных конструкции, и пока еще далеко не все задачи решены. Однако, мы надеемся, что проблематика статьи вдохновит разработчиков на создание разнообразных визуализаторов корпусных данных в помощь лингвистам.

2. Теоретическая база исследования

В центре исследования находятся именные конструкции — прежде всего, те, которые строятся вокруг имен существительных. Исследовались как конструкции отдельных лексем, так и конструкции, свойственные целым лексико-семантическим группам: обозначениям речевых действий (*дискуссия, комплимент, обращение, обсуждение, ответ* и т. д.), названиям эмоций (*апатия, благодарность, грусть, гнев, любовь* и т. д.), именам инструментов (*бритва, веник, весло, карандаш, коса* и т. д.).

Лексические конструкции — это наблюдаемые в речи последовательности лексических единиц, из которых одно (или несколько) — лексическая константа, а другие — переменные [Fillmore 1988a]. Предполагается, что слово в определенном значении способно структурно организовывать контекст вокруг себя — то есть характеризуется набором лексических конструкций,

которые строятся вокруг нее. Тем самым, основная функция конструкции — фиксировать регулярную сочетаемость целевого слова в определенном его лексическом значении (наполнение слотов ассоциируется с семантикой целевого слова).

Согласно идеологии Грамматики конструкций [Fillmore 1988b, Goldberg 1995, 2006, Tomasello 2003], лексическая конструкция, как и другие виды конструкций, обладает единством формы и значения. Форма лексической конструкции задается, с одной стороны, очевидно, лексически фиксированными единицами, а с другой стороны — ограничениями на заполнение переменных слотов: морфологическими, синтаксическими, лексико-семантическими. Форма конструкции может предусматривать и грамматические ограничения на форму ключевого слова — лексического центра конструкции. Конструкция может реализоваться в виде синтагмы: простого или сложного словосочетания, которое может, например, реализовать рамку валентностей целевого слова или даже выходить за ее пределы.

С точки зрения структурной организации, конструкция — это комбинация целевого слова и слотов, заполняемых регулярными контекстными соседями, среди которых могут быть леммы, грамматические (морфологические и синтаксические), лексико-семантические и т. п. признаки. Точнее говоря, по своей природе, конструкция — это абстрактный шаблон, предполагающий лексикализацию, т. е. различные реализации в виде комбинаций лемм/словоформ, ср. V|*дать, найти, предложить…* ОТВЕТ + PR|*на* + speech r:abstr|*вопрос*, r:qual|*простой, неоднозначный…* + ОТВЕТ, ОТВЕТ + t:hum r:concr|*академикам, мудрецам, отцу…*

Значение конструкции характеризуется большей или меньшей устойчивостью, варьирующей от регулярной свободной сочетаемости до высокой идиоматичности. Значение конструкции, как правило, некомпозиционно, то есть не выводится из значения составляющих элементов (в особенности если рассматривается абстрактный шаблон — комбинация целевого слова и лексико-семантических тегов классов). Лексикализованные конструкции могут удовлетворять принципу композиционности, если в них реализуется типовая свободная сочетаемость. Некомпозиционные сочетания (фраземы), в которых лексически фиксированы все элементы (ср. *любовь с первого взгляда*), также входят в фонд лексических конструкций, наряду с более свободными шаблонами, где ограничения на элементы задаются признаками типа «глагол», «инфинитив», «предлог *на* + предложный падеж».

Конструкция, понимаемая таким образом, это многоярусная структура, призванная компактно и в достаточной мере полно описать сочетаемостные возможности целевого слова, ассоциированные с его лексическим значением, и задать сочетаемость не только в терминах лемм/словоформ, но и с точки зрения грамматических и лексико-семантических классов. Данный взгляд на конструкции отражает идею взаимосвязи и взаимопроникновения различных уровней языка (от фонетического/графического до лексического) и позволяет рассмотреть языковые выражения не в их проекции на один из множества уровней (как представлялось бы с точки зрения модульного подхода), а как многоярусные структуры.

Подводя итог сказанному выше, можно заметить, что наше определение конструкции не противоречит традиционному, однако несколько выходит за его рамки. Принятое в нашем исследовании понимание конструкции позволяет, в отличие от метода *n*-грам, относиться избирательно к сочетаемостным возможностям целевых слов, учитывать тенденции в сочетаемости целевого слова и его соседей в контексте, описывать как лексическую сочетаемость, так и сочетаемость на уровне классов, не только устойчивые, но и свободные сочетания, важным образом отражающие типовое употребление слова в тексте.

Извлечение конструкций базируется на автоматической обработке множества корпусных контекстов, в которых употребляется целевое слово. Контекстные цепочки разбиваются на группы, идентифицируются типовые паттерны и, соответственно, выделяются признаки, обобщающие свойства элементов-соседей. Тем самым, компьютерная система «выучивает» конструкции по принципу генерализации свойств контекстного окружения (ср. гипотезу о генерализации конструкций при усвоении языка детьми [Tomasello 2003]). Нельзя не заметить, что на подобных же основаниях (у элементов с общим значением будет сходное контекстное окружение) действуют системы разрешения лексической неоднозначности (WSD), разметки семантических ролей (Semantic Role Labelling) и многие другие модули автоматической обработки текста. Однако в данном случае речь идет именно об экспликации информации в виде ограничений на элементы контекстных последовательностей.

Далее в статье мы опишем подходы к автоматическому выделению лексических конструкций, опишем эксперименты с выделением конструкций для имен эмоций, речи и инструментов и представим модуль визуализации структуры и наполнения конструкций.

3. Методика автоматического выделения конструкций в выборках НКРЯ

Итак, формально под конструкцией в нашем проекте понимается регулярная комбинация целевого слова (лексической константы) и различных тегов контекстного окружения, присутствующих в многоярусной разметке корпуса (см., в частности, [Lashevskaya et al. 2012, Mitrofanova et al. 2012]). В качестве основного лингвистического ресурса задействован Национальный корпус русского языка (НКРЯ), отличающийся богатством текстового наполнения, а также детальностью и многоплановостью лингвистической разметки. Акцент делается на использование в обучении корпусной разметки — на уровне лемм, частей речи, грамматических признаков, лексико-семантических признаков.

При выделении конструкций учитываются теги лемм, лексико-семантические и морфологические теги (*lex, sem, gr*). В этой связи, как конструкции нами рассматриваются, например, следующие сочетания целевых слов и элементов их контекстного окружения:

(1) ОТВЕТ + PR|*на* + t:speech r:abstr|*приветствие, вопрос, высказывание, рапорт, реплика*

(2) V pf tran inf act|*найти, дать* + A m sg acc inan plen|*простой, однозначный* + ответ + PR|*на* +S m inan sg acc|*вопрос*

(3) r:ord der:num t:ord r:qual|*первый* + ЛЮБОВЬ

(4) ЛЮБОВЬ + PR|*с* + ANUM m sg gen|*первый* + S f inan sg gen|*взгляд*

Семейства конструкций для отдельного слова ассоциируются с его значением, например:

(5) ДИСКУССИЯ + PR|*о, по, на*
r:qual|*горячий, долгий, жесткий, серьезнейший, старый, широкий* + дискуссия
ДИСКУССИЯ + PR|*о* + r:abstr|*вред, ценность, целесообразность, красота*
ДИСКУССИЯ + PR|*на* +t:pers|*тема*
V|*начать,организовать* + ДИСКУССИЯ
V + r:qual + ДИСКУССИЯ + PR + t:pers + r:abstr

В случае многозначности целевых слов каждое из их значений можно охарактеризовать определенным набором конструкций, следовательно, исследование семейств конструкций позволяет разграничивать (в том числе и автоматически) значения многозначного слова.

4. Эксперименты по автоматическому выделению конструкций в выборках НКРЯ

4.1. Инструмент автоматического выделения конструкций

Существует ряд проектов, в которых особое внимание уделяется формализации лексико-синтаксических связей единиц текста, например, PropBank (http://verbs.colorado.edu/~mpalmer/projects/ace.html), NomBank (http://nlp.cs.nyu.edu/meyers/NomBank.html), FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/), DeepDict (http://gramtrans.com/deepdict/), Sketch Engine (http://www.sketchengine.co.uk/), StringNet (http://nav3.stringnet.org/) и т. д. Данные ресурсы дают разноплановую информацию о сочетаемости лексических единиц, при этом форма представления результирующих данных, как правило, табличная. Исключение составляют PropBank и NomBank, где важнее всего оказывается семантико-синтаксическая разметка контекстов.

Для представления данных о конструкциях в рамках нашего проекта был создан специализированный модуль на языке Perl (разработчик С. В. Романов), где используются некоторые стандартные средства для обработки контекстных выборок с многоярусной лингвистической разметкой и для эффективной

выдачи данных (в частности, XML::LibXML, YAML, Log::Log4perl). Важнейший компонент нашего модуля — пакет Algorithm::Combinatorics, с помощью которого производится выявление частотных комбинаций тегов в контекстах для целевых слов.

На вход программы подается файл с выборкой контекстов с целевым словом, для которого требуется выявить конструкции. Затем пользователь определяет такие параметры обработки данных, как типы тегов, учитываемых при выделении конструкций (*lex, sem, gr*), ширина контекстного окна, в пределах которого ведется поиск частотных комбинаций тегов (от −5 до +5), а также число конструкций, попадающих в выдачу (от 1 до 50).

4.2. Анализ результатов работы инструмента автоматического выделения конструкций

Файл с результатами работы инструмента автоматического выделения конструкций содержит наиболее частотные сочетания целевого слова и различных тегов контекстного окружения (*lex, sem, gr*). В зависимости от назначенной ширины контекстного окна в выдачу попадают комбинации тегов в виде пар, троек, четверок, пятерок и т. д. Например, из троек в выдаче присутствуют частотные конструкции, организованные по схемам *sem+sem+sem, sem+sem+gr, sem+sem+lex, lex+sem+sem, gr+sem+sem, lex+sem+lex, lex+sem+gr, gr+sem+gr, gr+sem+lex* и т. д. Например, в случае конструкции *в азарте игры* мы получаем примерно следующие комбинации тегов в выдаче:

(6) *gr+lex+sem* PR|*в* + АЗАРТ + der:v r:abstr der:s|*игра*
gr+sem+gr PR|*в* + t:psych r:abstr + S f inan pl gen|*игра*
gr+sem+sem PR|*в* + t:psych r:abstr + der:v r:abstr der:s|*игра*
lex+sem+gr *в* + t:psych r:abstr + S f inan pl gen|*игра*
gr+gr+sem PR|*в* + S m inan sg loc + der:v r:abstr der:s|*игра*
gr+gr+gr PR|*в* + S m inan sg loc + S f inan pl gen|*игра*

и т. д.

В настоящий момент мы можем получать конструкции с двухслойной структурой, т. е. компоненты конструкции могут одновременно характеризоваться не более чем двумя признаками: морфологическими тегами и тегами лемм, или лексико-семантическими тегами и тегами лемм. Например,

(7) S f inan pl acc|*слеза* + УМИЛЕНИЕ
t:stuff r:concr t:liq|*слеза* + УМИЛЕНИЕ

(8) t:word r:concr r:abstr|*слово* + БЛАГОДАРНОСТЬ
S n inan pl ins|*слово* + БЛАГОДАРНОСТЬ
S n inan pl acc|*слово* + БЛАГОДАРНОСТЬ

(9) A,norm=acc,sg,f,plen|*опасный, тупой, механический, средневековый* + БРИТВА
A,norm=(gen,sg,f,plen|dat,sg,f,plen|ins,sg,f,plen|loc,sg,f,plen)|*опасный, безопасный* + БРИТВА
r:rel ev|*опасный* + БРИТВА
r:rel ev d:neg der:a|*безопасный* + БРИТВА
r:rel der:s|*механический* + БРИТВА

Одна из особенностей формата выдачи данных о конструкциях связана с тем, что у служебных слов (FW) отсутствует семантическая разметка, поэтому среди конструкций регулярно встречаются структуры вида FW+*lex*+FW, хуже интерпретируемые в комбинациях с лексико-семантическими тегами, но более очевидные в комбинациях с морфологическими тегами. Например, конструкции FW + ПОХВАЛА + FW может соответствовать структура типа PR|*к, за, после, в, с* + ПОХВАЛА + PR|*против, сверх, сквозь, в, на*.

Заметим, что большой интерес вызывают конструкции с компонентами, в состав которых входят лексико-семантические теги, поскольку чаще всего с ними ассоциируются группы лемм, выражающих общее значение и характеризующихся близкими дистрибутивными свойствами. Например:

(10) r:rel|*риторический, мировой, процедурный, спорный, шекспировский, практический, методический* + ВОПРОС
ОБСУЖДЕНИЕ + t:ment r:abstr|*проект, концепция* + r:abstr|*благоустройство, реформирование, реформа*
ОТВЕТ + FW + t:speech r:abstr|*запрос, призыв, вопрос, приветствие, просьба, высказывание, похвала, рапорт, реплика*

Наши данные позволяют проследить развертку простейшей структуры в сложную многокомпонентную конструкцию и исследовать видоизменение состава конструкции по пути движения от простого к сложному. Например,

(11) t:poss|*дать, получить, давать* + ОТВЕТ
r:qual|*простой, неточный, точный, вероятный, логичный, нужный, вразумительный, ясный, приличный* + ОТВЕТ
r:rel|*готовый, однозначный, стандартный, истинный, числовой, заданный, релевантный, эмоциональный, содержательный, необязывающий, отрицательный, утвердительный, хлесткий, окончательный, известный, конкретный, официальный, адекватный, отечественный, обстоятельный, определенный, реактивный, обоснованный, очевидный, зачаточный, энергичный, соответствующий, стойкий* + ОТВЕТ
t:move t:poss|*найти* + r:qual|*простой, точный, приличный* + ОТВЕТ + FW + t:speech r:abstr|*вопрос*
t:poss|*давать, дать* + r:rel|*конкретный, однозначный, окончательный* + ОТВЕТ + FW + *вопрос*

ОТВЕТ + PR|*на* + t:speech r:abstr|*приветствие, вопрос, высказывание, рапорт, реплика*
V pf tran inf act|*найти, дать* + A m sg acc inan plen|*простой, однозначный* + ОТВЕТ + PR|*на* ++S m inan sg acc|*вопрос*
найти, дать + простой, однозначный + ОТВЕТ + *на + вопрос*

Программа выделения конструкций позволяет получить основные статистические данные об их встречаемости в корпусе (абсолютные и относительные частоты), например:

(12) лемма: РЕКОМЕНДАЦИЯ
объем выборки: 193 контекста

sem+lex+sem
22 (19,13%) FW + РЕКОМЕНДАЦИЯ + FW
13 (11,30%) r:rel|*методический, негласный, технологический, подробный, конкретный, методологический* + РЕКОМЕНДАЦИЯ + FW
8 (6,95%) r:poss|*свой, их, его, наш, мой, ее* + РЕКОМЕНДАЦИЯ + FW
4 (3,47%) FW + РЕКОМЕНДАЦИЯ + t:hum r:concr t:prof|*специалист, политолог, стоматолог, журналист*
4 (3,47%) FW + РЕКОМЕНДАЦИЯ + t:hum r:concr|*старик, лекарь, спортсмен, член*
3 (2,60%) t:poss|*получить, давать, дать* + РЕКОМЕНДАЦИЯ + FW
2 (1,73%) r:dem|*этот* + РЕКОМЕНДАЦИЯ + r:rel|*особый, национальный*
2 (1,73%) t:speech r:abstr|*просьба, совет* + РЕКОМЕНДАЦИЯ + FW
2 (1,73%) r:abstr|*поступление, написание* + РЕКОМЕНДАЦИЯ + FW
2 (1,73%) r:abstr t:be:appear|*разработка* + РЕКОМЕНДАЦИЯ + FW

gr+lex+gr
6 (5,21%) CONJ|*и, однако* + РЕКОМЕНДАЦИЯ + PR|*по, в*
4 (3.47%) A pl gen plen|*общий, подробный, конкретный, методический* + РЕКОМЕНДАЦИЯ + PR|*на, по*
4 (3,47%) PR|*по* + РЕКОМЕНДАЦИЯ + APRO m sg gen|*этот, ваш, свой, один*
3 (2,60%) A pl nom plen|*методический, негласный* + РЕКОМЕНДАЦИЯ + PR|*по, на*
3 (2,60%) PR|*к, согласно, по* + РЕКОМЕНДАЦИЯ + S m anim sg gen|*политолог, производитель, журналист*
3 (2,60%) A pl ins plen|*методический, полезный* + РЕКОМЕНДАЦИЯ + PR|*по, относительно*
3 (2,60%) APRO pl acc inan|*свой, весь* + РЕКОМЕНДАЦИЯ + PR|*по, о*
3 (2,60%) PR|*на, за* + РЕКОМЕНДАЦИЯ + PR|*по, к*
2 (1,73%) V pf tran inf act|*подготовить, разработать* + РЕКОМЕНДАЦИЯ +PR|*о, по*
2 (1,73%) A pl acc inan plen|*соответствующий, лестный* + РЕКОМЕНДАЦИЯ + S m anim sg dat|*оператор, кот*

5. Визуализация структуры и наполнения конструкций

Наша нынешняя задача — из многообразия используемых в компьютерной лингвистике техник визуализации (ср., например, [Penn et al. 2009]) выбрать метод графического представления данных, отличающийся простотой и широкими иллюстративными возможностями, позволяющий отразить как состав конструкций, так и иерархию их компонентов.

Для получения графических представлений, отражающих структуру и наполнение конструкций, был задействован модуль pattern.graph (http://www.clips.ua.ac.be/pages/pattern-graph, [De Smedt 2012]), разработанный на языке Python и предназначенный для визуализации различных типов связей в тексте на естественном языке. На входе он принимает строку, обозначающую конструкцию, в формате, описанном в Разделе 2. На выходе создается граф, иллюстрирующий соответствующую конструкцию (см. рис. 2–3).

Визуализация производится в два этапа: во-первых, производится парсинг строки конструкции и выявление ее главных и второстепенных элементов с сохранением порядка следования, причем главными являются элементы, которые необходимым образом присутствуют и в своей линейной последовательности формируют конструкцию, в то время как второстепенные представляют из себя парадигматические варианты наполнения главных; формат входной строки позволяет однозначно провести данное разграничение; во-вторых, из них создается граф, отражающий данные структурные соответствия между элементами.

5.1. Парсинг конструкции и выявление ее элементов

В качестве исходных данных служат строки такого вида, как в примерах (1) и (4). В строках выявляются главные и второстепенные элементы. В главные элементы входит, во-первых, целевое слово, не имеющее тегов разметки; а также грамматические или лексико-семантические теги остальных слов в конструкции. Соответственно, первостепенными элементами в примерах (1) и (4) будут следующие (с учетом порядка следования):

(13) ОТВЕТ; PR; t:speech r:abstr

(14) ЛЮБОВЬ; PR; ANUM m sg gen; S f inan sg gen

Второстепенными элементами считаются леммы, обозначающие наполнение первостепенных грамматических и лексико-семантических элементов. Их количество может варьироваться от одного до нескольких десятков. Ср. соответствующие второстепенные элементы в примерах (1) и (4):

(15) *на*; *приветствие, вопрос, высказывание, рапорт, реплика*

(16) *с*; *первый; взгляд*

Для каждого второстепенного элемента сохраняется его соответствие «родному» первостепенному тегу.

Полученная структура из первостепенных, второстепенных элементов, порядка следования для первых и связей между первыми и вторыми передается для визуализации.

5.2. Визуализация элементов конструкции

В качестве узлов графа изображаются первостепенные и второстепенные элементы. При этом закрашенными узлами обозначаются главное слово конструкции, которое дополнительно выделяется красным цветом; а также лексемы, отражающие лексическое наполнение конструкции. Узлы лексико-семантических и морфологических тегов остаются пустыми.

Первостепенные элементы связываются направленными ребрами графа, в соответствии с порядком следования этих элементов в конструкции. Второстепенные элементы связываются с тегами, которые они наполняют, с помощью двунаправленных ребер. Для наглядности стрелки первого типа подсвечиваются зеленым, второго — синим цветом.

Задействованный нами модуль pattern.graph обладает особым удобством в использовании, так как не требует вычисления координат для изображения узлов и ребер. При указании размера узлов, наличия, длины и толщины ребер, их местонахождение вычисляется автоматически.

Примеры визуализации данных о конструкциях средствами модуля pattern.graph приведены на рис. 2–3.

Структура конструкции в графах отражается следующим образом: красным цветом помечен узел, содержащий целевое слово, зеленым цветом выделены ребра графа, связывающие между собой элементы разметки конструкции (лексико-семантические и морфологические теги), синим — ребра графа, связывающие теги лемм с лексико-семантическими и морфологическими тегами.

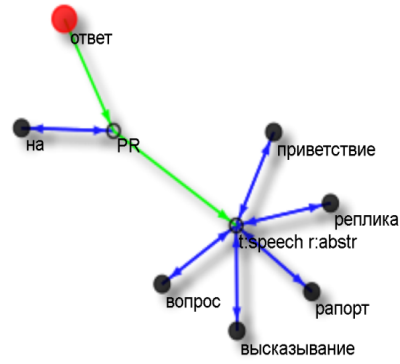


Рис. 2. Графическое представление конструкции ОТВЕТ + PR|*на* + t:speech r:abstr| *приветствие, вопрос, высказывание, рапорт, реплика*

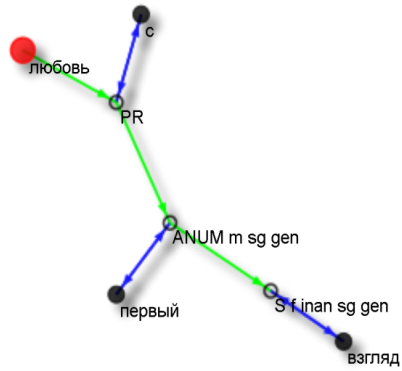


Рис. 3. Графическое представление конструкции
любовь + PR|*с* + ANUM m sg gen|*первый* + S f inan sg gen|*взгляд*

6. Заключение

Проведенные эксперименты дают основания утверждать, что

1) инструмент автоматического выделения конструкций приспособлен для обработки контекстных выборок из НКРЯ, его применение позволило получить списки конструкций для целевых существительных из лексико-семантических групп названий инструментов, обозначений речевых действий и названий эмоций;
2) полученные конструкции различаются по числу компонентов (это пары, тройки, четверки, пятерки, состоящие из тегов контекстного окружения) и по наполнению (это двухслойные структуры, в состав которых входят либо морфологические теги и теги лемм, либо лексико-семантические теги и теги лемм);
3) задача визуализации данных о выделенных конструкциях успешно решается с помощью модуля pattern.graph, позволяющего наглядно представлять организацию конструкций, иерархию и различные типы их компонентов.

В перспективе, мы бы хотели рассмотреть возможность представлять в конструкции три слоя разметки (леммы, грамматические теги, лексико-семантические теги) одновременно. Кроме того, хотелось бы учитывать статус факультативных элементов конструкции — в нынешней версии такой функционал не предусмотрен.

Модуль визуализации будет совершенствоваться с учетом пожеланий пользователей. Одной из дальнейших задач видится переход к динамической организации модуля визуализации — особенно в тех случаях, когда конструкции содержат много элементов и много лексических вариантов реализации. Предполагается рассмотреть вопрос о визуальном представлении нескольких конструкций в контексте, когда конструкции с разными лексическими

центрами «наслаиваются» друг на друга. Наконец, планируется провести сопоставление выделенных наборов лексических конструкций с наборами, который мог бы выделить лексикограф на тех же данных.

Литература

1. *Fillmore Ch. J., Kay P., O'Connor M. C.* (1988a), Regularity and idiomaticity in grammatical constructions: The case of "let alone", Language, Vol. 64–3.
2. *Fillmore Ch. J.* (1988b), The Mechanisms of Construction Grammar, Proceedings of the Berkeley Linguistic Society, Vol. 14.
3. *Firth J. R.* (1957/1968), A synopsis of linguistic theory 1930–1955, in Palmer F. R. (ed.), Selected Papers of J. R. Firth 1952–1959, Longman, London.
4. *Goldberg A. E.* (1995), Constructions. A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago, IL/London.
5. *Goldberg A. E.* (2006), Constructions at Work: the Nature of Generalization in Language, Oxford University Press, Oxford.
6. *Gries St. Th., Divjak D. S.* (2009), Behavioral profiles: a corpus-based approach towards cognitive semantic analysis, in Evans V., Pourcel S. S. (eds.), New directions in cognitive linguistics, John Benjamins, Amsterdam & Philadelphia, pp. 57–75.
7. *Lyashevskaya O. A., Mitrofanova O. A., Grachkova M. A., Shimorina A. S., Shurygina A. S., Romanov S. V.* (2012), Building the Inventory of Russian nominal Constructions [K postrojeniju inventar'a russkih imennyh konstrukcij], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2012» [Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog 2012»], RGGU, Moscow.
8. *Mitrofanova O. A., Lyashevskaya O. A., Grachkova M. A., Shimorina A. S., Shurygina A. S., Romanov S. V.* (2012), Experiments on Automatic Word Sense Disambiguation and Construction Identification (Based on Russian National Corpus) [Eksperimenty po avtomaticheskomu razresheniju leksiko-semanticheskoj neodnoznachnosti i vydeleniju konstrukcij (na materiale Nacional'nogo korpusa russkogo jazyka)], Strukturnaja i prikladnaja lingvistika. Vyp. 9 [Structural and Applied Linguistics. Vol. 9], St. Petersburg.
9. *Penn G., Carpendale Sh., Collins Chr.* (2009), Interactive Visualization for Computational Linguistics: Tutorial at ESSLLI-09, available at: esslli2009.labri.fr/documents/carpendale_penn.pdf.
10. *Sahlgren M.* (2008), The Distributional Hypothesis, Rivista di Linguistica [Italian Journal of Linguistics], Vol. 20 (1), pp. 33–53.
11. *de Smedt T., Daelemans W.* (2012), Pattern for Python, Journal of Machine Learning Research, Vol. 13.
12. *Tomasello M.* (2003), Constructing a Language: A Usage-Based Approach to Child Language Acquisition, Harvard University Press, Cambridge, MA.

ЧАСТОТНЫЙ ЛЕКСИКО-ГРАММАТИЧЕСКИЙ СЛОВАРЬ: ПРОСПЕКТ ПРОЕКТА[1]

Ляшевская О. Н. (olesar@gmail.com)
НИУ Высшая школа экономики, Москва, Россия

Обсуждается задача создания электронного частотного словаря, в котором будет отражено распределение грамматических форм в парадигме словоизменения русских имен существительных, прилагательных и глаголов, т. е. грамматический профиль индивидуальных лексем и лексических групп. В практике составления частотных словарей и квантитативных исследований стандартным объектом изучения является общая иерархия грамматических категорий, например, частотность частеречных классов или среднее соотношение частот именительного и творительного падежей. В данном проекте фокус переносится на распределение грамматических форм у конкретных лексем, выявление единиц с нестандартным перевесом тех или иных форм в парадигме. Словарь предназначен для исследований русской грамматики, грамматической семантики, а также изучения вариативности форм.

Ресурс строится на материалах Национального корпуса русского языка. В статье затрагиваются общие вопросы использования корпусов для создания частотных ресурсов подобного рода и технологии обработки данных. Предлагаются решения, связанные с отбором данных, уровнем дробности грамматических кластеров, параметрами мониторинга изменения грамматического профиля в зависимости от времени создания текста и жанрово-функционального регистра.

Ключевые слова: частотный словарь, грамматический профиль лексемы, словоизменение, грамматическая семантика, вариативность, русский язык, НКРЯ

LEXICO-GRAMMATICAL FREQUENCY DICTIONARY: A PRELIMINARY DESIGN

Lyashevskaya O. N. (olesar@gmail.com)
NRU Higher School of Economics, Moscow, Russia

A new electronic frequency dictionary shows the distribution of grammatical forms in the inflectional paradigm of Russian nouns, adjectives and verbs,

[1] В работе использованы результаты, полученные в рамках проекта № 11-01-0171, выполненного в рамках Программы «Научный фонд НИУ ВШЭ» в 2012–2013 гг.

i. e. the grammatical profile of individual lexemes and lexical groups. While the frequency hierarchy of grammatical categories (e. g. the frequency of part of speech classes or the average ratio of Nominative to Instrumental case forms) has long been the standard topic of research, the present project shifts the focus to the distribution of grammatical forms in particular lexical units. Of particular concern are words with certain biases in grammatical profile, e. g. verbs used mostly in Imperative, in past neutral or nouns used often in plural. The dictionary will be a source for many of the future research in the area of Russian grammar, paradigm structure, grammatical semantics, as well as variation of grammatical forms.

The resource is based on the data of the Russian National Corpus. The article addresses some general issues such as corpora use in compiling frequency resources and technology of corpus data processing. We suggest certain solutions related to the selection of data and the level of granularity of grammatical profile. Text creation time and language registers are discussed as parameters which may shape the grammatical profile fluctuations.

Key words: frequency dictionary, grammatical profile, inflection, semantics of grammar, form variation, Russian, Russian National Corpus

1. Введение

Частотный лексико-грамматический словарь продолжает серию частотных словарей, создаваемых на данных Национального корпуса русского языка, и является прямым продолжением частотного словаря (Ляшевская, Шаров 2009). В общем частотном словаре основная доля информации была представлена на уровне лексем. Из грамматической информации давались сведения о доле слов разных частей речи и о наиболее частотных словоформах русского языка. Вместе с тем, если смотреть с точки зрения конкретной лексемы, информации о частоте всех ее словоформ словарь не давал. Эту лакуну заполняет новый экспериментальный лексико-грамматический словарь. Он представляет грамматический профиль (т. е. распределение грамматических форм в парадигме словоизменения) 5000 наиболее частотных русских имен существительных, прилагательных и глаголов.

Далее в статье речь пойдет о задачах словаря, его структуре, а также о некоторых проблемных точках, связанных с обработкой и интерпретацией частотных данных.

2. Предназначение словаря

Квантитативныеисследованиянелексическихединицязыка—грамматических классов (например, иерархий падежного маркирования), грамматических форм внутри парадигмы конкретного слова, вариативности грамматических

и лексико-грамматических единиц, вариативности падежного и предложно-падежного оформления зависимых — были признаны необходимой составляющей лингвистического анализа еще в мировой лингвистике 50–70-х годов XX в. В русистике были получены замечательные результаты в классических работах Штейнфельдт 1963, Greenberg 1974, Граудина и др. 1976, Апресян 1967 и мн. др.). Однако именно появление представительных и сбалансированных лингвистических корпусов объемом от ста миллионов словоупотреблений и выше поставило эти исследования на принципиально новый уровень, как в плане используемых математических статистических моделей и компьютерных технологий, так и в плане осмысления частотных результатов и их устойчивости.

В теоретической лингвистике частотные исследования приобрели особую актуальность в связи с постулированием usage-based model — модели языка, предполагающей, что частота употребления языковых единиц оказывает непосредственное влияние на их конструктивные свойства, статус в системе, вариативность и изменение в истории языка (Kemmer & Barlow 2000). Еще одна гипотеза — о семантической мотивированности грамматических явлений — верифицируется в ходе исследований, изучающих сдвиги частот форм в разных лексико-семантических классах (см. об этом Janda, Lyashevskaya 2011): например, предполагается, что преобладание форм императива несовершенного вида связано с семантическими и функциональными особенностями лексических единиц. В когнитивных исследованиях изучается также гипотеза о том, что возможности языковой памяти таковы, что в частотных фрагментах человек оперирует единицами, большими чем слово (pre-fabricated units). Поднимается и вопрос, оперирует ли человек лексемами, т. е. единицами абстрактного уровня, или же это порождение грамматической схоластики, и человек опирается в своем языковом опыте исключительно на словоформы (Newman 2008). Наконец, изучение грамматических частотных профилей в разных языках помогло бы извлечь новые факты для лингвистической типологии и истории развития языков.

В грамматике русского языка, и теоретической, и практической, традиционно большую роль играет вопрос о дефектных парадигмах, а также о вариативных формах словоизменения. Несмотря на получившую общее признание точку зрения о градуальности таких явлений, как, например, singularia и pluralia tantum, выявление ассоциированных с ними лексических единиц и описание их функционирования все еще нуждается в эмпирических данных. То же можно сказать и о проблематике появления, со-существования и исчезновения вариативных форм типа род. мн. *помидор/помидоров,* прош. ед. *стыл/стынул*, статусе «вторых» падежей и т. п.

В преподавании родного и иностранных языков знание о частотных фактах грамматики позволяет выстроить правильную последовательность изучения грамматических тем (например, порядок изучения падежей), соотнести грамматические категории с теми лексемами, при которых они чаще всего встречаются, изучать лексику в контексте (знать самые частотные сочетания), выбирать для образца тексты, наиболее подходящие по жанрово-стилевому признаку к изучаемой грамматической теме и т. п.

И, конечно, неоценимую роль играют частотные данные в разработках систем автоматической обработки текста. Особенно это стало очевидно в эпоху стремительного развития алгоритмов машинного обучения, построенных на вероятностях. Грамматические и сочетаемостные преференции слов учитываются в синтаксических парсерах, системах разрешения неоднозначности, средствах исправления орфографии, распознавания текста, в онтологических расширениях поисковых систем и др.

Несмотря на то, что задача построения частотной русской грамматики и фронтального изучения грамматической вариативности осознана и ставится в литературе (Мустайоки 1973, Baerman et al. 2010), в настоящее время не существует ни одного сколько-нибудь полного лексикографического ресурса, приближающего нас к этой цели. Ресурс на материале НКРЯ дает уникальную возможность ответить на многие исследовательские вопросы, исходя из современных возможностей корпусной лингвистики.

3. Общая температура по больнице, или почему не всегда помогает статистика падежей

Когда говорят о частотной грамматике языка, в первую очередь, имеют в виду соотношения частот частеречных классов, падежей и других грамматических категорий. Особенно популярна тема частотного распределения падежей — в работе Копотев 2008 цитируются три исследования, появившихся только в 1959-1961 гг., что касается настоящего времени, то, как показывает веб-поиск, аналогичные работы, построенные на разных текстовых выборках, плодятся с невиданной скоростью. Работа самого М. Копотева привлекает внимание к устойчивости частотных данных на больших корпусах (см. табл. 1). Его вывод — в том, что современные корпуса довольно хорошо согласуются друг с другом в оценке средней вероятности появления падежей, а различия кроются в жанровой принадлежности текстов.

Табл. 1. Частотное распределение шести падежей по данным (Копотев 2008)

	И	Р	Д	В	Т	П
□ НКРЯ	27,06	29,23	5,98	18,66	8,44	10,63
■ ХАНКО	24,30	32,62	5,50	17,73	8,08	11,78
▨ J. 1953	38,80	16,80	4,70	26,30	6,50	6,90
▬ Št. 1963	33,60	24,60	5,10	19,50	7,80	9,40

Однако, легко видеть, что принцип «выбирай родительный, если забыл — не ошибешься» может сыграть злую шутку со студентом РКИ, в случае, если ему нужно употребить слово *шепот*. Как показывает табл. 2[2], распределение

[2] Здесь и далее в таблицах приведены данные по корпусу со снятой лексико-грамматической омонимией НКРЯ.

частот падежей у некоторых существительных может разительно отличаться от средней картины.

Табл. 2. Частотный грамматический профиль лексем *шепот, поза, тропинка* (падежные формы ед. числа)

	И	Р	Д	В	Т	П	Всего (F.abs)
шепот	10,9%	3,7%	0,9%	8,3%	**75,6%**	0,6%	349
поза	15,9%	6,3%	0,8%	19,0%	4,0%	**54,0%**	126
тропинка	27,6%	2,0%	**52,0%**	5,1%	5,1%	8,2%	98

Дж. Гринбергу принадлежит наблюдение, что разные семантические группы должны иметь разную дистрибуцию падежей (как в предложных, так и в беспредложных употреблениях), иными словами, средние значения падежных показателей в группе имен абстрактных качеств (или имен частей тела, или названий мер) должны отличаться от средних значений по всему массиву лексики (Greenberg 1974/1991). Выбор русского языка как объекта исследования Гринберга был не случаен — именно в тот момент русский язык, один из немногих, располагал частотным списком форм падежей и предложно-падежных сочетаний имен существительных, входившим в состав замечательного частотного словаря Э. Штейнфельдт (Šteinfeldt 1963). Гринберг искал «волшебное» соотношение, которое позволяло бы отнести слово к тому или иному семантическому классу — и, естественно, не нашел его. Позднее его наблюдение было реинтерпретировано как семантически мотивированный сдвиг частот грамматических форм. Например, большую долю форм творительного падежа *шепотом* легко объяснить пересечением в семантике грамматической формы (творительный способа) и семантике лексемы (*шепот* как способ произнесения); форм предложного падежа *(в) позе* — связью между стативной семантикой существительного и семантикой локативной группы *в + S.loc*, наиболее типичном контекстном варианте употребления этого слова. Аналогичным образом, преобладание форм датива у существительного *тропинка* объясняется тем, что слова со значением траектории — идеальный лексический наполнитель предложной группы *по + S.dat*.

В работе (Janda, Lyashevskaya 2011) мы ввели понятие грамматического профиля лексемы — как инструмента для изучения семантических и функциональных причин девиаций грамматических форм. Исследование поведения форм вида, времени и наклонения, частности, показало предсказуемые частотные эффекты в разных клетках парадигмы: в императиве несовершенного вида — для глаголов привлечения внимания, вежливой просьбы, лексики, относящейся к культурному фрейму встречи гостей и т. п., ср. *раздевайтесь, садитесь, присоединяйтесь, закусывайте, закуривайте, ступайте, прощайте*; в инфинитиве совершенного вида — для глаголов, в которых заложена презумпция труднодостижимого результата (вследствие этого они часто употребляются в контексте глаголов попытки, модальных предикативов, в целевых придаточных и т. п., ср. *попытался/тяжело было/чтобы восполнить*) и т. п.

В исследовании (Kuznetsova 2013) вводятся классы типичных «женских» и типичных «мужских» глаголов — соотношение форм мужского и женского рода у глаголов типа *вышивать* и глаголов типа *надвинуть* будет разным.

На материале BNC С. Райс и Дж. Ньюман (Rice, Newman 2005, Newman 2008) сделали наблюдение, что разброс грамматического распределения может присутствовать и внутри лексических групп. Они показали, что даже близкие по смыслу слова, английские *think, know* и *mean,* могут иметь значительную диспропорцию форм времени, лица и числа, и назвали это явление “inflectional islands”. Объяснение этого явления кроется в индивидуальных семантических особенностях каждого глагола, в способности присоединять разные типы субъектов и т. п. В (Janda, Lyashevskaya 2011) указывается также большой вклад устойчивых конструкций в формирование тех или иных грамматических «флюсов» у индивидуальных лексем, ср. *мне плевать, мне наплевать, на чужой каравай рот не разевай, хоть залейся, поминай, как звали*.

Однако, наиболее удивительный факт русской лексической системы состоит в том, что почти не существует существительных, грамматический профиль которых соответствовал бы «среднему» профилю нарицательной лексики, глаголов со «средней» пропорцией форм времени-лица-числа и т. п. По сути, мы имеем дело со сложным наслаиванием семантических особенностей, сочетаемостных и конструктивных свойств, которые суммарно влияют на частотный выход.

4. Обработка корпусных данных

Основная часть словаря строится на данных 1900–2010 гг., в диахронической части привлекаются данные, начиная с 1800 г. Данные для «малого» словаря были собраны по корпусу со снятой лексико-грамматической омонимией (5,4 млн словоупотреблений, стандартная коллекция), для «большого» словаря — по основному, газетному, поэтическому и устному корпусу. Сбор осуществлялся с учетом функциональных стилей и жанров текста, а также с учетом времени создания.

Прежде всего, была собрана статистика по словоформам с лексико-грамматическим разбором (лемма, часть речи, словоизменительные характеристики)[3], разметкой лексико-семантического класса капитализации написания. Были также собраны 2- и 3-грамы, отражающие статистику предложно-падежных сочетаний существительных и местоимений.

Для «борьбы» с грамматической омонимией словоформ внутри парадигм и между парадигмами использовалась автоматически дизамбигуированная версия основного, газетного, поэтического и устного корпуса. Она была создана с применением двух программ — модуля на эвристиках и HMM-модуля, обученного на текстах снятого вручную корпуса. Небольшая часть данных дополнительно корректировалась вручную.

Особо отметим, что большую проблему для дизамбигуации представляют ингерентно-пересеченные парадигмы, например, парадигмы мужского

[3] Использовались стандартные соглашения словаря (Ляшевская, Шаров 2009).

и женского рода имени *рояль* или парадигмы прилагательных вида *запасной* и *запасный*. Устаревший вариант женского рода существительного распознается словарем лишь в формах, не предусмотренных в парадигме мужского рода (*роялью*), и тем самым, в словаре отражается искусственно дефектная парадигма. Пересеченные парадигмы прилагательных, различающихся лишь в именительном падеже, также разводятся плохо, поскольку модели дизамбигуации не предусматривают столь тонкой настройки, да и вручную в письменном корпусе далеко не всегда удается однозначно определить лексему. Такие точечные места в словаре, где информация может быть недостоверна по причине несовершенной дизамбигуации, снабжаются специальной пометой.

5. Виды частотной информации в словаре

Пользователь имеет возможность пользоваться двумя наборами данных. «Малый» словарь представляет наиболее аккуратные результаты в смысле разведения омонимов. Однако в корпусе со снятой вручную омонимией многие многие грамматические формы частотных лексем могут быть либо не представлены вообще, либо встречаются редко, и следовательно, не могут показать достоверное распределение форм. «Большой» словарь строится на корпусах НКРЯ, в десятки раз превосходящих «снятник», однако следует учитывать, что в некоторых зонах (например, в зоне противопоставления родительного и винительного падежа одушевленных существительных) информация в нем менее достоверна.

5.1. Грамматические категории

Пользователь может выбрать данные как по всем грамматическим формам парадигмы, так и по более крупным кластерам форм. Например, могут быть приведены суммарные данные по формам полных пассивных причастий (без учета признаков падежа, числа и рода), по четырем формам прошедшего времени глагола, по всем формам единственного VS множественного числа существительного. Информация о падежных распределениях существительных и местоимений дополнена сведениями о распределении предложных конструкций, в которых задействован тот или иной падеж. Кроме того, можно получить сопоставительные данные для написаний с прописной VS строчной буквы.

5.2. Омонимия и вариативность

Из всей парадигмы пользователю могут быть выданы сведения только об омонимичных формах (в т. ч. внутрипарадигматическая омонимия, ср.

солдат — им. ед. и род. мн., омонимия форм, принадлежащих разным парадигмам, ср. *заплыв* — формы имени существительного и глагола, см. Венцов, Касевич 2004). Предоставляются сведения о соотношении частот вариантов грамматических форм (например, *сильней* и *сильнее*, *дверями* и *дверьми*), так наз. «основных» и «вторых» падежей, различающихся на письме (ср. *без толка* и *без толку*), и других секундарных форм (ср. *сильней* и *посильней*).

5.3. Распределение по годам и жанрам

Информация об изменении грамматических профилей во времени дается в 10-летних интервалах; в газетном корпусе учитываются интервалы в 1 год. Пользователь может увидеть распределения в художественной прозе, в поэзии, в периодике, в бытовой, учебно-научной и т. п. сферах нехудожественной литературы, в электронной коммуникации, а также в устной непубличной речи.

5.4. Единицы измерения

Пользователь может выбрать один или несколько вариантов представления частотной информации:

- количество текстов корпуса, в которых встретились формы;
- абсолютная частота вхождений и размер корпуса;
- частота в ipm;
- иерархия форм у рассматриваемой единицы/класса вида

  Loc > Gen > Nom > Acc > Dat > Ins;

- процентное распределение (см. табл. 3) и попарное соотношение форм;
- квинтильное распределение каждой из форм, например, положение формы предложного падежа единственного числа слова *велосипед* в первой, второй… пятой порции списка, в котором представлены формы предложного падежа единственного числа всех существительных (а — самые редкие, д — самые частые, см. табл. 4).

Табл. 3. Профиль падежных форм лексемы *влияние*: абсолютное и относительное распределение

	И	**Р**	**Д**	**В**	**Т**	**П**	**Всего (F.abs)**
sg	98	128	29	170	137	14	576
pl	4	9	3	7	2	2	27
	И	**Р**	**Д**	**В**	**Т**	**П**	**Всего (%)**
sg	17,0%	22,2%	5,0%	29,5%	23,8%	2,4%	100,0%
pl	14,8%	33,3%	11,1%	25,9%	7,4%	7,4%	100,0%

Табл. 4. Квинтильное распределение падежных форм ед. числа в группе имен транспортных средств

Лемма	И	Р	Д	В	Т	П	Всего (F.abs)
метро	а	д	г	а	а	д	185
корабль	д	в	б	б	а	в	231
грузовик	д	г	в	б	б	в	134
пароход	д	д	а	б	в	г	121
автомобиль	г	г	в	б	б	г	441
поезд	д	в	в	б	б	г	618
самолет	г	в	г	в	в	г	385
трамвай	г	б	в	г	в	г	198
лодка	г	в	б	г	б	г	280
вагон	а	г	г	в	а	д	473
велосипед	б	в	а	г	б	д	206
автобус	г	б	в	в	б	д	281
такси	в	а	б	д	а	д	174

Оговорим, что пользователь может выбрать разные методики расчета соотношений частот в парадигме, известных из литературы. За основу сравнения (100%) может быть принята вся парадигма (т. е. сумма всех частот грамматических форм), некоторая базовая часть (например, парадигма глагола за вычетом форм причастий и деепричастий), приоритетная форма (например, сумма форм прошедшего времени), а также доля употреблений двух форм относительно друг друга (например, отношение частоты форм женского рода к частоте форм мужского рода).

5.5. Сравнение лексем. Классы

Информация в словаре разнесена на несколько уровней. Первый уровень — индивидуальные грамматические профили лексем. На втором уровне даются сведения для крупных лексико-семантических классов (в классификации НКРЯ), например, для глаголов движения, имен инструментов и т. п. Третий уровень — распределение грамматических частот на уровне частеречного класса (словарь также дает справочную информацию о встречаемости самих частеречных классов, а также именных и глагольных грамматических категорий).

Таким образом, информация об индивидуальных лексемах может быть сопоставлена с данными по их лексико-семантическому классу и, шире, со средним грамматическим профилем части речи. Возможно сопоставление грамматических профилей нескольких лексем между собой.

6. Заключение

Словарь адресован, в первую очередь, исследователям русского словоизменения, грамматической семантики, тем, кто изучает вариативность грамматической нормы. Вместе с тем, нужно заметить, что «лексикоцентричный» подход, несмотря на ресурсоемкость и неплотность данных, может оправдывать себя и в автоматической обработке текста. В частности, в экспериментах (Данилова и др. 2013) показано, что учет лексического фактора позволяет повысить качество автоматической дизамбигуации лексико-грамматической омонимии на 3 %.

Электронная форма словаря позволяет постоянно совершенствовать его. Во-первых, планируется развивать функционал с учетом пожеланий пользователей, в частности, дополнить словарь модулем графического представления результатов, подключить внешние словари (словарь вариантов, словообразовательный и т. п.) и др. Во-вторых, будет совершенствоваться качество даных за счет улучшения дизамбигуации корпусных данных и работы с сообщениями пользователей об ошибках. В-третьих, увеличение объема словаря: включение новых лексических данных, добавление информации об авторе и т. п., — требует дополнительных исследований, поскольку работа с малыми частотами (sparse data) требует особой осторожности и особых техник.

Главный вопрос — в том, как интерпретировать полученные данные, каким образом переносить сведения о статистических вероятностях на другие текстовые корпуса и как научиться делать аккуратные выводы о функционировании грамматических форм в целом. Предлагаемый словарь — лишь первый опыт составления большого лексико-грамматического ресурса, и, соответственно, станет благодатным материалом для исследования достоверности корпусных данных. Безусловно, мы должны лучше понимать структуру выборок, как она связана с устойчивостью статистических данных, научиться балансировать выборки для разных временных срезов, провести множество экспериментов с полученным лексическим материалом для того, чтобы достоверность интерпретации корпусных данных перестала вызывать вопросы.

Литература

1. *Baerman M., Brown D., Corbett G. G., Krasovitsky A., Williams P.* (2010), Predicate agreement in Russian: A corpus-base approach, Wiener Slawistischer Almanach, Sonderband 74, pp.109–121.
2. *Greenberg J. H.* (1974/1990), The relation of frequency to semantic feature in a case language (Russian), in Denning K., Kemmer S. (eds), On Language, Selected Writings of Joseph H. Greenberg, Stanford, pp. 207–226.
3. *Ilola E., Mustajoki A.* (1989), Report on Russian Morphology as it Appears in Zaliznyak's Grammatical Dictionary, (Slavica Helsingiensia 7), Helsinki.
4. *Janda L. A., Lyashevskaya O.* (2011), Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian, Cognitive Linguistics, 22 (4), pp. 719–763.
5. *Kemmer S., Barlow M.* (2000), A Usage-Based Conception of Language, Essen, 2000.
6. *Kuznetsova J.* (2013), Linguistic Profiles: Correlations between Form and Meaning. Ph.D. diss., University of Tromsø.
7. *Newman J.* (2008), Aiming low in linguistics: Low-level generalizations in corpus based research. Proceedings of the 11th International Symposium on Chinese Languages and Linguistics, National Chiao Tung University, Hsinchu, Taiwan, May 24, 2008.
8. *Rice S., Newman J.* (2005), Inflectional islands, ICLC-9, Yonsei University, Seoul, Korea.
9. *Šteinfeldt E.* (1963), Russian Word Count, Moscow.
10. *Апресян Ю. Д.* (1967), Экспериментальное исследование семантики русского глагола, М.
11. *Венцов А. В., Касевич В. Б.* (ред.) (2004), Словарь омографов русского языка, СПб.: Филологич. ф-т СПбГУ.
12. *Граудина Л. К., Ицкович В. А., Катлинская Л. П.* (1976), Грамматическая правильность русской речи. Стилистический словарь вариантов. М.
13. *Данилова В., Волков О., Ладыгина А., Привознов Д., Сербинова И., Сим Г.* (2013). Снятие омонимии методом НММ (рукопись).
14. *Копотев М.* (2008), К построению частотной грамматики русского языка: падежная система по корпусным данным // Мустайоки А., Копотев М. В., Бирюлин Л. А., Протасова Е. Ю. (ред.), Инструментарий русистики: корпусные подходы, Хельсинки.
15. *Ляшевская О. Н., Шаров С. А.* (2009), Частотный словарь современного русского языка (на материале Национального корпуса русского языка), М.: Азбуковник.
16. *Мустайоки А.* (1973), Опыт составления частотной грамматики русских существительных, Хельсинки, (рукопись).

References

1. *Apresjan Ju. D.* (1967), Experimental research on the semantics of the Russian verb [Eksperimental'noe issledovanie semantiki russkogo glagola], Moscow.
2. *Baerman M., Brown D., Corbett G. G., Krasovitsky A., Williams P.* (2010), Predicate agreement in Russian: A corpus-base approach, Wiener Slawistischer Almanach, Sonderband 74, pp. 109–121.
3. *Danilova V., Volkov O., Ladygina A., Privoznov D., Serbinova I., Sim G.* (2013). Disambiguation with HMM [Snjatie omonimii metodom HMM] (manuscript).
4. *Graudina L. K., Ickovich V. A., Katlinskaja L. P.* (1976), Correct Russian speech: Stylistical dictionary of grammatical choices [Grammaticheskaja pravil'nost' russkoy rechi. Stilisticheskiy slovar' variantov]. Moscow.
5. *Greenberg J. H.* (1974/1990), The relation of frequency to semantic feature in a case language (Russian), in Denning K., Kemmer S. (eds), On Language, Selected Writings of Joseph H. Greenberg, Stanford, pp. 207–226.
6. *Ilola E., Mustajoki A.* (1989), Report on Russian Morphology as it Appears in Zaliznyak's Grammatical Dictionary, (Slavica Helsingiensia 7), Helsinki.
7. *Janda L. A., Lyashevskaya O.* (2011), Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian, Cognitive Linguistics, 22 (4), pp. 719–763.
8. *Kemmer S., Barlow M.* (2000), A Usage-Based Conception of Language, Essen, 2000.
9. *Kopotev M.* (2008), Towards the frequency grammar of Russian: corpus evidence on the grammatical case system [K postroeniju chastotnoy grammatiki russkogo jazyka: padezhnaja sistema po korpusnym dannym] // Mustayoki A., Kopotev M. V., Birjulin L. A., Protasova E. Ju. (eds.), Instruments of Russian linguistics: corpus approaches [Instrumentariy rusistiki: korpusnye podkhody], Helsinki.
10. *Kuznetsova J.* (2013), Linguistic Profiles: Correlations between Form and Meaning. Ph.D. diss., University of Tromsø.
11. *Lyashevskaya O. N., Sharoff S. A.* (2009), Frequency dictionary of modern Russian based on the Russian National Corpus [Chastotnyy slovar' sovremennogo russkogo jazyka (na materiale Nacional'nogo korpusa russkogo jazyka)], Azbukovnik, Moscow.
12. *Mustajoki A.* (1973), On compiling the frequency dictionary of Russian nouns [Opyt sostavlenija chastotnoy grammatiki russkikh suschestvitel'nykh], Hel'sinki, (manuscript).
13. *Newman J.* (2008), Aiming low in linguistics: Low-level generalizations in corpus based research. Proceedings of the 11th International Symposium on Chinese Languages and Linguistics, National Chiao Tung University, Hsinchu, Taiwan, May 24, 2008.
14. *Rice S., Newman J.* (2005), Inflectional islands, ICLC-9, Yonsei University, Seoul, Korea.
15. *Šteinfeldt E.* (1963), Russian Word Count, Moscow.
16. *Ventsov A. V., Kasevich V. B.* (eds.) (2004), Dictionary of Russian homographs [Slovar' omografov russkogo jazyka], St.-Petersburg.

ВМЕСТЕ ИЛИ РАЗДЕЛЬНО? ЗАМЕТКИ О СЕМАНТИЧЕСКОЙ КАТЕГОРИИ ПАРНОСТИ В РУССКОМ ЯЗЫКЕ[1]

Микаэлян И. Л. (irina-mikaelian@yandex.ru)
Университет штата Пенсильвания, Стейт Колледж, США

Зализняк Анна А. (anna.zalizaniak@gmail.com)
Институт языкознания РАН,
Институт проблем информатики РАН, Москва, Россия

Задача работы состоит в уточнении, на основании корпусных данных, наших представлений о грамматических и семантических свойствах русских собирательных числительных. В центре внимания находится слово *двое*, которое рассматривается в сопоставлении с другими количественными словами, включающими в свое значение сему ‘два’ — числительными *два* и *оба*, а также с существительным *пара*. Продемонстрирована актуальность для русского языка семантической категории «парности», которая обнаруживает себя в особенностях сочетаемости с числительными неодушевленных существительных типа *туфли, сапоги, глаза* и одушевленных типа *родители, супруги* (которые предлагается обозначить термином gemina tantum). Семантический анализ слов *двое* и *оба* в контексте имен лиц показал, что эти слова практически никогда не взаимозаменимы в силу расхождения их презумпций и импликаций, при сходстве ассерций.

Ключевые слова: русский язык, собирательные числительные, Национальный корпус русского языка, pluralia tantum, категория парности, семантика, грамматика

TOGETHER OR SEPARATELY? ON THE SEMANATIC CATEGORY OF TWONESS IN RUSSIAN

Mikaelian I. L. (irina-mikaelian@yandex.ru)
The Pennsylvania State University, State College, PA, USA

Zalizaniak Anna A. (anna.zalizaniak@gmail.com)
Institute of Linguistics, Russian Academy of Sciences;
Institute of Informatics, Russian Academy of Sciences,
Moscowm Russia

[1] Работа выполнена при финансовой поддержке РГНФ, грант № 11-04-00105а.

This paper attempts to refine our understanding of the grammatical and semantic features of the Russian collective numerals using data of corpora. The focus of our attention is the word *dvoe* considered in comparison with other quantity words comprising the meaning 'two' in their semantics, i. e. the numerals *dva* 'two' and *oba* 'both', as well as the noun *para* 'pair, couple'. The importance for the Russian language of the semantic category of "twoness" has been shown, and a new term *gemina tantum* has been introduced to designate the class of nouns that tend to be used in plural form and normally refer to two objects forming a pair or a couple, cf. *shoes, boots, eyes, parents, spouses.* Semantic analysis of the words *dvoe* and *oba* in the context of human nouns has shown that these words practically never interchange because, despite similar assertions, they carry different presuppositions and implications.

Key words: Russian language, collective numerals, Russian National Corpus, pluralia tantum, category of twoness, semantics, grammar

1. Вводные замечания

В состав собирательных числительных традиционно включают числительные *оба, двое, трое, четверо, пятеро, шестеро, семеро, восьмеро, девятеро, десятеро*. Обычно упоминаются также еще десять (реже, одиннадцать) «потенциальных», редко употребительных лексем (*одиннадцатеро — двадцатеро* и *тридцатеро*). К этому иногда присовокупляют также разговорное *по́лтороо*. С другой стороны, не принято включать в список собирательных числительных слово *одни*, хотя для этого как раз имеются системные основания, ср. *одни сутки, одни сапоги* = 'одна пара сапог', *одни проводы*. Слово *оба* обладает набором существенных морфосинтаксических и семантических особенностей, которые позволяют рассматривать его как отдельную, не входящую в класс собирательных числительных, лексему. Так, И. А. Мельчук причисляет его к количественным числительным, признавая, что такое решение также не совсем удовлетворительно [Мельчук 1985: 38]; в [Сичинава 2012] оно упоминается среди местоимений-числительных. Каждое из наиболее употребительных малых собирательных числительных (*двое*, *трое* и *четверо*) также обладают индивидуальными свойствами (см., в частности, [Янко 2002]), и все три функционально противопоставлены «большим» собирательным числительным.

Изучение русских собирательных числительных имеет достаточно обширную традицию. Наиболее последовательный и глубокий анализ проделан в книге [Мельчук 1985]; однако, как признает сам Игорь Мельчук, "<…>в области употребления числительных и в частности лично-количественных (т. е. собирательных, *прим. авторов*) числительных колебания узуса особенно сильны. Поэтому правила [их] употребления должны были бы опираться на результат широкого лингвостатистического и социо-лингвистического обследования" [Мельчук 1985: 403]. Такого рода исследования частично проводились еще в 50е — 60е годы, в частности, в работах А. Е. Супруна и Ю. И. Щербакова (см. обзор и библиографию в [Nikunlassi 2000]), но стали особенно актуальны

в последние 15 лет, когда появилась возможность работы с корпусами и с большими массивами языкового материала, доступного в интернете. Самое полное описание употребления собирательных числительных на сегодняшний день представлено в экспериментальном исследовании [Nikunlassi 2000]. Существенные дополнения содержатся в работах [Янко 2002], [Никунласси 2006], [Добрушина, Пантелеева 2008], [Наконечная-Лаланн 2013]. Особо отметим статью о числительных, опубликованную на сайте rusgram.ru [Сичинава 2012], которая написана на базе Национального корпуса русского языка (далее — НКРЯ).

Во всех перечисленных работах собирательные числительные изучаются в сопоставлении с количественными. Однако не меньший интерес представляет сопоставление слова *двое* с другими количественными словами, обозначающими два предмета. С одной стороны, это слово *оба*, традиционно относимое также к собирательным числительным (см. например, [Грамматика-80]), а с другой — слово *пара*. Слова *двое* и *оба*, при видимой близости их значения (так, например, на английский язык оба слова могут переводиться как *both* или (*the*) *two* (*of them*)), почти никогда не бывают взаимозаменимы. Как выясняется, в семантической структуре слов *двое* и *оба* имеется лишь один общий компонент (см. раздел 4). Слово *пара* конкурирует со словом *двое* и частично вытесняет последнее в контексте неодушевленных имен (см. раздел 2).

2. Сочетаемость собирательных числительных. Множественность и парность

Как известно, существует два полюса употребления малых собирательных числительных, соответствующие их двум различным функциям. С одной стороны, они заменяют числительные *два, три* и *четыре* в контексте неодушевленных имен класса pluralia tantum, ср. *двое суток, трое часов, четверо ножниц*, в случае, если числовое выражение стоит в прямом падеже (им. или вин.); в косвенных падежах в данном контексте возможны только количественные числительные, ср.: *прождал* ***двое*** *суток*, но: *после* ***двух*** *суток ожидания уехал*. Поскольку такое распределение обусловлено исключительно морфосинтаксическими причинами, собирательные числительные в данном употреблении могут рассматриваться как члены парадигмы числительных *два*, *три* и *четыре* (см., например, [Nikunlassi 2000]).

В другой своей функции малые собирательные числительные сочетаются исключительно с одушевленными именами, причем тяготеют к сочетанию с существительными, обозначающими лиц мужского пола невысокого социального ранга (ср. *двое солдат, трое студентов*); при этом именные группы с собирательными числительными обладают конкретно-референтным статусом ([Мельчук 1985] и др). Что касается сочетаемости с существительными, обозначающими лиц женского пола, то хотя учебные грамматики запрещают такое употребление, даже нормативные академические грамматики признают, что оно возможно и даже весьма распространено, что подтверждается

корпусными данными (ср. *двое дочерей*). Сочетание собирательных числительных с именами лиц весьма высокого ранга тоже вполне допустимо, хотя и нуждается в более сильной контекстной поддержке, чем более распространенные словосочетания (ср. *двое солдат* vs. *двое генералов*), см. раздел 3.3.

Термин pluralia tantum в русской грамматической традиции употребляется в значении «существительное, не имеющее формы ед. числа», и применяется к таким словам как *сани, сени, ножницы, часы, сутки, похороны, посиделки* и т. п., обозначающим единичные объекты, а также к существительным несчетным, таким как *щи* или *будни* (ср. [Виноградов 2001], [Грамматика-80]). Все эти существительные являются неодушевленными. В книге [Мельчук 1985: 386] вводится категория «одушевленные pluralia tantum», к которой автор причисляет слова *дети, девчата, ребята* (и их уменьшительные варианты). Такое словоупотребление, хотя и не является общепринятым и расширяет границы класса pluralia tantum, но представляется вполне допустимым. И. А. Мельчук относит к данной категории в том числе слово *дети* — вопреки традиции считать его супплетивной формой мн.числа к *ребенок*. По-видимому, сюда же следует отнести слово *люди*. Хотя традиционно считается, что *люди* — это супплетивное мн. число от *человек*, против этой точки зрения имеются достаточно серьезные аргументы [Шмелев 2009]. К классу «одушевленные pluralia tantum» примыкают, с одной стороны, такие слова как *родители, молодожены, новобрачные, молодые,* которые преимущественно используются во множественном числе и обозначают пару, состоящую из лиц противоположного пола, а с другой — *родные* и *близкие*, которые не имеют формы ед. ч.[2]

Как справедливо отмечает А. Вежбицкая в связи с обсуждением слов *очки, ножницы, трусы, брюки*, *сани*, *носилки* и т. п., «семантика парности» в русском языке обладает особым, выделенным статусом [Wierzbicka 1996: 396] [3]. Мы предлагаем ввести категорию pluralia gemina — куда попадают существительные, которые в форме множественного числа имеют, наряду с обычным значением мн. числа, значение «естественной пары»; сюда относятся, в частности: *сапоги, туфли, ботинки, башмаки, тапочки, валенки, лыжи, коньки, носки, гольфы, перчатки, варежки, рукавицы; губы, глаза, уши, руки, ноги*, а также *супруги, молодожены, родители, бабушки* и *дедушки* в значении ‘родственники одного и того же лица’. Так, *родители* (в значении ‘мама и папа <определенного лица>’) — это pluralia gemina, в отличие от обычного мн. числа, используемого, например, для обозначения множества людей, пришедших на родительское собрание.

Класс pluralia gemina естественным образом распадается на одушевленные и неодушевленные существительные, которые по-разному ведут себя

---

[2] Компания Билайн пытается преодолеть эту дефектность слова *близкие*, используя формулы типа «Чтобы пополнить счет *близкого* с вашего мобильного телефона, подключите услугу и т. д.», однако пока такое словоупотребление представляется ненормативным.

[3] Обратим внимание на не имеющее аналогов в европейских языках русское слово *сутки*, обозначающее пару, которую образуют день и ночь, буквально: «стык дня и ночи» [Фасмер 1996, III: 811].

отношении квантификации. И те и другие более частотны во множественном числе и имеют тенденцию реализовать парное значение, однако между ними имеются существенные различия.

В контексте числительного *двое* неодушевленные существительные в форме мн. числа (*сапоги, перчатки*) в норме реализуют парное значение, т. е. соотносятся с парой предметов; соответственно, *двое сапог* означает ‘две пары сапог’. Такое употребление воспринимается как стилистически сниженное, хотя оно является относительно распространенным. Еще более ограничено употребление числительного *двое* в контексте парных частей тела для обозначения двух пар этих предметов (??*двое глаз = две пары глаз,* ??*двое рук = две пары рук*). В НКРЯ такие словосочетания не встречаются, однако, как это ни странно, они приводятся без особых стилистических помет в словарях Ушакова и в МАС. И. А. Мельчук считает оба употребления решительно устаревшими [Мельчук 1985: 385].

Одушевленные существительные класса pluralia gemina ведут себя иначе. В конструкции с собирательным числительным реализуется значение обычного мн. числа: *На собрание пришло только четверо родителей* (= ‘четыре человека’, но не: ‘четыре пары’). Соответственно, сочетание *двое родителей* обозначает двух человек (а не две пары); в зависимости от контекста, это могут быть родители одного и того же ребенка или двух разных, ср.: *на доверенности требуется подпись двоих родителей* (= ‘мамы и папы <одного ребенка>’), *на собрание пришло только двое родителей* (= ‘два родителя <возможно, разных детей>’)[4].

3. Числительное *ДВОЕ*

Итак, употребления числительного *двое* включают два полюса: с одной стороны, неодушевленные существительные pluralia tantum, чье сочетание с собирательными обусловлено чисто морфосинтаксическими причинами, а с другой — сочетания с именами лиц, где выбор собирательного числительного обусловлен семантикой (имени или контекста в целом). Между этими полюсами лежит промежуточная зона одушевленных существительных (pluralia tantum и pluralia gemina, см. выше). При этом, с одной стороны, семантические механизмы выбора собирательного числительного действуют также в зоне неодушевленных имен, а именно pluralia gemina, а с другой — морфосинтаксис играет некоторую роль и в выборе типа числительного в контексте имен лиц (см. раздел 3.2).

3.1. *ДВОЕ* в контексте неодушевленных имен

В контексте неодушевленных существительных pluralia tantum конструкция с собирательным числительным конкурирует с конструкцией со словом

[4] При том, что словосочетание ?*два родителя* малоупотребительно.

пара. В таблице 1 представлен материал, полученный поиском в НКРЯ (цифра слева от косой черты), и поиском в блогах (цифра справа от косой черты)[5].

Таблица 1. Количество вхождений для некоторых существительных pluralia tantum в контексте слов *двое* и *две пары* (по данным НКРЯ и по результатам поиска в блогах)[6]

двое суток	2080/4900	*две пары суток	0/4
двое ворот	29/1937	две пары ворот	0/52
двое саней	26/219	две пары саней	1/5
двое часов	15/69	две пары часов	4/439
двое носилок	6/138	две пары носилок	1/5
двое очков	6/1038	две пары очков	11/49
двое брюк	6 (из них 3 в 19 в.)/998	две пары брюк	9/1189
двое трусов	3/1235	две пары трусов	3/843
двое весов	1 (19 век)/1097	две пары весов	0/6
двое ножниц	0/289	две пары ножниц	0/165
двое родов	2/824	*две пары родов	0/0
двое валенок	1/20	две пары валенок	4/202
двое перчаток	1/200	две пары перчаток	7/1261
двое сапог	0/322	две пары сапог	13/58
двое туфель	0/109	две пары туфель	3/38

На основании этих данных прежде всего можно сделать вывод о том, что слово *сутки* является особой точкой в системе[7]. Действительно, количество вхождений словосочетания *двое суток* в НКРЯ почти на два порядка превышает количество вхождений следующего по частотности словосочетания *двое ворот*. Очевидно, это связано с тем, что слово *сутки* представляет уникальную среди слов категории pluralia tantum комбинацию свойств непредметности и исчисляемости: сутки обычно именно считают. В отличие от других непредметных pluralia tantum, сочетание с собирательным числительным у слова *сутки* не имеет никаких стилистических ограничений. Так, хотя слово *каникулы* также обозначает временной отрезок, сочетание *двое каникул* воспринимается как разговорно-сниженное (и в НКРЯ не встречается). То же относится и к названиям событий или ритуальных действий (*роды*, *гонки*, *похороны,* и т. п.).

---

[5] Об использовании такого сегмента Интернета как блоги в качестве лингвистического ресурса см. [Беликов, Ахметова 2009]. Поиск производился при помощи системы Яндекс; данные здесь и далее приводятся на 1 апреля 2013 г.

[6] Результаты по НКРЯ получены путем сложения числа вхождений, найденных в основном и в газетном корпусе.

[7] Заметим также, что экзотическое собирательное числительное *по́лторо* используется, по-видимому, исключительно в составе словосочетания *по́лторо* суток.

В работе [Никунласси 2006] содержится статистический анализ употребления собирательных числительных с неодушевленными именами на материале, с одной стороны, опроса школьников, а с другой — баз данных Интегрум-Техно, значительно превышающих по объему НКРЯ, но содержащих исключительно тексты СМИ. Сопоставляя эти данные, А. Никунласси делает вывод, что в русском языке расширяется употребление нумеративных классификаторов (таких, как *пара, штука, нитка* <бус>), которые вытесняют собирательные числительные. В частности, парно-симметричные предметы, хотя и допускают сочетание с *двое*, более охотно сочетаются со словом *пара*, которое, по аналогии, захватывает и некоторые другие существительные pluralia tantum (как, например, *часы*). В то же время, неодушевленные существительные, которые обозначают собственно парные предметы (*сапоги, перчатки*) еще более явно предпочитают сочетание со словом *пара*.

Как показало наше исследование, материал НКРЯ в данном отношении оказывается недостаточным, а материал блогов свидетельствует о том, что в разговорной речи употребление слова *пара* в качестве классификатора для существительных pluralia tantum достаточно широко распространено; с другой стороны, как в самом узусе, как и в оценке нормативности употребления слов *пара* и *двое* для обозначения парных предметов или одного предмета сложной конструкции (*две пары сапог, две пары брюк, две пары часов* и т. д.) имеется большой разброс.

3.2. *ДВОЕ* как оператор, переводящий существительное в разряд личных имен

За пределами классов pluralia tantum и примыкающих к ним имен детенышей животных собирательные числительные могут квантифицировать только имена лиц. Более того, слово *двое* может использоваться именно как показатель личности существительного (т. е. как оператор, переводящий данное существительное в разряд личных имен) — например, в следующих контекстах:

1. При сочетании с существительными адъективного склонения: *два неизвестных* — это, скорее всего, числа, а *двое неизвестных* могут быть только люди; *два белых* — это грибы, а *двое белых* — люди, и т. д.

2. При сочетании с одушевленными существительными (кроме детенышей животных): *двое медведей* — это антропоморфизм, в отличие от *двое медвежат*. Ср.:

(1) *Место сбитого моментально без шума заняли* ***двое рыжих муравьёв***.
[Валерий Медведев. Баранкин, будь человеком! (1957)]

В примере (2) именно числительное *двое* указывает на то, что слово *горилла* использовано в переносном значении — ‘огромный, устрашающего вида мужчина’:

(2) *Я нажала кнопку первого этажа, и лифт начал плавно опускаться. Когда он остановился, я открыла дверцы и вывезла коляску на площадку вестибюля, где мне тут же преградили дорогу* ***двое горилл****, дежуривших на первом этаже.* [Марина Серова. Я больше не шучу] http://www.loveread.ec/read_book.php?id=3915&p=10

3. В контексте собирательного числительного неодушевленное существительное может метонимически обозначать человека:

(3) *— Значит, это трое. Один придет из гнойной с резекцией кишечника по поводу рака. Один из сосудистой после протезирования бедренной артерии, это пять. Ну и от вас* ***двое желудков****, — взглянул он на Архипова. — Это семь. И четверо старых клиентов. Итого одиннадцать.* [Влада Валеева. Скорая помощь (2002)]

3.3. *ДВОЕ* в контексте личных имен

В статье [Добрушина, Пантелеева 2008] показано, что существует определенная корреляция между частотностью употребления существительных во множественном числе и частотностью их сочетаемости с собирательными числительными. Это обстоятельство служит объяснением того факта (неоднократно обсуждавшегося в литературе ср. [Мельчук 1985], [Янко 2002]), что имена лиц «высокого ранга» употребляются с собирательными числительными реже, чем имена лиц «низкого ранга». А именно, для лиц «низкого ранга» более характерно коллективное поведение (*боевики, жители, пассажиры, солдаты*), с чем хорошо совместима семантика собирательного числительного. В целом, такое распределение действительно имеет место, однако существенно, что при желании говорящий может объединить в группу людей любого ранга, на основании какого-то общего признака, ср.:

(4) *На обеде* ***двое сенаторов-демократов****, 76-летний Эдвард Кеннеди (брат Джона Кеннеди) и 91-летний Роберт Бэрд, почувствовали себя плохо.*
[Полищук Оксана. Эх, раз, да еще раз // Труд-7, 2009.01.23]

(5) *В первой десятке еще* ***двое президентов*** *из 19 века: Джон Тайлер (1841–45), 8,51 млн долларов и Джеймс Монро (1817–25), 10, 27 млн долларов.* [Захар РАДОВ. Самым богатым президентом США оказался Джордж Вашингтон // Комсомольская правда, 2011.02.23]

(6) ***Двое великих биологов*** *заложили две ветви иммунологии: Пастер — химическую, а Мечников — клеточную* […].
[Юрий Чайковский. Юбилей Ламарка — Дарвина и революция в иммунологии // «Наука и жизнь», 2009]

(7) *Однако в период раннего Средневековья* ***двое пап****, Григорий I (590–604) и Николай I (858–867), особенно решительно выступали за укрепление и развитие папства.*
(http://slovarionline.ru/entsiklopediya_kolera/page/papstvo.5178)

Приведем некоторые (округленные) цифры, полученные поиском с помощью системы Яндекс для формы номинатива (на всем пространстве Интернета): *двое генералов* — 900; *двое заместителей* — 14 тыс., *двое депутатов* — 15 тыс., *двое членов* <чего-то> — 38 тыс. Интересно отметить, что для всех перечисленных существительных основной корпус НКРЯ дает ответ «ноль». Даже с учетом всех оговорок, касающихся цифр, выдаваемых поисковыми системами, указанные выше результаты, очевидно, не могут быть проигнорированы (ср. соображение, высказанные в [Шмелев 2010], о том, что отсутствие некоторого явления в корпусе, даже очень большом, не означает его отсутствия в языке).

4. *ДВОЕ* vs. *ОБА*

Как уже говорилось, русские числительные *двое* и *оба* весьма существенно различаются по своим семантическим и грамматическим свойствам. В данном разделе мы рассмотрим этот вопрос более подробно; начнем с сопоставления их сочетаемости (на фоне числительного *два*), см. Таблицу 2:

Таблица 2. Сочетаемость числительных *два*, *двое*, *оба* с разными классами существительных

	одуш. все падежи	**неодуш. не pl.tant. все падежи**	**pl.tant. «парные предметы» им.,вин.пад.**	**pl.tant. «парные предметы» косв. пад.**	**pl.tant. «мероприятия» им., вин.**	**pl.tant. «мероприятия» косв. пад.**
Два	+ два студента	+ два стола	–	+ режу *двумя ножницами*	–	+ участвовал *в двух гонках*
Оба	+ оба студента	+ оба стола	– (суппл.: *и те и другие ножницы*)	+ пользуюсь *обоими ножницами*	– (суппл.: *и те и другие гонки*)	+ участвовал *в обоих гонках*
Двое	+ двое студентов	–	+ двое ножниц	–	+ на воскресенье назначено *двое гонок*	–

Как можно видеть из этой таблицы, за исключением сочетаемости с одушевленными существительными, которую допускают все три слова, сочетаемостные свойства числительных *двое* и *оба* расходятся во всех выбранных позициях, и при этом свойства числительного *оба*, с точностью до возможного употребления супплетивных форм, совпадают со свойствами числительного *два*.

Обратимся теперь собственно к семантике.

4.1. *Двое*

Семантика слова *двое* двухфокусная. С одной стороны, *двое*, как и остальные собирательные числительные, имплицирует наличие некоторого множества лиц, принадлежащих к определенной категории (обозначенной данным существительным), из которого по некоторому параметру «выбираются» несколько элементов. Этот компонент составляет презумпцию слова *двое*. Тем самым, собирательные числительные имеют латентную валентность «из кого?»: *двое из них*[8]. По этой причине вне контекста естественно звучит *двое солдат, матросов, студентов* или *прихожан* и хуже — *двое президентов, капитанов, императоров* или *патриархов*: лица «высокого ранга» нормально мыслятся вне множества одноименных им лиц (ср. выше).

Второй фокус определяется тем, что эти выбранные из множества два элемента, в свою очередь, тоже оказываются между собой связанными — но уже не узами принадлежности к множеству одноименных лиц, а фактом данного совместного действия (или общностью некоторого признака), называемого в данном предложении.

4.2. *Оба*

Семантика слова *оба* также двухфокусная, но состоит из других компонентов. Словарь В. И. аля дает очень точное определение: *оба* — «тот и другой, каждый из двух; те двое, о коих идет речь» (Даль 1994: II, 1467). И. А. Мельчук описывает семантику слова *оба* так: «<…> *оба* обозначает не просто число, а включает в свой смысл еще и значение квантора ‘все’: *оба* = ‘все два’» [Мельчук 1985: 37][9]. Другими словами, слово *оба* содержит презумпцию существования закрытого класса, состоящего ровно из данных двух элементов, а также придает именной группе определенный референтный статус. Именно поэтому парные части тела, составляющие ядерную зону категории «парности», так естественно сочетаются со словом *оба*, ср. *ослеп на оба глаза*. В *оба* не только нет никакого открытого класса «таких же» объектов, более того, есть противоположный смысловой компонент ‘других таких же нет’; тем самым данные два объекта (или человека) составляют «исчерпывающий класс», формируемый

---

[8] В [Успенский 2004: 40] отмечается данное свойство собирательных числительных в связи с обсуждением категории «часть — целое».

[9] Ср. выражение *оба два* (которое в русском языке относится к просторечию), а также укр. *обидва*, белорусск. *абодва*, итал. *ambidue* и др. *Оба два* имеет ту же внутреннюю структуру, что *все три* (*все четыре и т. д.*): первое безударное местоимение-числительное выражает идею «исчерпанности» множества, а второе ударное количественное числительное указывает на его мощность.

некоторым признаком. Признак этот может быть «естественным» (ср. *требуется согласие обоих родителей*; *обе его бабушки из дворян*), он может быть упомянут в предтексте (*У Марии два сына. **Оба** они учатся в университете*), но может и высказываться, в форме презумпции, впервые: ***Оба** ее сына учатся в университете* (презумпция: ‘у нее есть только два сына’).

Именно эта презумпция существования двухэлементного класса блокирует взаимозамену между *двое* и *оба*. Предложение ***Двое** ее сыновей учатся в университете* не содержит никакой информации относительно наличия у нее еще других сыновей (тем самым часть смысла теряется). И наоборот, если вместо *В этот момент в комнату вошли **двое** полицейских* (где речь идет о новых персонажах) сказать *В этот момент в комнату вошли **оба** полицейских*, появится смысловой компонент, в исходной фразе отсутствовавший.

Таким образом, слова *двое* и *оба* имеют в некотором смысле противоположные презумпции. При этом ассерция и импликация у них кажутся сходными: ассертивным является компонент количества, а импликацию составляет идея некоторой близости, возникающей между упомянутыми двумя лицами в силу обладания одним и тем же признаком, предицируемым в данном предложении (‘учатся в университете’, ‘вошли в комнату’ и т. п.). Здесь, однако, имеется еще одно весьма существенное различие.

А именно, если слово *двое* обычно указывает на то, что два человека осуществляют **одно совместное действие**[10] (и в силу этого между ними возникает определенная связь), то *оба*, наоборот, указывает на **идентичность двух разных** действий. Так, например, фраза *Двое супругов имеют в собственности квартиру* значит «вместе», у них одна квартира на двоих, а *Оба супруга имеют в собственности квартиру* — «раздельно», у каждого по квартире[11]. *Обе девочки засмеялись*, *Оба сына поступили в университет* — это не одно совместное действие (в отличие от *Вошли двое полицейских* или *Двое неизвестных напали на бизнесмена*), а два идентичных, и это принципиальная разница. Соответственно, характер «близости», возникающей между референтами существительных, квантифицируемых словами *оба* и *двое*, различный: для *двое* это участие в совместном действии, а для оба — это сходство (поступков, судеб, характеров и т. п.). Тем самым, тождественным у *двое* и *оба* является лишь тривиальный собственно количественный компонент; все остальные составляющие значения различны.

Подводя итог проведенному сопоставлению, можно сказать, что слово *двое* (в конструкции с существительным) используется в ситуации, когда два «раздельных» человека, связанных между собой лишь принадлежностью к некоторой большей общности (солдаты, депутаты, матросы, студенты, ученики одного класса или члены любого множества, произвольно выделенного говорящим, как,

---

[10] О категории «совместного действия» см. [Зализняк, Шмелев 1999, 2013 (в печати)]; с предикатами, относящимися к данной категории, не сочетаются, в частности, слова *вместе* и *оба*.

[11] Ср. значение ‘и тот, и другой’ у др.-русского прилагательного *обои* (сохранившегося в выражении *обоего* пола) [Срезневский 1989: II, 532].

например, все президенты США), в данной ситуации оказываются объединены неким общим признаком, т. е. в некотором смысле действуют «вместе». Между тем *оба*, наоборот, употребляется в ситуации, когда два человека, жестко связанные между отношением «вместе» (так как они составляют уникальное двухэлементное множество), действуют в описываемой ситуации строго «раздельно».

Авторы благодарны анонимным рецензентам за доброжелательную и конструктивную критику.

Литература

1. *Беликов В. И., Ахметова М. В.* Статистическая оценка функциональных свойств лексики по материалам интернета // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2009. Бекасово, 27–31 мая 2009. М., 2009. С. 25–30.
2. *Виноградов В. В. Русский язык. М., 2001.*
3. *Грамматика-80* — Русская грамматика. Т. 1–2. Н. Ю. Шведова (гл. ред.) М., 1980.
4. *Даль В. И.* Толковый словарь живого великорусского языка. В 4-х тт. М., 1994.
5. *Добрушина, Н. Р., Пантелеева, С. А.* Собирательные числительные: коллектив как индивидуализация множественности. // А. Мустайоки и др. (ред.) Инструментарий русистики: корпусные подходы (Slavica Helsingiensia, 34). Хельсинки, 2008. С. 107–124.
6. *Зализняк Анна А., Шмелев А. Д.* О том, чего нельзя сделать вместе // Типология и история языка. К 60-летию А.Е.Кибрика. М., 1999. С. 450–457.
7. *Зализняк Анна А., Шмелев А. Д.* О двух лигвоспецифичных единицах русского числового кода. // Логический анализ языка. Числовой код: универсальное и лингвоспецифичное. М., УРСС, 2013 (в печати).
8. *Мельчук И. А. Поверхностный синтаксис русских числовых выражений. Wiener Slawistischer Almanach. Sonderband 16. Vienne, 1985.*
9. *Наконечная-Лаланн В.* К вопросу о функционировании собирательных числительных в современном русском языке // Russian Linguistics, 2013, Vo 37, 1, (in press). Электронная версия http://link.springer.com/article/10.1007/s11185-012-9103-5
10. *Никунласси А.* К вопросу об употреблении собирательных числительных при неодушевленных существительных в русском языке (норма и узус) // Scando-Slavica, 2006, vol. 52. З. 5–18.
11. *Сичинава Д. В.* Числительное // http://rusgram.ru. 2012.
12. *Срезневский И. И.* Материалы к словарю древнерусского языка. В 3-х тт. М., 1994.
13. *Успенский Б. А.* Часть и целое в русской грамматике. Москва: Языки славянской культуры, 2004.
14. *Фасмер М.* Этимологический словарь русского языка. В 4-х тт. М., 1996.
15. *Шмелев А. Д. Человек, люди, народ* в числовых конструкциях. // Von grammatischen

16. *Kategorien und sprachlichen Weltbildern* — Die Slavia von der Sprachgeschichte bis zum Politsprache. Festschrift Daniel Weiss zum 60. Geburtstag. Wiener Slavistischer Almanach, Sbd 73, München-Wien, 2009. P. 569–586.
17. *Шмелёв А. Д.* Языковые факты и корпусные данные // Русский язык в научном освещении, 2010, 19 (1), 236–265.
18. *Янко Т. Е.* Русские числительные как классификаторы существительных // Русский язык в научном освещении. 2002. №1(3), с. 168–181.
19. *Nikunlassi A.* The use of collective numerals in contemporary Russian: an empirical approach // Wiener Slawistischer Almanach 45 (2000), 209–246.
20. *Wierzbicka A.* Semantics: Primes and Universals. Oxford: Oxford Univ. Press, 1996.

References

1. *Belikov V. I., Ahmetova M. V.* (2009), Statisticheskaja ocenka funkcional'nyh svojstv leksiki po materialam Interneta [Statistical evaluation of functional properties of lexical units based on the Internet data], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" [Komp'iuternaja Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferencii "Dialog 2009"], Bekasovo, pp.25–30
2. *Vinogradov V. V.* (2001), Russkij jazyk [The Russian language], Moscow.
3. *Russkaja grammatika* (1980) [The Russian grammar]. T.1–2. N.Ju, Shvedova (red.) M., 1980.
4. *Dal' V. I.* (1994) Tolkovyj slovar' russkogo jazyka [Explanatory dictionary of Russian]. T. 1–4. M., 1994.
5. *Dobrushina E. R., Panteleeva S. A.* (2008) Sobiratel'nye chislitel'nye: kollektiv kak individualizacija mnozhestvennosti [Collective numerals: community as individualization of multiplicity], Mustajoki A. et al. (eds.) Instrumentarij rusistiki: korpusnye podhody [Toolset of Russian studies: corpus-based approaches], Slavica Helsingiensia, 34, Helsinki. P. 107–124.
6. *Zalizniak Anna A., Shmelev A. D.* (1999) O tom, chego nel'zja sdelat' vmeste [What cannot be done together?], Tipologija i istorija jazyka. Sbornik statej k 60-letiju A. E. Kibrika [Typology and the history oa language. Festschrift to A. E. Kibrik], M., P. 450–457.
7. *Zalizniak Anna A., Shmelev A. D.* (2013, in print). O dvuh lingvospecifichnyh edinicah russkogo chislovogo koda [On two language-spicific units of the Russian numeral code], Logicheskij analiz jazyka. Chislovoj kod: universal'noe I lingvospecifichnoe [Logical analysis of language: universal and language-specific], Moskva.
8. *Ljashevskaja O. N., Sharov S. A.* (2009) Chastotnyj slovar' sovremennogo russkogo jazyka (na materialax Nacional'nogo korpusa russkogo jazyka) [Frequency dictionary of Russian (based on the Russian National Corpus)]. Moskva. http://dict.ruslang.ru/
9. *Mel'chuk I. A.* (1985) Poverhnostnyj sintaksis russkih chislovyh varazhenij [Surface syntax of Russian numeral expressions]. Wiener Slawistischer Almanach. Sonderband 16. Vienne.

10. *Nakonechnaja-Lalann V.* (2013) K voprosu o funkcionirovanii sobiratel'nyh chislitel'nyh v sovremennom russkom jazyke [On the functioning of collective numerals in Russian], Russian Linguistics, 2013, vol. 37, №1, P. 91–101. http://link.springer.com/article/10.1007/s11185-012-9103-5
11. *Nikunlassi A.* K voprosu ob upotreblenii sobiratel'nyh chislitel'nyh pri neodushevlennyh sushchestvitel'nyh v russkom jazyke [On the use of collective numerals with inanimate nouns in Russian], Scando-Slavica, 2006, vol. 52. P. 5–18.
12. *Sichinava D. V.* (2012) Chislitel'noe [Numeral], http://rusgram.ru.
13. *Sreznevskij I. I.* (1994) Materialy k slovarju drevnerusskogo jazyka. T. 1–3. Moskva.
14. *Uspenskij B. A.* (2004) Chast' i celoe v russkoj grammatike [Part and whole in the Russian grammar], Москва: Jazyki russkoj kul'tury.
15. *Fasmer M.* Etimologicheskij slovar' russkogo jazyka [Etymological dictionary of Russian]. T. 14.
16. *Shmelev A. D.* (2009) *Chelovek. ljudi, narod* v chislovyh konstrukcijah [*Chelovek. ljudi, narod* in numeral constructions] Von grammatischen Kategorien und sprachlichen Weltbildern — Die Slavia von der Sprachgeschichte bis zum Politsprache. Festschrift Daniel Weiss zum 60. Geburtstag. Wiener Slavistischer Almanach, Sbd 73, München-Wien, 2009. P. 569–586.
17. *Shmelev A. D.* (2010) Jazykovye fakty i korpusnye dannye [Linguistic facts and corpus data], Russkij jazyk v nauchnom osveshchenii [Russian language in the scientific coverage], 19 (1), pp. 236–265.
18. *Yanko T. E.* (2002) Russkie chislitel'nye kak klassifikatory sushchestvitel'nyh [Russian numerals as nominal classifiers], Russkij jazyk v nauchnom osveshchenii [Russian language in the scientific coverage], №1(3), pp. 168–181.
19. *Nikunlassi A.* (2000) The use of collective numerals in contemporary Russian: an empirical approach // Wiener Slawistischer Almanach 45, pp. 209–246.
20. *Wierzbicka A.* (1996) Semantics: Primes and Universals. Oxford: Oxford Univ. Press.

ДА ЧЕРТ ЛИ В ДЕТАЛЯХ?.. МЕРА ДЛЯ ОЦЕНКИ СОВПАДЕНИЯ ЭЛЕМЕНТОВ ИДИОСТИЛЯ В ТЕКСТАХ ОДНОГО — ИЛИ ДВУХ РАЗНЫХ? АВТОРОВ (Агеев — Сирин/Набоков — Леви)[1]

Михеев М. Ю. (m-miheev@rambler.ru)

НИВЦ МГУ, Москва, Россия

Рассматривается гипотеза Н. Струве о том, что автором текста *«Роман с кокаином»* (опубликованного в 1936 под псевдонимом М. Агеев) был В. Набоков. Сравниваются некоторые идеостилевые черты этого произведения и всех текстов Набокова, а также то, что имеется в Национальном корпусе русского языка (опубликованное как до выхода произведений Агеева и Набокова, так и после). Общий итог — в пользу того, что Набоков к этому тексту, скорее всего, все-таки непричастен. Никита Струве отверг биографические аргументы и потребовал, чтобы ему были предъявлены именно аргументы «филологические», — литературоведческого, или поэтического, характера. Таковые здесь и рассматриваются.

Ключевые слова: поэтические приемы, идиостиль писателя, уникальные словосочетания, текстовые конструкции, архаизмы, авторские неологизмы, установление авторства

THE DEVIL IN THE DETAILS THERE?.. MESURE TO ASSESS THE MATCH IDIOSTYLE ELEMENTS IN THE TEXTS OF THE SAME — OR TWO DIFFERENT? AUTHORS (Ageev — Sirin/Nabokov — Levi)

Mikheev M. Yu. (m-miheev@rambler.ru)

NIVC MGU, Moscow, Russia

[1] Работа выполнена при финансовой поддержке РФФИ, грант № 13-06-00402.

We analyze N. Struve's hypothesis that the author of the text of *The romance with cocaine* (published in 1936 under the pseudonym M. Ageev) was Vladimir Nabokov. We compare the idiostyle features of this text and all of Nabokov's texts, as well as what is available in the Russian National Corpus, published before the Ageev and Nabokov works and after them. The general conclusion is that Nabokov seems not to be involved in this text. This problem was stated by Nikita Struve, who rejected biographical arguments and required that "philological", literary or poetic arguments should be given. We consider all of these arguments.

Keywords: poetic techniques, idiostyle, unique phrases, textual constructions, archaisms, neologisms, attribution

Вся набоковская поэтика зиждется
на тро́пе уточнения… (Ив. Толстой)

Попытаюсь решить задачу, поставленную Никитой Струве, который отверг аргументы Габриэля Суперфина и Марины Сорокиной, приведенные ими в пользу того, что автором повести «Роман с кокаином», скрытым за псевдонимом М. Агеев[2], никак не мог быть Владимир Набоков, а был, на самом деле, некий практически ничего более не написавший — Марк Людвигович (или Лазерович) Леви (или Левит)[3]. Как известно, Струве потребовал от своих оппонентов, чтобы ему были предъявлены не биографические и не «краеведческие» факты — то, что, например, Набоков прекрасно знал Петербург, где прожил до своего отъезда в Берлин, но не Москву, которая в подробностях описана в повести; или что, например, по сохранившимся в архивах документам многие из лиц, чьи имена даны в тексте и которые в самом деле обучались или преподавали в конкретное время в соответствующей гимназии как соученики автора или его преподаватели[4], но — именно «филологические» аргументы, литературоведческого, поэтического характера. На этот «вызов» Суперфин и Сорокина не ответили. Попытался ответить вместо них — Иван Толстой. Многие его аргументы действительно основательны[5], но не со всем и у него можно

---

[2] С написанием первого Е как Ять.

[3] *Сын газенпотского 1-й гильдии купца* (Суперфин, Сорокина 1994: 263).

[4] Суперфин считает, что сюжет «Романа…» носит явно автобиографический характер и что *художественный текст этого автора более документален, чем любые исходящие от него официальные материалы* (там же, с. 268–9).

[5] "Отметим прежде всего главный порок «Романа с кокаином» — книга Агеева в каждом своем слове серьезна, исповедальна, ибо лишена чувства юмора. (…) [это] — плоский автор с выраженной склонностью к многословию и общим местам" (Толстой 1995: 198).

согласиться[6]. Ниже я рассмотрю некоторые уже обсуждавшиеся критиками слова и выражения двух авторов и хотел бы привести несколько новых, усиливающих ту или другую позицию.

Перед нами, с одной стороны, литератор-любитель Агеев-Леви, опубликовавший всего лишь только одну, сделавшую его знаменитым (хотя и посмертно) повесть «Роман с кокаином» (далее — РсК), а с другой стороны, Сирин-Набоков, всемирно известный автор множества рассказов, повестей, пьес, романов. Объем текста РсК, подвергаемый проверке, — 45 200 словоупотреблений, а текстов Набокова, которые мы будем с ним сравнивать, — 1 042 200 (по крайней мере, у меня, на самом деле их еще больше), что превышает объем РсК более чем в 20 раз.

Ситуация в целом кажется в чем-то сходной с той, с которой мы сталкиваемся в случае рассмотрения романа Шолохова «Тихий Дон» и так называемой *отслойки* от него — *прототекста*, возможно лежавшего некогда в основании (во всяком случае, как его предтекст) и принадлежавшего, по предположениям наиболее многочисленных «антишолоховедов», т. е. сторонников версии «плагиата», донскому писателю Федору Крюкову, умершему в начале 1920 г. Однако дилемма — Набоков или Леви? — все-таки более простой случай по сравнению с «Тихим Доном», так как в нем никакой «отслойки» внутри текста производить не надо — т. е. решать, какие из выражений принадлежат одному, а какие другому автору. Достаточно, кажется, только проверить поэтические средства, которые использовались в повести «Агеевым» и оценить, насколько они совпадут со средствами, которыми пользовался в своих произведениях Набоков — по крайней мере, до 1934 года, т. е. до начала публикации повести (точнее, до 1933-го, когда ее рукопись была послана Агеевым-Леви — из Константинополя в Париж[7]).

Замечу, что по интуитивным ощущениям Струве (да и по моим собственным, пока я не стал разбирать примеры из текста подробно), многие из метафор или уникальных игровых «словечек» Набокова совпадают с метафорами в РсК. Это заслуживает проверки. Ведь, вообще говоря, мистификации Набокову всегда были свойственны. Насыщенность совпадений в тексте РсК с тем или

---

[6] Например, слово *шибко* очень часто встречающееся у Агеева в РсК, 24 раза, но оно, по мнению Толстого, "употребляется при этом именно в «зощенковском» значении, как 'очень, сильно', а не, как у Набокова — 'быстро' (там же, с.203). Но у Даля все-таки было такое значение *шибкий*, как 'сильный, резкий'; в современном же языке оба слова с этим корнем стали просторечными — как наречие *шибко* 'сильно, очень', так и прилагательное *шибкий* 'скорый, быстрый' (МАС). Хотя возможно подмеченное различие и характерно для идиолектов Агеева и Набокова.

[7] Рукопись под названием «Повесть с кокаином» была прислана по почте из Константинополя в феврале 1933 г. в парижский журнал «Иллюстрированная Россия» для участия в конкурсе произведений новых авторов на имя Адамовича. Но ее тогда лишили права публикации, так как произведения подписывались псевдонимом или девизом, о чем автор не позаботился (Рагозина 2000). Отрывки РсК как «Повесть с кокаином» [по-видимому, с авторским названием] были первоначально опубликованы в парижской «Иллюстрированной Жизни» (март 1933), а затем (1 часть) — в журнале «Числа», ну, а в 1936-м она вышла отдельным изданием уже как «Роман с кокаином» (Парфенов 1995: 5).

иным «коронным» выражением Набокова действительно впечатляет (ее еще предстоит досконально оценить будущему исследователю, у меня здесь для этого не хватит ни времени, ни места)[8]. Создается впечатление, будто Агеев специально копирует стиль, подлаживаясь под Набокова. — Так что это, намеренное стилизаторство? Естественно, что сравнивать надо будет только такие элементы текста, которые наиболее характерны для обоих авторов и в то же время наименее общеупотребительны у остальных пишущих на русском языке, а в идеале — вообще никем, кроме них, не употреблялись. Иначе говоря, то, что в окружающем контексте наименее представлено, а у данных авторов совпадает, составляя как бы единый идиолект. Рассмотрим пристальнее эти наиболее похожие метафоры или просто экзотизмы — слова того «вывозного сорта» эмигрантского русского языка, на котором предпочитал писать Набоков и которые как будто наследовал за ним Агеев[9]. — Будем сравнивать их с тем, что имеется в Национальном корпусе русского языка (далее — НК): в нем на сегодняшний момент (фев. 2013) около 229 968 800 слов, что более чем в 200 раз превышает объем рассматриваемых текстов обоих писателей.

0. Никита Струве (Струве 1990) подметил среди множества прочих, например, такие, как он их называет, «разительные совпадения»: у Агеева в РсК кто-то из героев *по-рыбьи опускал углы губ*, и у Набокова такое выражение уже было — *по-рыбьи открытые рты* — в «Подвиге» (1931–32). Да и позже: *к старческим рыбьим губам* («Приглашение на казнь» 1935–36), всего же — не менее 5 раз: то есть перед нами устойчивая черта его стиля. Но есть пример более раннего использования этого сравнения — причем, в тексте вполне доступном для обоих (Набокова и Агеева как читателей), его находим по НК: *Иволгин так же молча и машинально ловил его за локоть и, как будто молча, бормотал что-то судорожно, по-рыбьи открывая и закрывая рот* (Арцыбашев «Ужас» 1905).

00. Другое из «разительных» совпадений: *лицо у него было, от морского солнца, как ростбиф* («Подвиг» 1931–32); *с… пухлыми, цвета ветчины, губами* (М. Агеев «Паршивый народ» 1934[10]); *Бэрнес был крупного сложения, светлоглазый шотландец с прямыми желтыми волосами и с лицом цвета сырой ветчины* («Другие берега» 1954) — т. е. у Набокова опять же встречается совпадающее выражение как до, так и после текста Агеева (хотя первое — совпадающее не в точности). Но и у этого выражения находится более точный предшественник: *Это были вполне честные немецкие лица: одинаковые губы цвета ветчины,*

---

[8] Струве даже считает, что некоторые метафоры поздних произведений Набокова повторяют метафоры из РсК, которые ранее у Сирина не встречались. Наиболее рискованное его заявление звучит так: что РсК не только стоит на уровне бунинско-набоковского мастерства (так считал Мережковский), но что РсК уже «содержит в себе **все** темы и **все** приемы набоковского мастерства» (Струве 1995: 171).

[9] Или же: написал все-таки сам Набоков, а выдал за написанное — Леви[том]? Или же: первоначально написал Леви[т], а отредактировал — Набоков? и т. д.

[10] Этот рассказ — единственный текст, опубликованный Агеевым, помимо РсК.

глаза вялые, как трава, — медлительные, уверенные в каждом своем движении люди (Г. В. Алексеев «Подземная Москва» 1924).

С другой стороны, упреки, высказанные Зинаидой Шаховской (а еще вдовой Набокова, Верой Набоковой[11], и также ранее — В. Ходасевичем): что де у Агеева в тексте встречаются совершенно недопустимые для мэтра-Набокова погрешности языка, вплоть даже до грамматических ошибок вроде *звенение рубля; натуживать волю*, — Струве отважно отвергает, объявляя мнимыми, и говорит, что эти погрешности как раз скорее сближают автора РсК с его знаменитым двойником, поскольку Набоков вовсе не классик, а прежде всего — модернист, экспериментатор, и определенный налет безвкусицы, фокусничанья в любом его тексте почти всегда был налицо (Струве 1986: 161–3). Вот это мне кажется вполне заслуживающим внимания (как и сама гипотеза, что под видом чьего-то чужого текста Набоков мог кинуть такой пробный шар, как РсК, где эти «погрешности» просто достигли некоторого максимума). Свое предположение Струве подытоживает так: «Не привез ли Марко Леви с собой в Константинополь рукопись, данную ему Набоковым в Берлине, чтобы оттуда отослать ее в Париж?» (Струве 1986: 175). Иными словами, Набоков мог по каким-то соображениям вступить с Агеевым в сговор, чтобы тот издал рукопись под своим именем. Зачем это было нужно, не будем «вдаваться», оставив выяснение подробностей для любителей. Но в принципе, такой оборот дела кажется допустимым: Набоков захотел опубликовать свою очередную экспериментальную прозу под чужим именем и до самой смерти никому (даже жене) не признался, что авторство в ней принадлежало все-таки ему.

Рассмотрим собственно текстовые, «поэтические» аргументы — как в пользу этой версии, так и в ее опровержение, на тех примерах, которые наиболее сближают два идиостиля. Учтем, что если в случае сравнения Крюкова

---

[11] В своем письме в редакцию газеты «Русская мысль» Вера Набокова категорически утверждала: «Мой муж, писатель Владимир Набоков, «Романа с кокаином» не писал, псевдонимом «М. Агеев» никогда не пользовался, в журнале «Числа», нагрубившем ему в одном из своих первых номеров, не печатался, в Москве никогда не был, в жизни своей не касался кокаина (ни каких-либо других наркотиков) и писал, в отличие от Агеева, на великолепном, чистом и правильном, петербургском русском языке. О слабости русского языка Агеева можно судить не только по слову «шибко» — слову, совершенно недопустимому в серьезном литературном произведении, — но и по таким словам, как «зачихал» в смысле ‘чихнул’ или «использовывать» и тому подобное. Я нарочно не вхожу в рассмотрение примитивности замысла и грубости его выполнения далеко, впрочем, не бездарным господином Агеевым, но не могу не удивляться тому, что Н. Струве, сорбоннский специалист по русскому языку и литературе, мог спутать вульгарный и часто неправильный слог Агеева со слогом тончайшего стилиста В. Набокова» (Волчек 1989). Однако на обложке журнала «Числа» № 10, 1934 красовалась рекламное объявление об участии Набокова в журнале *(В вышедших номерах напечатали оригинальные произведения и ответы на анкету следующие авторы: Г. Адамович, М. Агеев,…, В. Сирин)*. По поводу кокаина — Струве указывает на рассказ Набокова «Случайность» (1924), где фамилия героя — *Лужин*, которая потом перекочует в роман «Защита Лужина» (1929): “Не растроúлся ли «забытый» рассказ: кокаин перешел к студенту Вадиму Масленникову [главный герой РсК], самоубийство к шахматисту Лужину, невстреча с приезжей из СССР бывшей возлюбленной — к Ганину из «Машеньки»” (Струве 1986: 159).

с Шолоховым мы имели примерно одинаковые по объему массивы текстов, то тексты Набокова значительно превышают текст РсК. И если там стоял вопрос, есть ли в составе «Тихого Дона» фрагменты идиостиля Крюкова, «разбавленные» или растворенные собственноручной правкой текста Шолохова, то здесь следует сформулировать проблему иначе: является ли РсК стилизацией «под» Набокова, выполненной самостоятельным литератором Агеевым-Леви, или же все-таки следует опознать в нем текст самого Набокова? — Ответ почти предопределен, так как существует множество признанных способов проверки по совершенно, казалось бы, независимым критериям (употребление автором предлогов, союзов, частиц, средняя длина слова, предложения, соотношение в нем придаточных итп.), доказывающих, например, что как структура, так и количество вводных оборотов и вводных конструкций (*может быть, впрочем, конечно, вероятно…*) в тексте РсК и в шести русскоязычных романах Набокова, написанных приблизительно в тот же период, сильно не совпадают (Мухин 2001)[12]. Есть и иные статистические методы идентификации авторства. Но в случае серьезного редактирования текста, его «обработки», или переписывания своими словами, все они очевидно работать уже не будут, поскольку текст станет отвечать скорее редакторским, нежели авторским нормам словоупотребления.

Устаревшие, архаизированные глаголы (или их значения)

1. *Натуживать* у Агеева — волю, у Набокова — голос (а у других, по НК: мускулы, лицо…) и еще *натуживаться / натуженный / натуженность*. В РсК — при описании наркотического опьянения: 1. *хотя я натуживаю всю силу воли и приказываю рукам двигаться быстро, руки не слушаются* (и совсем рядом в тексте следующие два употребления этого слова). Вероятно, это заимствовано Агеевым — из прочитанных сравнительно недавно романа и повести Сирина, где встретились похожие, с нашей современной точки зрения, архаизмы (или неологизмы). Наиболее яркий пример — в раннем рассказе: *такая грузная боль давила на грудь, такая боль — и ничего не видать, кроме зыбкой лампы, — и в сердце упираются ребра, мешают вздохнуть,– и кто-то, перегнув ему ногу, ломает ее, натужился, сейчас хряснет* («Катастрофа» 1924). Слово повторяется у Набокова 7 раз в разных его текстах, причем некоторые примеры с точным повтором образа, самоповтором, попаданием «один-в-один», а у Агеева мы их не наблюдаем: *Успешно, хоть и медлительно, росъ на балконе круглый, натуженный, седовласый кактусъ* («Отчаяние» 1934) — и через 20 лет: *однажды, в пустыне, где-то в Новой Мексике, среди высоких юкк в лилейном цвету и натуженных кактусов, за мною шла в продолжение двух-трех миль огромная вороная кобыла* («Другие берега»

[12] Сравнение по романам «Король, дама, валет» 1928, «Защита Лужина» 1930–1, «Подвиг» 1932, «Камера обскура» 1932–3, «Отчаяние» 1936 и «Дар» 1937. Это сходно с наличием «статистического» подтверждения авторства Шолохова скандинавскими исследователями под руководством Г.Хьетсо, которое, правда, позднее оспаривалось.

1954)[13]. Кроме того, Набоковские возвратные формы глагола или причастия звучат все же более нормативно, не так резко, как Агеевское *натуживать*.

При этом конечно возможно, что оба, и Набоков и Агеев, черпали, так сказать, просто из языка, порождая вполне независимо каждый свои мелкие отступления от нормы. Ведь и в НК, еще ранее 1924 года: *лишь слабое сравнение с картонной цирковой гирей, ухватясь за которую профан заранее натуживает мускулы, но, вмиг брошенный собственным усилием навзничь, еще не в состоянии понять, что случилось, — может быть уподоблено впечатлению, с каким отступили и разбежались все, едва Крукс поднялся вверх* (А. Грин «Блистающий мир» 1923)[14].

000. Но вот зато таких оставленных улик, явно и исключительно Набоковских словечек, например, как глагол *отпахнуть* (употребляемый вместо *раскрыть* или *распахнуть* — для обозначения открывания двери), у Агеева мы почему-то не находим. Или, скажем, *переглотнуть* (в значении ‘с трудом/судорожно сглотнуть один раз’), который представляет собой совершенно уникальное выражение, своеобразный фирменный знак, или «тавро» Набокова и повторен им 12 раз в 10 текстах! Последний не встречается более ни у кого по НК, и у Агеева его нет. С другой стороны, имеются исключительно Агеевские гапаксы: *зачихнул, опыхивать*, *отплеснуть* (ну, а также используемые уже другими вместе с ним, но при этом — не Набоковым: *запух, выпыхивать, неизведомый, обглаживать*), уникально Набоковских же «следов» в РсК все-таки не находим.

2. Устаревший к середине или даже к началу XX века глагол *поворотить* (глаза / на) / *поворотиться к* (кому или чему-либо) вместо *повернуть/-ся*[15] — употреблен дважды в РсК и множество раз у Набокова, но почему-то уже только после РсК: в текстах, написанных автором на английском, т. е. переводных! Вот у Агеева: 1. *медленно-медленно поворотил изумленно выпученные глаза прямо на Семенова*. И второй раз там же: 2. *Помилуйте, Софья Петровна, — поворотился к ней вместе со столиком Яг*. — В текстах Набокова почему-то этот «вновь архаизированный» глагол возникает только после его

---

[13] И еще дважды: *Между тем рыба начинала клевать, – и, пренебрегая удочкой, попросту держа в пальцах лесу, натуженную, вздрагивающую, Василий чуть-чуть подергивал, испытывая прочность подводных судорог, и вдруг вытаскивал пескаря или плотву…* («Круг» 1934); *М-сье Пьер поднимал крепко закушенный стул, вздрагивали натуженные мускулы, да скрипела челюсть* («Приглашение…» 1935–6).

[14] С другой стороны, если посмотреть на частоту именно этой группы слов, т. е. *натужить/натуживать-ся/натуженный*, в целом по НК (а там встречаем 22 употребления — т. е. на весь объем НК, до сегодняшнего дня, 229 968 800 слов, их средняя частота 0,001 промилле), то у Набокова она в 7 раз выше (0,007), а у Агеева — даже в 66 раз (0,066). На мой взгляд, это аргумент в пользу сходства, или даже «единства» двух идиостилей, именно в этом компоненте.

[15] Во всем НК последний глагол встречается почти в 30 раз чаще, чем первый (поворотить — 1050 употреблений против 28 000 поворачивать) — у Набокова выдерживается такое же соотношение (всего 14 раз против 390, т. е. 1/28), однако если сравнивать по современным текстам, написанным после Набоковских (1966–2012), то это соотношение – еще в 3 раза реже (130 против 11.000, т. е. 1/85).

перехода на английский: 1. *Крестьянин поворотился, поглядел на пустое сиденье и сообщил…* («Под знаком незаконнорожденных», англ. 1947) и там же еще дважды: 2. *потом один поворотился и стал глядеть куда-то вбок ;* 3. *Кол поворотился к Кристалсену* (на рус. — в пер. С. Б. Ильина); 4. *спросила, поворотясь в направлении грохота: «Что вы там ищете, Тимофей?»* («Пнин» 1954–5; англ. изд. 1957; в рус. переводах — Б. Носика и С. Ильина[16]); 5. *докатился до Тихого Океана; поворотил на север сквозь бледный сиреневый пух калифорнийского мирта, цветущего по лесным обочинам* («Лолита» 1955; русский перевод самого Набокова, 1967); 6. *Не желая быть свидетелем супружеской сцены, я поворотился, чтобы уйти, но она остановила меня* («Бледное пламя» 1962, пер. С. Ильина и А. Глебовской) и там же еще 7 раз. — Но почему это словечко появляется у Набокова только в текстах, написанных после Агеева? Может быть это и надо признать влиянием Агеева? — Тут скорее всего «архаизация» произошла задним числом, была внесена переводчиками, что и сам Набоков принял под их влиянием.

3. Интересен у обоих редкий глагол *промахивать / промахнуть* — в значении (а) ‘пройти, проехать, пробежать, пролететь, пронестись, проскочить’ <мимо и на большой скорости, возможно даже не заметив — кого-то или чего-либо>. В МАСе *промахивать* в несов. виде образуется только от *промахать* — т. е. ‘пройти какое-либо большое расстояние’ (б), но не в указанном выше значении (а). В искомом же значении этот глагол фиксирует словарь говоров (СРНГ) — как сов. вид перех. и неперех. (со ссылкой на словарь Академии 1822 и более поздний пример: *Промахнул нас губернатор, поскакал.* Смол. 1914)[17]. У Агеева глагол встречается дважды, и первое употребление — как раз в интересующем нас устаревшем значении (а): 1. *Когда промахнули Яр и стала видна вышка трамвайной станции…* А второе довольно странно, не укладываясь ни в одно даже из набора диалектных: 2. *когда, наконец, случайно промахнув по карману, я звякнул в нем ее неиспользованными десятью серебряными пятачками, и тут же вспомнил ее губки…* — т. е. очевидно ‘задев, хлопнув, проведя вдоль’ (брюк с карманом)? У Набокова этот глагол — как в сов., так и в несов. виде — устойчивый маркер стиля, встречающийся более 15 раз (что более чем в 50 раз превосходит средний уровень по НК), с явным культивированием архаизированного значения (а). Интересно, что глаголов предписываемого словарем сов. вида в этом значении становится у него со временем все меньше[18]. Вот «правильные» формы: 1. *Промахнуло восемь столетий: саранчой налетели татары* (Набоков «Кэмбридж» 1921); 2. *Молодой белый пекарь в колпаке промахнул на трехколесном велосипеде: есть что-то ангельское в человеке, осыпанном мукой* («Путеводитель…» 1925); 3. *Макс, не понимая, видел издали,*

---

[16] Но не в пер. Г. Барабтарло (1983)!

[17] Там же фиксируются не интересные нам значения ‘промахнуться’ Арх., Хабар.; ‘быстро продать ч-л, променять’ Перм. и ‘упускать, пропускать’ ч-л. – Ряз. Хабар.

[18] Напомню, что в МАСе *промахивать* допускается только как несов. к *промахать*.

как промахнули сплошной полосой освещенные окна («Случайность» 1924); 4. *Огромное небо, налитое розоватой мутью, темнело, мигали огни, промахнул трамвай и разрыдался райским блеском в асфальте* («Сказка» 1926); 5. *На берегу где-то заиграли зорю, промахнули над пароходом две чайки, черные как вороны, и с плеском легкого дождя, сетью мгновенных колец прыгнула стая рыб* («Машенька» 1926); 6. *Промахнула мелкая станция, платформа под черным навесом, и снова стало темно* («Король…» 1928).

А вот четыре примера с явным нарушением узуса: 7. *А знаешь ли, с каким великолепным грохотом промахивает через мост, над улицей, освещенный, хохочущий всеми окнами своими поезд*? («Письмо в Россию» 1925); 8. *Изредка, вскрикнув оленьим голосом, промахивал автомобиль* («Машенька» 1926) и там же: 9. *Сейчас она спит в вагоне, промахивают в темноте телеграфные столбы, сосны, взбегающие скаты…* — здесь очевидно ‘пролетают так быстро, что их на успеваешь рассмотреть’; 10. *За ослепительной пустыней площади, по которой изредка с криком, новым, столичным, промахивал автомобиль* («Король…» 1928). Вот и это «мо» Набокова Агеев почему-то тоже не подхватывает.

То есть соотношение «правильных» к «неправильным» формам было 6 : 4. А вот после РсК у Набокова складывается почему-то явное предпочтение именно в пользу отклонения от нормативного значения (ниже только последний из примеров — с глаголом в сов. виде) и соотношение меняется на 1 : 5. Вот примеры: 1. *Была черная ветреная ночь, каждыя несколько секундъ промахивалъ надъ крышами бледный лучъ радіобашни, — световой тикъ, тихое безуміе прожектора* («Отчаяние» 1934); 2. *Гремел телефон, промахивал, развеваясь, метранпаж, театральный рецензент всё читал в углу приблудную из Вильны газетку* («Дар» 1937–8); 3. *с новым, беспокойным любопытством я взглянул на мостовую, на белый ее покров, по которому тянулись черные линии, на бурое небо, по которому изредка промахивал странный свет* («Посещение музея» 1938); 4. *Она гремела по асфальту среди других, сильно наклоняясь вперед и в ритм качая опущенными руками, промахивала с уверенной быстротой, ловко поворачивалась, так что перехлест юбки обнажал ляжку* («Волшебник» 1939); 5. *видишь скучного начальника небольшой станции, стоящего в одиночестве на платформе, мимо которого промахивает твой поезд* («Другие берега» 1936–67); 6. *Автомобиль семейного типа выскочил из лиственной тени проспекта, продолжая тащить некоторую ее часть с собой, пока этот узор не распался у него на крыше, за край которой держался левой рукой, высунутой из окна, полуголый водитель машины; она промахнула идиотским аллюром* («Лолита» 1955–65).

В целом и в НК форма этого глагола несов. также встречаются, но всего лишь однажды, у Александра Малышкина: *Но тут же за путями проступали кирпично-красные тылы заводов, окраины, утопающие в индустриальной мгле, за березами промахивали голенастые железные конструкции электросети; чуялось невдалеке разноцветное и могучее возбуждение Москвы* («Люди из захолустья» 1938). — Но здесь, в свою очередь, логично подозревать заимствование у Набокова.

Необычные сочетания или измененные внутри сочетания значения слов

4. Глагол *жмуриться / зажмуриться* встречается в текстах Набокова 50 раз, а в РсК — 4 раза, где находим одно уникальное (три остальные употребления вполне нормативны) — когда герой неожиданно встречает свою возлюбленную: *Я переступил порог. И вдохнув сырой и душистый сумрак, — вдруг мысленно зажмурился от внутреннего и страшного удара: в магазине стояла Соня*.

Сочетания *мысленно зажмурившись* у Набокова нет. Пропуская все более или менее нормативные употребления глагола, укажу наиболее метафорические. Во-первых, в переходе от «зажмуренного» поцелуя к целиком уже «зажмуренной» — душе: 1. *Он давал себя укачивать, баловать, щекотать, принимал с зажмуренной душой ласковую жизнь, обволакивавшую его со всех сторон* («Защита…» 1929–30). Повтор того же в «Даре»: 2. Федор Константинович *старался сосредоточиться, представить себе недавнюю теплоту их живых отношений, но душа не желала шевелиться, а лежала, сонная и зажмуренная, довольная своей клеткой* («Дар» 1937–38)[19]. А в следующем примере оттуда же можно при желании увидеть влияние Агеева на Набокова: *лежа неподвижно и даже не жмурясь, я мысленно вижу, как моя мать, в шеншилях и вуали с мушками, садится в сани* («Дар»). Этот пример как раз в тексте более позднем, чем РсК, так что его можно засчитать в пользу гипотезы Струве. И все-таки он не «дотягивает» до *мысленно зажмуриться*. Зато у Набокова возникает еще более смелый неологизм — словечко *разжмуривать*: 1. об открытом новом номере журнале — *Илья Борисович хотел распахнуть один из них, книга сладко хрустнула, но не разжмурилась — еще слепая, новорожденная* («Уста к устам» 1929); 2. *Все слилось окончательно, но он еще на один миг разжмурился, оттого что зажегся свет, и Родион на носках вошел, забрал со стола черный каталог, вышел, погасло* («Приглашение…» 1935–36). И еще раз: 3. *Девочка во сне вздохнула, разожмурив пупок, и медленно, с воркующим стоном, дыхание выпустила* («Волшебник» 1939). Агеев же его не использует. Но и этот неологизм в НК все-таки не уникален: он встречался ранее, согласно — у Жаботинского, Замятина и А. Белого.

5. Набоков использует глагол *целиться* для описании такого специфического действия, когда человек только прикидывает, примеривается, готовится что-то сделать. В таких случаях обычно говорят *нацелиться* (+ инф.) или *нацелиться на* (что-то), уподобляя всю ситуацию прицельному метанию или стрельбе в цель. Этому словоупотреблению следует Агеев в единственном примере: 1. *Нелли, которая, с лицом внезапно заболевшего человека, в нетерпении то опускалась локтями на стол, то снова выпрямлялась, при этом не спуская глаз с Хирге, словно прицеливалась, откуда лучше откусить: сверху или снизу*. У Набокова же элементы этой конструкции (или целого фрейма) с глаголами

---

[19] Именно такого рода «дуплеты», повторы один в один, на мой взгляд, отчетливее всего и выдают единого автора.

целиться / прицелиться / попасть / промазать повторяются во множестве текстов — в целом более 10 раз (6 раз до РсК и еще 5 одновременно или после него). Вот только два из них: 1. о Подтягине, который от сердечного приступа падает в обморок — *Старик мутно глянул на него, сделал движение рукой, как будто целился на муху, и вдруг с легким клекотом зашатался, повалился вперед* («Машенька» 1926) — то есть, видимо, старик так наклонил голову, что можно было подумать, будто он хочет *поймать* (или *хлопнуть*) севшую муху. Тут, кстати, и характерное Набоковское подтрунивание, даже издевательство над таким героем, больным стариком, к которому принято выражать сочувствие, и языковое нарушение: по МАС, *целиться на* ч-л. — это ‘направлять свои действия на к-л. цель, метить’, но все-таки ведь не на такую конкретную цель, как муха! Еще пример — с действием уже без обозначения первого компонента как такового (*целиться*), зато с ответной на него реакцией, позволяющей его идентифицировать (такой как *попал* или *не попал*): герой рассказа кивает кому-то в толпе в знак приветствия, чтобы получить такой же кивок в замен, но получает ответ уже совсем от другого человека, которого и не думал приветствовать, метафорически это уподоблено стрельбе в цель. *Он увидел, среди чужих, некоторые знакомые лица, — вон Кочаровский — такой милый, круглый, — кивнуть ему… кивнул но не попал: перелет, — в ответ поклонился Шмаков* («Музыка» 1932). Но и этого метонимического развития ситуации у Агеева опять же не происходит.

Специфическая метафоричность

6. Вот иллюстрация того, что Струве проницательно назвал «динамизацией источника света» у Набокова: *На улице асфальт отливал лиловым блеском; солнце путалось в колесах автомобилей* («Машенька» 1926). Как будто похожее встречаем и у Агеева: *На балконе от заходящего, выпуклого как желток сырого яйца, солнца, хоть и зацепившего за крышу, однако видимого целиком, словно оно прожигало эту крышу насквозь, — лица стали махрово-красными*. Но и в НК находим, с одной стороны, еще более сходное с примером Набокова: *В грузной синеве золотым пылающим колесом ярилось промытое талыми ветрами солнце* (Б. Лавренев «Сорок первый» 1924). А с другой стороны, на мой взгляд, более сходное с примером Агеева: *В шестом часу утра, — когда солнце чуть пробует зацепиться красноватыми лучиками за шпили московских церквей, <…> с ходынского аэродрома выехала извозчичья пролетка с поднятым верхом* (Г. В. Алексеев «Подземная Москва» 1924). С третьей же стороны, еще один возможный источник фразы Агеева — это ранние издания 2-й части «Тихий Дона» (1928): (2:21 — в рук., с.96) *болтался в синеватой белеси неба солнечный желток*. Согласно (Гура 1960: 137), Шолохов был «склонен изобретать вычурные сравнения». После 1947 это место было исправлено на вполне нейтральное: *плыло в синеватой белизне неба солнце*. Возможно также заимствование и из более близкого к РсК по времени текста (в свою очередь взятое скорее всего у того же Шолохова): *И хотя в небе полное, как желток в эмалевой сковороде,*

лежало солнце, до настоящей весны было еще далеко (Л. Леонов «Скутаревский» 1930–1932)[20].

7. Для Набокова характерно постоянное пристальное внимание к **музыке** — при несколько шутливом к ней отношении. Сам писатель заявлял, может быть, демагогически, что в музыке ничего не смыслит. К примеру, арфа у него предстает — как нога страуса: 1. *Когда мы вошли в кафе, там играл дамский оркестр; я мимоходом заметил, как в одной из граненых колонн, облицованных зеркалами, отражается страусовая ляжка арфы* («Весна…»); а оркестр арфисток (там же) — как группа ткачих за работой: 3. *оркестр из полудюжины прядущих музыку дам* (к тому же не знающих, по его выражению, *куда девать грудь, лишнюю в мире гармонии*). Ну, или ветер — как дирижер оркестра: 4. *норд-ост, рассыпающий ноты оркестра в городском саду* («Машенька»). Нечто подобное «птичьим» у Набокова образам для арфы или рояля встречаем и в РсК: *И тотчас раздались тоненькие звуки и такие разрозненные, словно курица гуляла по арфе*. Однако точного подобия здесь нет.

Возможное мелкое «заимствование» у Набокова в следующем примере из РсК: *Яг, через гулкую залу, где кресла, рояль и люстра были в белых чехлах, провел нас в свою комнату*. Однако все же «лошадиной» метафоры тут у Агеева нет — такой, какая была у Набокова: *отец, осторожно прикоснувшись губами к его хохолку, уходил, — мимо позолоченных стульев, мимо обширного зеркала, мимо копии с купающейся Фрины, мимо рояля, большого безмолвного рояля, подкованного толстым стеклом и покрытого парчовой попоной* («Защита…» 1929–30). Т.е. «первоисточник» выглядит гораздо более поэтично. Тот же образ, кстати, с авторским самоповтором, встретится у Набокова и еще через полтора десятка лет: *ты перешла крыльцо; остановилась; локтем мягко открыла стеклянную дверь; миновала чепраком покрытый рояль, пересекла вереницу прохладных, пропахших гвоздикой комнат* («Под знаком…» 1947).

Герой в повести Агеева: 1. *сел на кончик стула, поставив графин себе на колено и держа его за горлышко — совсем как отдыхающий скрипач…* Тут же метафора звукового ряда как бы уступает сравнению со зрительным образом: 2. *Как раз, когда я протиснулся в зал, скрипач, уже со скрипкой, вставленной под подбородок, торжественно поднял смычок и, привстав на цыпочках и подняв плечи, — вдруг опустился, и (движением этим рванув за собой пианино и виолончель) заиграл*. И у Набокова много «музыкальных» сравнений, метафор с каламбурами и контаминациями, игрой на фразеологии, как, например: *бритое лицо музыканта, державшего почему-то телефонную трубку, как скрипку, между щекой и плечом* («Защита…» 1929–30). — Возможно Агеев в приведенном выше примере и вдохновлялся этим, но повтора в точности Набоковского сравнения там не было, а у последнего не было повторов Агеева.

---

[20] Но также сама метафора солнца как *желтка* была и в ранних стихах — у Эренбурга, Зенкевича (как раз у Набокова этой последней грубоватой метафоры не встречалось).

8. Попытаемся теперь подвести примеры обоих авторов под единый «образец» словосочетания (как это было сделано в Михеев 2012): *Провода* (телеграфные / телефонные / столбы / проволоки) + (как *канаты // струны*) + (взмывать / взлетать / взмах / взлет // стоять). У Агеева под этот образец ложится пример: *В такой жаркий московский вечер, когда падает первый снег, когда щеки в брусничных пятнах, а в небе седыми канатами стоят провода*. А у Набокова ему соответствует сразу несколько примеров (это опять-таки устойчивый маркер его стиля): 1. *В окно, сквозь стеклянную дверь в проход видать было, как взмывают ровным рядом телеграфные струны* («Случайность» 1924) и там же 2. *Черный телеграфный столб пролетел, перебил плавный взмах проволок*; 3. *Он лег навзничь на полосатый тюфяк лавки и в пройму дверцы видел, как за коридорным окном поднимаются тонкие провода среди дыма горящего торфа и смуглого золота заката* («Машенька» 1926); 4. *Мартынъ нашелъ все, что любилъ — телеграфные столбы, обрывающіе взлетъ проводовъ, вагонъ-ресторанъ* («Подвиг» 1931–32). Все четыре примера Набокова — из текстов предшествующих РсК, а после него этот образ у него почему-то пропадает. Хотя более близких данному образцу примеров, чем в РсК, во всем НК найти мне не удалось, но все же расхождение образов у Агеева и Набокова налицо: у последнего (по Струве) вместо статичной формы сравнения оказывается динамичная: *провода*, или *столбы*, на которых висят провода, не просто *стоят* как натянутые *струны*, но — *взмывают*, *поднимаются* или *взлетают*, так как герой движется, едет в поезде.

9. И последний пример: проверим соответствие образцу — *дверь, сосущая воздух*. У Агеева: *я же, сбежав по уже опустевшей лестнице и открывая тугую, шумно сосущую воздух дверь, хоть и оглянулся и посмотрел на мать, однако, сделал это не потому вовсе, что мне стало ее сколько-нибудь жаль, а всего лишь из боязни, что она в столь неподходящем месте расплачется*. Почти тот же образец встречаем и у Набокова, но в тексте, написанном позднее РсК («Весна в Фиальте» 1936): *и от нашего сквозняка всосался и застрял волан белыми далиями вышитой кисеи промеж оживших половинок дверного окна, выходившего на узенький чугунный балкон, и лишь тогда, когда мы заперлись, они с блаженным выдохом отпустили складку занавески*. — Мог ли тут уже сам Набоков «вдохновиться» образом Агеева (надо признать, более «топорным»), и доработав его по-свойски, превратить — в «сквозняк», пронизывающий отношения героя и его легкомысленной возлюбленной? В НК наиболее близок этому образу оказывается пример из Булгакова: *Сосущая с тихим змеиным свистом воздух пружина-цилиндр на железной двери выпускала меня* («Театральный роман» 1936–1937).

Подведем итог. В самом деле, некоторая вероятность того, что «Роман с кокаином» первоначально был написан Набоковым, все же остается, но — самая минимальная. Практически во всех случаях, когда повтор уникального идиостилистического приема у Набокова вслед за Агеевым интуитивно кажется очевидным, при ближайшем рассмотрении оказывается, что или сам

прием не уникален (а «изобретен» и использован кем-то еще, ранее их обоих), или же точного «копирования» его Набоковым не происходит — он всегда его как-то развивает или деформирует. Когда же, наоборот, прием вначале был употреблен Набоковым, еще до РсК, то и его точного копирования у Агеева тоже не происходит, такого, которое было бы похоже на самоповтор. Ну, а доказывать, что диапазон экспериментирования со словарем у Набокова в несколько раз шире, чем у Агеева, излишне.

(Замысел этой работы обязан своим возникновением возвращению с «Диалога-2012» и поездке автора в одной машине — с Еленой Гришиной и Светланой Савчук, которым моя искренняя благодарность.)

Литература

1. *Агеев* 1934 — Агеев М. Паршивый народ // Встречи. Париж, апрель 1934, с. 161–7.
2. *Агеев* 1936 — Агеев М. Роман с кокаином. Париж: Издательская коллегия парижского объединения писателей, 1936.
3. *Агеев* 1990 — Агеев М. Роман с кокаином. М. 1990.
4. *Агеев* 2000 — Агеев / Марк Леви. Роман с кокаином. М.: Согласие, 2000.
5. *Волчек* 1989 — Дмитрий Волчек. Загадочный господин Агеев // Родник, 1989.
6. *Гура* 1960 — Гура В. В. Жизнь и творчество М. А. Шолохова. М. 1960.
7. *Михеев* 2012 — М.ШОЛОХОВ или все-таки — Ф.КРЮКОВ? Неформальные процедуры при установлении авторства «Тихого дона» // Диалог. Международная конференция по компьютерной лингвистике. М. 2012.
8. *Мухин* 2001 — Мухин Н. Ю. Кто написал «Роман с кокаином»? Опыт лингвостатистического исследования. Екатеринбург, 2001.
9. *Рагозина* 2000 — Ксения Рагозина. Детектив с романом // Русский журнал. Предисловие к книге: М. Агеев / Марк Леви. Роман с кокаином. — М.: Согласие, 2000 [Дата эл. публикации: 23.3.2000].
10. *Парфенов* 1995 — Мишель Парфенов. М. Агеев (1898–1973). Загадка в пяти действиях // М. Агеев. Роман с кокаином и Паршивый народ. Париж 1995.
11. *Струве* 1986 — Никита Струве. Спор вокруг В.Набокова и «Романа с кокаином» // Вестник РСХД. Париж 1986 № 146.
12. *Струве* 1990 — Никита Струве. Роман-загадка [в предисловии к книге] // Марк Агеев. Роман с кокаином. М. 1990.
13. *Струве* 1995 — Никита Струве. Еще об авторстве «Романа с кокаином» // Вестник РСХД. Париж 1995 № 172.
14. *Суперфин, Сорокина* 1994 — Суперфин Г. Г., Сорокина М. Ю. «Был такой писатель Агеев…» Версия судьбы, или о пользе наивного биографизма // Минувшее. СПб. Вып.16. 1994.
15. *Толстой* 1995 — Иван Толстой. Тропóю трóпа, или почему Набоков не был автором «Романа с кокаином» // Звезда 1995 № 3.

References

1. *Agejev* 1934 — Agejev M. Parshyvyj narod // Vstrechi. Paris 1934, p.161-7.
2. *Agejev* 1936 — Agejev M. Roman s cocainom. Paris, 1936.
3. *Agejev* 1990 — Agejev M. Roman s cocainom. Moscow, 1990.
4. *Agejev* 2000 — Agejev / Marc Levi. Roman s cocainom. Moscow, 2000.
5. *Gura* 1960 — Gura V. V. Zhizn' i tvorchestvo M.A.Cholohova. Moscow, 1960.
6. *Miheev M. Ju.* Cholohov ili Krjukov? (…) 2012, available at http://www.dialog-21.ru/digest/2012/?type=main
7. *Muhin* 2001 — Muhin N. Ju/ Kto napisal "Roman s cocainom"? Opyt lingvostatisticheskogo issledovanija. Ekaterinburg, 2001
8. *Ragozina* 2000 — Xenija Ragozina/ Detektivy s romanom // Russkij zhurnal. Moscow, 2000
9. *Parfenov* 1995 — Mishel' Parfenov. M. Agejev (1898–1973). Zagadka v 5 dejstvijah // M. Agejev. Roman s cocainom i parshyvyj narod. Paris, 1995.
10. *Nikita Struve* 1986 — Nikita Struve. Spor vokrug V.Nabokova… // Vestnik RSHD. Paris, 1986 № 146.
11. *Nikita Struve* 1995 — Nikita Struve. Roman-zagadka // Agejev M. Roman s cocainom. Moscow, 1990.
12. *Nikita Struve* 1995 — Nikita Struve. Eshcho ob avtorstve… // Vestnik RSHD. Paris, 1995 № 172.
13. *Superfin, Sorokina* 1994 — Superfin G. G., Sorokina M. Ju. "Byl takoj pisatel' Agejev…" Versija sud'by, ili o pol'ze naivnogo biografizma // Minuvsheje. SPb. Vyp.16, 1994.
14. *Tolstoj* 1995 — Ivan Tolstoj. Tropoju tropa, ili pochemu Nabokov ne byl avtorom "Romana s cocainom" // Zvezda 3, 1995.
15. *Volchek* 1989 — Dmirtij Volchek. Zagadochnyj gospodin Agejev // Rodnik, 1989.

A COREFERENTIALLY ANNOTATED CORPUS AND ANAPHORA RESOLUTION FOR CZECH

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz),
Mírovský J. (mirovsky@ufal.mff.cuni.cz),
Novák M. (mnovak@ufal.mff.cuni.cz)

Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic

The paper presents an overview of a finished project focused on annotation of grammatical, pronominal and extended nominal coreference and bridging relations in the Prague Dependency Treebank (PDT 2.0). We give an overview of existing similar projects and their interests and compare them with our project. We describe the annotation scheme and the typology of coreferential and bridging relations and give the statistics of these types in the annotated corpus. Further we give the final results of the inter-annotator agreement with some explanations. We also briefly present the anaphora resolution experiments trained on the coreferentially annotated corpus and the future plans.

Keywords: anaphora, annotation, bridging relations, coreference, coreference resolution

1. Introduction

Coreferential and bridging relations between discourse entities are of major importance for establishing and maintaining textual coherence. The ability to automatically resolve these kinds of relations is an important feature of text understanding systems. The Prague Dependency Treebank (PDT 2.0) (Jan Hajič et al., 2006) is a manually annotated corpus of Czech. The texts are annotated in three layers — morphological, analytical and tectogrammatical. The most abstract (tectogrammatical) layer includes among other mark-ups the annotation of coreferential links. The whole corpus contains almost 50 thousand sentences. In this paper we present an overview of the projects of annotating different types of coreference and bridging relations in the Prague Dependency Treebank, speak about the results of inter-annotator agreement and summarise some anaphora resolution experiments made on Czech data.

Section 2 describes the state of the art concerning annotating, analysing and using coreferentially annotated corpora. Section 3 gives a short overview of the types of coreference and bridging relations annotated in PDT. In Section 4, we give the statistics and discuss the results of inter-annotator agreement. Section 5 describes some anaphora resolution experiments that were made using the Czech coreferentially annotated data. We make conclusions in section 6.

2. PDT coreference and similar projects

The experiments on anaphora resolution, referential choice prediction, etc. are made using the annotated corpora for coreference. There are a number of different large-scale annotated corpora for coreference and anaphoric relations on which the experiments for coreference resolution are made. The largest annotated corpora for English include MUC (Hirschman and Chinchor, 1997), ACE (Doddington et al., 2004), OntoNotes (Pradhan et al., 2007), GNOME (Poesio, 2004), ARRAU (Poesio and Artstein, 2008). The coreference annotations for other languages than English are more limited. The most well-known corpora including anaphoric informations are AnCora (Recasens and Marti, 2009) for Spanish and Catalan, VENEX (Poesio et al., 2004) for spoken and written Italian, the Italian Live Memories Corpus (Rodríguez et al., 2010), TüBA-DZ Treebank (Hinrichs et al., 2004) and Postdam Commantary Corpus (Stede, 2008; Krasavina and Chiarcos, 2007) for German, PdiTB (Poláková et al., 2012) etc.

Determining coreference is a highly complicated task, and even between human annotators there is a lot of disagreement leading to a relatively low number of inter-annotator agreement, especially concerning nominal coreference and bridging relations. The cases of vagueness and referential ambiguity were a subject of a rich discussion in computational linguistics and anaphoric community during the last few years. There were discussed such topics as e.g. justified sloppiness hypothesis in Poesio et al. 2006, the reliability of anaphoric annotation in Poesio and Artstein 2005, examples and reasons for vagueness and referential ambiguity in Versley 2008, so-called near-identity relation in Recasens et al. 2010. Some discussion on ambiguous cases of coreference and the reasons for inter-annotator disagreement for Czech were presented in Nedoluzhko 2010.

3. Types of coreference and bridging relations annotated in PDT

In PDT 2.0, two types of coreference (grammatical and textual) and six types of bridging relations have been annotated. The **grammatical coreference** typically occurs within a single sentence, the antecedent being able to be derived on the basis of grammar rules of a given language. It includes relative pronouns, verbs of control, reflexive pronouns, reciprocity and verbal complements. The detailed description of the types of grammatical coreference and the examples may be found in Mikulová et al. 2006.[1] **Textual coreference** is generally taken to mean the use of various linguistic means (pronouns, synonyms, generalizing nouns etc.) which function as anaphoric (occasionally cataphoric) reference devices. This reference is not expressed by grammatical means alone, but also via context. As for textual coreference in PDT, it has been annotated in two time periods. At first, the so-called pronominal textual coreference was manually annotated. It was restricted to cases in which a demonstrative

[1] The resumed typology of grammatical coreference in PDT was also presented at DIALOG in Nedoluzhko 2009.

this or an anaphoric pronoun of the 3rd person, also in its zero form, are used (Kučová and Hajičová, 2004). Afterwards, the annotation of textual coreference was extended to cases where the anaphor is expressed by other means such as full noun phrases (definite descriptions, repetitions, synonyms etc.), adverbs (there, then etc.) and some types of numerals and pronouns neglected in the first stage. This stage of the project was called the Extended Textual Coreference and described in detail during the annotation period in (Nedoluzhko et al. 2009; Nedoluzhko, 2011; Nedoluzhko and Mírovský, 2011). Annotation of extended textual coreference and bridging relation is a project related to PDT 2.5 (Bejček et al., 2011), which is a revised, updated and extended version of PDT 2.0.

The textual coreference is further classified into two types — coreference of NPs with specific (type SPEC) or generic (type GEN) reference. Compare examples (1) and (2):

(1) *Mary and John went together to Israel, but Mary* [type SPEC] *had to return because of the illness.*

(2) *Lions live in a forest. They are not vegetarians* [type GEN].

Special cases of textual coreference. Two special cases of reference are annotated in PDT. First, the textual coreference covers the cases of endophoric references to discourse segment of more than one sentence, including also the cases where the antecedent is understood by inferencing from a broader co-text. This kind of relation has no explicitly marked antecedent, it just proves the fact that the given anaphoric NP corefers with some discourse antecedent of more than one sentence. We consider this decision to be provisional and we plan to complete it later. Second, a specifically marked link for exophora denotes that the referent is "out" of the co-text, it is known only from the actual situation. In the same way as for segments, the new nominal and adverbial links are being added.

For the **bridging relations**, the following types are distinguished: part-of relation (*room — ceiling*), set — subset (*students — some students*) and FUNCT (*trainer — football team*) traditional relations, CONTRAST for coherence relevant discourse opposites, ANAF for explicitly anaphoric relations without coreference and the further underspecified group REST. The more detailed description of types can be found for example in Nedoluzhko and Mírovský 2011.

4. Statistics and inter-annotator agreement

By the end of 2011, the whole PDT data was annotated for coreference and bridging relations (see Nedoluzhko et al. 2011).[2] Table 1 shows the statistics of the annotated data.

[2] The completed and corrected version was published together with the annotation of discourse relations in the Prague Discourse Treebank in 2012 (see Poláková et al. 2012).

Table 1. Statistics of the annotated data

Total number of sentences (in the annotated documents)	49,431
Total number of tokens	833,195
Number of coreferring nodes — grammatical coreference	23,272
Number of coreferring nodes — textual coreference	86,349
Number bridging relations	32,171
% of co-referring nodes	17,6%

As for the distribution of types of textual coreference and bridging relations, the proportion is represented in Table 2:

Table 2. The distribution of types of textual coreference and bridging relations

Type	**Number**
Textual coreference (specific)	20,243 (pronouns) + 50,593 (nouns) = 70,836
Textual coreference (generic)	3,095 (pronons) + 12,418 (nouns) = 15,513
All textual coreference links	86,349
All bridging links	32,171

As seen from the table, textual coreference makes the significant majority of the annotated relations and inside the group of textual coreference the coreference of specific noun phrases significantly prevail. The reason for relatively low percentage of bridging relations may be mainly the small number of types and their precise delimitating (even for annotation of the bridging relation of type REST, very precise rules were set). As for the significant dominance of textual coreference between specific noun phrases over generic ones, the reasons are mainly empiric. Also postulating coreference between generic noun phrases is a much more complicated task than coreferring specific noun phrases, so in most existing projects it is excluded from the annotation of coreference (Poesio, 2004; Recasens, 2010 etc.).

We have measured the inter-annotator agreement in the annotation of coreference and bridging anaphora in PDT on a small part of the data that had been annotated in parallel by two annotators. To evaluate the agreement, we have used the chain-based F1-measure, a simple ratio, and Cohen's κ (Cohen, 1960). The chain-based F1-measure has been used for measuring the agreement on the recognition of a coreference or bridging relation, a simple ratio and Cohen's κ have been used for measuring the agreement on the type of the relations in cases where the annotators recognized the same relation.

In the chain-based measure, we consider the annotators to be in agreement on recognizing a coreference or bridging relation if the two nodes connected by an arrow by one of the annotators have also been connected by the other annotator; coreference chains are taken into account, i.e. it is sufficient for the agreement if the arrow starts in or goes to a node that is coreferentially connected (possibly transitively) with the node used for the relation by the other annotator.

Table 3. Results of the inter-annotator agreement

Measurement	F1	Agreement on types	Kappa on types
All parallel data — coreference	0.72	0.90	0.73
All parallel data — bridging anaphora	0.46	0.92	0.89

Table 3 shows that the results for inter-annotator agreement are not particularly high. In our measurements and analyses of inter-annotator agreement, we observe the three main types of disagreement: (a) disagreement concerning the decision if the relation in question should be annotated as a coreference (or bridging) relation, (b) disagreement on choosing the antecedent and (c) disagreement in the type of the annotated relation. The reasons for relatively low numbers of inter-annotator discrepancies and the typology of disagreements with the examples were discussed in Nedoluzhko 2010.

5. Automatic experience on the annotated data

The main objective of our annotation effort has been to provide data for developing automatic techniques for resolution of anaphoric relations. PDT has served as a source of gold standard data for testing as well as a source of training data for tools utilizing machine learning methods.

Antecedents in grammatical coreference can be usually derived with high accuracy from grammatical rules. Nguy 2006 presented a set of rules for various types of grammatical coreference, achieving more than 90% F1-measure for every type.

In Nguy and Žabokrtský 2007, a rule-based system was employed to resolution of pronominal textual coreference. Higher complexity of this task affects the success rate which is substantially lower (74% F1-measure) than what can be reached in the task of grammatical coreference resolution. Applying machine learning methods, particularly perceptron ranking in Nguy et al. 2009, on the same task outperformed the rule-based method with F1-measure over 79%.

However, the features used in these experiments were extracted from the manually annotated tectogrammatical layer of PDT 2.0. Thus the system could take advantage of perfectly correct information on various linguistic attributes which are not available in a real-world scenario. In Bojar et al. 2012, the authors used the same perceptron ranker and the same feature set for training and testing, this time extracted from the automatically analyzed data though. Unreliability of information on tectogrammatical gender and number as well as uncertainty of presence of a subject omitted on the surface[3] resulted in a substantial drop in performance to 50% F1-measure.

It confirms that correct identification of an unexpressed subject and determination, whether it is anaphoric, is central to resolution of the zero variant of pronominal coreference. This and a corresponding issue in English — determination of whether a personal pronoun "it" is anaphoric — were addressed in the work of Veselovská et al. 2012

[3] Czech is a pro-drop language.

by a set of rules tested on Prague Czech-English Dependency Treebank 2.0 (PCEDT). Some of these rules made use of parallel nature of the treebank by providing information from the English side to facilitate identification of Czech unexpressed subjects.

Annotation work on the Extended Textual Coreference project encouraged research on noun phrase (NP) textual coreference resolution. Novák 2010 carried out the first experiments on NP coreference in Czech. The approach of maximum entropy ranking was further elaborated in Novák and Žabokrtský 2011, where authors compared systems based on classification and ranking approaches in machine learning. As a result, the best system achieves 44.4% F1-measure on coreference with specific reference. Novák 2010 also paid his attention on coreference with generic reference as well as bridging relations of the type PART. Despite the unsatisfying results, his work introduces a novel feature inspired by Hearst patterns (Hearst, 1992) that is supposed to capture a PART-WHOLE relation by exploiting a large morphologically annotated corpus.

Knowledge of anaphoric relations in a text can be crucial to solving more complex tasks. Multiple tools mentioned above have been integrated with a modular NLP framework Treex (Popel and Žabokrtský, 2010) that is used in various scenarios. For instance, the rules for resolving grammatical and pronominal textual coreference contribute on English to Czech translation in TectoMT system (Žabokrtský et al., 2008). In addition, some of these tools and their counterparts for English helped to form both sides of the automatically annotated Czech-English parallel corpus CzEng 1.0 (Bojar et al., 2011), consisting of over 15 million sentence pairs.

The overview of performance of the tools mentioned above can be found in Table 4.

Table 4. The overview of performance of the tools

Type of the task	Published	Data	Success rate
Grammatical coreference, verbs of control	Nguy 2006	PDT 2.0	91.5%
Grammatical coreference, reflexive pronouns	Nguy 2006	PDT 2.0	97.1%
Grammatical coreference, relative pronouns	Nguy 2006	PDT 2.0	99.6%
Grammatical coreference, reciprocity	Nguy 2006	PDT 2.0	94.7%
Pronominal coreference, rule-based	Nguy and Žabokrtský 2007	PDT 2.0	74.2%
Pronominal coreference, perceptron ranking, gold features	Nguy et al. 2009	PDT 2.0	79.4%
Pronominal coreference, perceptron ranking, system features	Nguy et al. 2009	PDT 2.0	50.3%
Identification of an anaphoric unexpressed subject, rule-based	Veselovská et al. 2012	PCEDT 2.0	61.5%
Identification of an anaphoric unexpressed subject, rule-based, expoliting English side	Veselovská et al. 2012	PCEDT 2.0	69.5%
NP coreference, maximum entropy ranking	Novák 2010	PDT 2.5	39.4%
NP coreference, perceptron ranking, improved features	Novák and Žabokrtský 2011	PDT 2.5	44.4%

6. Conclusion and future work

We presented the finished project of the Czech annotation of different types of coreference and bridging relations. We compared our project to other similar projects, gave the statistics of coreference and bridging types and the results for inter-annotator agreement. We briefly presented the anaphora resolution experiments trained on coreferentially annotated corpus.

At present, we are completing the annotation for the first and second person coreference. In future, other corpora for Czech (e.g. the Prague Dependency Treebank of Spoken Czech, Prague Czech-English Dependency Treebank) are to be supplied with some types of coreferential relations.

Acknowledgements

We gratefully acknowledge the support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875) and GAUK 4226/2011.

1. *Bojar, Ondřej; Žabokrtský, Zdeněk; Dušek, Ondřej; Galuščáková, Petra; Majliš, Martin; Mareček, David; Maršík, Jiří; Novák, Michal; Popel, Martin; Tamchyna, Aleš:* CzEng 1.0. Data, Charles University in Prague, UFAL, 2011.
2. *Bojar, Ondřej; Žabokrtský, Zdeněk; Dušek, Ondřej; Galuščáková, Petra; Majliš, Martin; Mareček, David; Maršík, Jiří; Novák, Michal; Popel, Martin; Tamchyna, Aleš:* The Joy of Parallelism with CzEng 1.0. In Proceedings of LREC 2012, İstanbul, 2012.
3. *Cohen, Jacob:* A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1), 1960.
4. *Doddington, George; Mitchell, Alexis; Przybocki, Mark; Ramshaw, Lance; Strassel, Stephanie; Weischedel, Ralph:* The Automatic Content Extraction (ACE) program — tasks, data, and evaluation. In Proceedings of LREC 2004, Lisbon, 2004.
5. *Hajič, Jan et al.:* Prague Dependency Treebank 2.0.CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.
6. *Hearst, Marti A.:* Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th Conference on Computational linguistics — Volume 2, Nantes, France, 1992.
7. *Hinrichs, Erhard; Kübler, Sandra; Naumann, Karin; Telljohann, Heike; Trushkina, Julia:* Recent developments in linguistic annotations of the TüBa-D/Z treebank. In Proceedings of the Third Workshop on Treebanks and Linguistic Theories, Tübingen, 2004.
8. *Hirschman, Lynette; Chinchor, Nancy:* MUC-7 Coreference Task Definition — Version 3.0, 1997.
9. *Krasavina, Olga; Chiarcos, Christian:* PoCoS — Potsdam Coreference Scheme. In Proceedings of the Linguistic Annotation Workshop, Prague, 2007.

10. *Kučová, Lucie; Hajičová, Eva:* Coreferential Relations in the Prague Dependency Treebank. In Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium, S. Miguel, 2004.
11. *Mikulova Marie et al.:* Anotace na tektogramatické rovině Pražského závislostního korpusu. Referenční příručka. Technical report no. 2006/31, Charles University in Prague, UFAL, 2006.
12. *Nguy, Giang Linh:* Proposal of a set of rules for anaphora resolution in Czech. Master thesis, Charles University in Prague, 2006.
13. *Nguy, Giang Linh; Novák, Václav; Žabokrtský, Zdeněk:* Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In Proceedings of the SIGDIAL 2009 Conference, London, 2009.
14. *Nguy, Giang Linh; Žabokrtský, Zdeněk:* Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data. In Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium, Lagos, 2007.
15. *Nedoluzhko, Anna:* Coreferential relationships in text — comparative analysis of annotated data. In Papers from the Annual International Conference "Dialogue 2010" Issue 9 (16), Moscow, 2010.
16. *Nedoluzhko, Anna:* Razmetka koreferencii na sintaksičeski annotorovannom korpuse češskich tekstov. In Papers from the Annual International Conference "Dialogue 2009" Issue 8 (15), Moscow, 2009.
17. *Nedoluzhko, Anna; Mírovský, Jiří; Ocelák, Radek; Pergler, Jiří:* Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank. In Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium, Goa, 2009.
18. *Nedoluzhko, Anna; Mírovský, Jiří; Pajas, Petr:* Annotation Tool for Extended Textual Coreference and Bridging Anaphora. In Proceedings of LREC 2010, Malta, 2010.
19. *Nedoluzhko, Anna; Mírovský, Jiří:* Annotating extended textual coreference and bridging relations in the Prague Dependency Treebank. Technical report no. 2011/44, Charles University in Prague, UFAL, 2011.
20. *Nedoluzhko, Anna; Mírovský, Jiří, Hajičová, Eva, Pergler, Jiří, Ocelák, Radek:* Extended Textual Coreference and Bridging Relations in PDT 2.0. Data. Charles University in Prague, ÚFAL, 2011.
21. *Nedoluzhko, Anna:* Rozšířená textová koreference a asociační anaphora (Koncepce anotace českých dat v Pražském závislostním korpusu). ÚFAL, Praha, 2011.
22. *Novák, Michal:* Machine Learning Approach to Anaphora Resolution. Master thesis, Charles University in Prague, 2010.
23. *Novák, Michal; Žabokrtský, Zdeněk:* Resolving Noun Phrase Coreference in Czech. In Lecture Notes in Computer Science 7099, Springer-Verlag Heidelberg, 2011.
24. *Poesio, Massimo; Delmonte, Rodolfo; Bristot, Antonella; Chiran, Luminita; Tonelli, Sara:* The Venex corpus of anaphora and deixis in spoken and written Italian. Manuscript, 2004.

25. *Poesio, Massimo; Artstein, Ron:* The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, Ann Arbor, 2005.
26. *Poesio, Massimo; Sturt, Patrick; Artstein, Ron; Filik, Ruth:* Underspecification and Anaphora: Theoretical Issues and Preliminary Evidence. In Discource Processes 42(2), 2006.
27. *Poesio, Massimo:* The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In Proceedings of The 5th SIGdial Workshop on Discourse and Dialogue, Boston, 2004.
28. *Poesio, Massimo; Artstein, Ron:* Anaphoric annotation in the ARRAU corpus. In Proceedings of LREC 2008, Marrakech, 2008.
29. *Poláková, Lucie; Jínová, Pavlína; Zikánová, Šárka; Hajičová, Eva; Mírovský, Jiří; Nedoluzhko, Anna; Rysová, Magdaléna; Pavlíková, Veronika; Zdeňková, Jana; Pergler, Jiří; Ocelák, Radek:* Prague Discourse Treebank 1.0. Data, Charles University in Prague, ÚFAL, 2012.
30. *Popel, Martin; Žabokrtský, Zdeněk:* TectoMT: Modular NLP Framework. In Lecture Notes in Computer Science, Vol. 6233, Springer-Verlag Heidelberg, 2010.
31. *Pradhan, Sameer S.; Hovy, Eduard; Marcus, Mitch; Palmer, Martha; Ramshaw, Lance; Weischedel, Ralph:* Ontonotes: A unified relational semantic representation. In Proceedings of the International Conference on Semantic Computing, Washington DC., 2007.
32. *Recasens, Marta; Hovy, Eduard; Martí, M. Antònia:* A typology of near-identity relations for coreference (NIDENT). In Proceedings of LREC 2010, Valletta, 2010.
33. *Recasens, Marta; Martí, M. Antònia:* AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation, 2009.
34. *Rodríguez, Kepaj; Delogu, Francesca; Versley, Yannick; Stemle, Egon W.; Poesio, Massimo:* Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In Proceedings of LREC 2010, Valletta, 2010.
35. *Stede, Manfred:* Disambiguating Rhetorical Structure. In Research on Language and Computation, Vol. 6, Issue 3–4, Springer Netherlands, 2008.
36. *Versley, Yannick:* Vagueness and referential ambiguity in a large-scale annotated corpus. In Research on Language and Computation 6 (3–4), Springer Netherlands, 2008.
37. *Veselovská, Kateřina; Nguy, Giang Linh; Novák, Michal:* Using Czech-English Parallel Corpora in Automatic Identification of 'It'. In The Fifth Workshop on Building and Using Comparable Corpora, İstanbul, 2012.
38. *Žabokrtský, Zdeněk; Ptáček, Jan; Pajas, Petr:* TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, 2008.

THE PROSPECTS OF APPLICATION OF SEMANTIC MARKUP TO THE NAMED ENTITY RECOGNITION PROBLEM

Nekhay I. V. (nekhayiv@gmail.com)

Department of image recognition and text processing,
DIHT MIPT, Moscow, Russia

The paper describes an attempt to construct a Named Entity classifier upon ABBYY Compreno Syntactic and Semantic Parser that was presented at the "Dialogue" conference in 2012. The classifier employs supervised learning technique, namely the Conditional Random Fields model, developed under heavy constraints on the available feature set: no external NE lists or non-local features are used. The system is evaluated on the NER field's "gold standard" evaluation corpus of CoNLL-2003 achieving F-scores of 91.61% on dev and 87.51% on test set. The classifier outperforms several other systems developed under the same constraints on features, but underperforms a single system that makes use of significantly richer local context. The gain of individual classifier features based on parser attributes is explored; it is demonstrated that Compreno's semantic hierarchy and surface (syntactic) slots provide classifier with the most valuable features used to locate and classify NEs. This reliance on parser results, however, leads to error propagation from parser to classifier, as shown in the section on error analysis. Final conclusions make an attempt to offer several topics for following research.

Key words: semantic classification, named entity recognition, semantic and syntactic parser

1. Introduction

Continuing expansion of the Internet supports practical interest in methods of information extraction from unstructured texts. One subtask of information extraction is called named entity recognition (subsequently NER). This subtask consists of two problems: identification of named entity boundaries in text and further classification of named entities in a usually finite set of categories.

NER systems are usually based on text analysis systems of a much broader purpose. Analysis can vary in depth from shallow lexical or morphological, as in (Klein, Smarr, Nguyen, Manning, 2003), to integration of NER subsystem into a text parser (deep, syntactic or semantic) as described by Finkel and Manning in (Finkel, Manning, 2009). A research of capabilities of NER systems based on deep text analysis is of certain interest. Therefore a syntactic and semantic parser based on ABBYY Compreno technology, which was introduced at the "Dialog" conference in 2012 [(Anisimovich, Druzhkin, Minlos, Petrova, Selegey, Zuev, 2012) and (Bogdanov, 2012)] is the object of research described by this paper.

As the authors of the survey (Nadeau, Sekine, 2007) noted, NER solutions employ two major approaches: rule-based and statistics-based, mainly machine learning, approach. The development of a rule-based system is labour-intensive and the resulting system often trades recall for precision. The use of statistical methods, given that a sufficient amount of data is available, can significantly decrease the labour-output ratio for some tasks. That is why a statistical NER approach, using results of Compreno parser execution as source data, is considered in the current paper.

An influential comparison of language-independent NER algorithms was performed during CoNLL-2003 (Tjong Kim Sang, De Meulder, 2003) as the conference's shared task. Special corpora of news articles in English and German were prepared for this comparison. The English language CoNLL-2003 corpus has become de facto a standard for evaluation of works in the field of NER. A number of papers present results of assessments of different systems on this corpus.

Evaluation of Compreno parser applicability to the NER task in German would suffer from parser's immaturity at this moment, thus only English corpus of CoNLL-2003 was used for this research.

2. Task setup and main limitations

Evaluation method. A NER system evaluation methodology based on measurements of precision, recall and F-score was developed in the course of CoNLL conferences. The methodology will be described in greater detail in the following section. We will stick to this methodology for evaluation of our system. As our experiments show and some authors (e.g. (Tkachenko, Simanovsky, 2012), (Rosenfeld, Feldman, Fresko, 2006)) point out, the choice of features plays the most important role in the development of machine-trained NER classifiers. Therefore, the dependence of the integral F-score on the choice of features, obtained from the Compreno parser, will interest us in the first place. All our results were achieved by tuning the feature set solely; no changes or settings for a particular corpus were introduced into the learning algorithm or the parser.

Lexical features. Our approach to feature selection can be characterized as a refusal of local textual features in favour of a more generic, less language-dependent approach. The use of specific lexical features like specific separate words, word combinations, prefixes and suffixes is observed in the majority of works (CoNLL-2003 systems survey (Tjong Kim Sang, De Meulder, 2003) and later (Ratinov, Roth, 2009), (Tkachenko, Simanovsky, 2012)). In current research we don't use such features, relying on semantic descriptions of words in the semantic hierarchy, built into the parser. The resulting features are assumed to be more portable across distinct text genres, topics or even languages due to the universal inter-language nature of semantic descriptions.

External sources of information. The problem of incorporating external sources (they are also sometimes called gazetteers) such as the Wikipedia (Ratinov, Roth, 2009), DBPedia and YAGO (Tkachenko, Simanovsky, 2012) to extract list of named entities, that are later applied for NER, is another problem drawing attention of researchers. We do not use external sources, because we strive to obtain an evaluation of the parser's capabilities in «pure» form. Therefore only the published

values of F-measure which are achieved by researchers without use of external lists are chosen for comparison of results.

Local and non-local features. Due to the time constraints, we do not incorporate different types of non-local features into the system. All the explored features use only the current token, its left context, and its parent token according to the analysis tree, so the features are local only.

3. Corpus and evaluation method of CoNLL-2003

Description of the corpus. English corpus of CoNLL-2003 was created by complementing texts of Reuters news sub-corpus (about 300,000 words in size) with named entity markup according to 4 categories: person names (**PER**), organization names (**ORG**), locations (**LOC**) and all other NEs (**MISC**).

Corpus source texts were broken down into tokens, and each token has a label of respective category. CoNLL evaluation method takes into consideration the accuracy of both named entity bounds detection and classification into category. Only NEs with correctly identified bounds and category increase the values of precision, recall and F-score.

Integral F-scores, most often given in comparative papers, are calculated by micro-averaging in all four categories. More detailed description of the applied tokenization method and integral score calculation are given in (Tjong Kim Sang, De Meulder, 2003).

Corpus parts and concept drift. Initially the corpus is subdivided by its authors into three parts: training, development, and test. Training and development parts are chosen from news messages of August 1996, and test part — from reports of December 1996. As a consequence of this partition an effect, known as **concept drift**, occurs, caused by a significant change in primary persons and events appearing in news publications. A decrease of NER systems scores between development and test parts can be observed in all the papers devoted to this corpus (see table 1).

Table 1. Concept drift demonstrated by the results of top two CoNLL-2003 systems, F-score

Paper	**Florian, Ittycheriah, Jing, Zhang, 2003**			**Chieu, Ng, 2003**		
Corpus part	development	test	**drift**	development	test	**Drift**
NE label						
MISC	89.06	80.44	**−8.62**	88.41	79.16	**−9.25**
ORG	90.24	84.67	**−5.57**	88.56	84.32	**−4.24**
LOC	96.12	91.15	**−4.97**	95.57	91.12	**−4.45**
PER	96.60	93.85	**−2.75**	95.89	93.44	**−2.45**

Table 1 shows that the lowest concept drift effect can be observed in the **PER** category. It is supposedly caused by the fact that this category is the most easily formalized (and, therefore, has the most accurate markup in the corpus), and its tokens are well identifiable even by superficial lexical features, as in our study (Nekhay, 2012).

Manual exploration of classifier errors showed us that the **MISC** category, formed by residual principle, is the worst formal and contains in fact a union of other categories: nationalities, names of sports competitions, movies, etc. NEs of these categories often appear only in news messages of a narrow time period, when sufficient public interest in these subjects exists. We should also note that most errors in corpus markup ("Nato" and some other obvious NEs of category **MISC**, marked up as non-NEs) are also associated with this category.

Following assumptions can be made about the **ORG** and **LOC** categories. The usage frequency of names of important geographical objects (countries, capitals) depends at less extent on the time of publication, and name existence duration is maximal compared to **PER** and **MISC** categories. Similarly, a lot of mentions of organizations belong to large, international organizations which become international newsmakers at a constant frequency. Probably that is the reason why the **LOC** and **ORG** categories take intermediate position in table 1.

Counteracting the concept drift. In the survey (Nadeau, Sekine, 2007) its authors mention the change of text genre or domain (the latter observed between parts of CoNLL corpus) as one of the major challenges in the development of NER systems. On the other hand, in the latest works (Tkachenko, Simanovsky, 2012) the concept drift causes less significant drop of F-measure from 93.78 to 91.02. The drift is probably compensated by the use of external NE lists (gazetteers), as these lists must include with the same probability NEs appearing both in August and December of 1996. We presume that our refusal of use of local lexical features (words, suffixes, affixes, etc.) allows to reduce the effect of temporal shifts, but consider that this subject requires further research.

4. Classifier implementation

Description of the Compreno parser, upon which our classifier is based, goes beyond the current paper. The structure of the parser is described in works of (Anisimovich, Druzhkin, Minlos, Petrova, Selegey, Zuev, 2012) and (Bogdanov, 2012). As becomes clear further, the concepts of surface slot, parent and child constituent and path in the semantic hierarchy are the key concepts that form the most important features for our classifier.

4.1. Token synchronization

During the research a number of differences between data representations in the corpus markup and in the parser results were found. The most significant difference is tokenization method.

Since the parser allows adaptation of itself to a lot of different applications, we decided to keep corpus data in the original form and synchronize parse results to that form. A special algorithm that associates each corpus token to one or more parser tokens was developed for this task. However, due to the synchronization being imperfect, a certain fraction of classifier errors is caused by bad token alignment.

4.2. Classifier algorithm

Implementation. A Conditional Random Fields (CRF) model, trained by a L-BFGS algorithm provided by MALLET library (McCallum, 2002), was applied for NE identification and classification without introducing any changes to the algorithms.

Feature limitations. All features for the used algorithm are encoded as string values and treated as Booleans (true/false). Thus all the Boolean features are encoded in the most natural way — as a present feature for "true" value and as an absent feature for the opposite. Each of N-valued features is mapped into a set of N Boolean features, each corresponding to the source feature taking Nth value. This approach excludes the use of non-integer features and significantly limits available integer features to only those having a relatively small value set.

Alternative algorithms. Besides CRF, attempts were made to use decision trees and SVM for token classification. Upon increase of the number of features the main advantage of decision trees — their direct interpretation by a human — was lost, and feature mapping became rather hard to implement. SVM implementation that was applied required far more time for classifier training, though didn't provide a significant decrease of errors.

4.3. Features used by the algorithm

Before beginning to describe the features, it is important to note that the research used somewhat simplified and limited XML parser interface. This interface provides only one parse structure for each sentence (the "best" one according to the parser model) and a generalized view of the parser attributes. Attribute generalization has both positive and negative effects. While, on the one hand, it can lose significant information, it can increase classifier resilience to overfitting on the other. Access to more detailed internal parser structures was not implemented due to time limitations.

Surface-lexical features. Of the most often applied in NER field surface-lexical features, determined by token spelling, we use word case (**WCASE**: *first letter capital, all letters capital*, …) as well as a more detailed word case feature called **SHAPE**. The value of **SHAPE** is formed with a series of replacements: capital letters to the symbol "X", lowercase — to "x", digits — to "*d*". The replacements of first and last pairs of symbols remain on their places. Repetitions are excluded from other replacements, and the remaining symbols are sorted alphabetically. For example, the token "*Ireland-born*" has shape value of «*Xx-xxx*», and token «*1996-02-12*» — «*dd-ddd*». **WCASE** is considered for the current and preceding token; **SHAPE** — in the token window [−2..0].

Gazetteer-like features. External lists of NEs are not used; however, all NEs encountered during training are used to implement features named **PART_OF_(MISC|ORG|PER|LOC)**. Such feature is "true", if the current token is part of a NE in the corresponding category. To avoid overfitting these lists, randomly chosen 50% of them are used in the training phase, while at the test phase all 100% are looked through. The features are applied to the current token.

Surface-morphological features. For each word the parser determines part of speech, which is represented by our **POS** feature in [−1..0] token window.

Surface-syntactic features. For each word the parser defines two syntactic attributes: a surface slot (**SURFSL**: *Modifier_NominalEntityLike, Modifier_Attributive, Object_Indirect* …) and a simplified representation of the word's syntactic function in the sentence (**SYNTF**: *Subject, Preposition, AdverbialModifier*, …). For each token we consider these attributes of the token itself and of its parent (**PAR_SURFSL, PAR_SYNTF**), determined according the parse tree. These features are, perhaps, more dependent on the text language than others.

Deep-semantic features. The most significant for our work are features associated with semantic descriptions of words. The Compreno parser has at its foundation a semantic hierarchy (SH), which is a tree with semantic classes (SC) as nodes and lexical (roughly equivalent to word) classes (LC) as leaves. For each word the parser indicates its most probable LC and a few parent SCs along the path toward root in SH. This set of classes comprises the value of **EXLEXCLASS** feature, whose value is a vector of Booleans, corresponding to each of the parent SCs and showing, which of the SCs lie in the hierarchy path. For example, a lexical class "SOCCER" has a following set of semantic classes: *FOOTBALL : TYPES_OF_GAMES : SPORT : AREA_OF_HUMAN_ACTIVITY : … (*further parents omitted by simplified parser interface). Besides, we use several types of hierarchy path generalizations:

- Parser-defined attribute "NearestSensibleParent" (**NSP**), eliminating a lot of minor SCs. For the *soccer* example above its value would be *TYPES_OF_GAMES*.
- Artificially invented feature **ABR_EXLEXCLASS**, calculated by cutting from hierarchy path all lexical classes and semantic classes, appearing below a hardcoded list of classes, e.g. *COUNTRY_BY_NAME, PERSON_BY_FIRSTNAME* etc.
- **LEXCLASS_CONT** — a set of Boolean features, associated with the appearance in the hierarchy path of several manually selected semantic classes most correlated with NE labels in the training set.

The parser also provides an attribute named **NOUN_TYPE** that divides nouns into proper and common ones according to guesses in semantic hierarchy.

We presume that generalization of hierarchy paths plays an important role in maintaining the balance between preserving significant information and overfitting the classifier. We consider an "ideal" generalization such one that would choose for each word the most general semantic class, whose descendants in the hierarchy possess some kind of equivalence in terms of the problem being solved. However, this generalization still requires further research.

Feature combinations. Experiments demonstrate that some features (**NOUN_TYPE** and **WCASE**, **NOUN_TYPE** and **NSP**) show significantly higher results when they are combined into single feature. It is clear from intuition, that one feature with values like *(NOUN_TYPE=Common,WCASE=Lower), (NOUN_TYPE=Proper,WCASE=AllUpper)*, … possesses more information than two features valued *(Common, Proper, …)* and *(Lower, AllUpper, …)* in terms of a CRF model, based on a weighted sum of feature values. On the other hand, the values set size of a combination of several multiple-valued features may even exceed the number of words in the training set, what leads, obviously, to overfitting the data. For this reason, our classifier uses only two simple abovementioned combinations. Still, accurate feature combinations may yield higher results given enough research effort.

5. Experimental results

Table 2 shows the results achieved by our system on the CoNLL-2003 corpus and a comparison to other recent results. Results that overcome ours are highlighted with bold. It follows from the table that, with the exception of two feature sets of (Rosenfeld, Feldman, Fresko, 2006), higher results are achieved only with the help of external NE lists or document- and collection-level features.

Table 2. Comparing results to other systems

Research	Feature set	Devel. data	Test data
Our results		91.61	87.51
Tkachenko, Simanovsky, 2012	Local features	88.91	82.89
	Word + Wikipedia and DBPedia lists	85.21	78.16
	Word + Brown, Clark, LDA and phrase clusters of full Reuters corpus	90.87	87.00
	Full set (including lists)	**93.78**	**91.02**
Ratinov, Roth, 2009	(3): local features: token, word case, prefixes, suffixes, tokens in [-2..+2] window, word case in same window, two previous labels	89.25	83.65
	(3) + external lists	91.61	87.22
	(3) + Brown clusters of full Reuters corpus	90.85	86.82
	(3) + all external sources	**92.49**	**88.55**
	(3) + all non-local features	90.69	86.53
	(3) + all external + all non-local features	**93.50**	**90.57**
Rosenfeld, Feldman, Fresko, 2006	Lexical features, current and previous token		87.38
	Lexical features and combinations in [-2..0] token window		87.36
	(1): lexical features and combinations in [-2..+2] token window		**87.76**
	(2): (1) + suffixes and prefixes of previous token		**89.11**
	(2) + document- and collection-level classification results (non-local)		**90.72**
Chieu, Ng, 2003	Used training set only	91.60	86.84
	Training set and external NE lists	**93.01**	**88.31**

6. Exploring individual feature contributions

Exclusion of individual features from the common set (table 3) allows to evaluate the significance of each feature for classification. We can conclude that semantic features for current and previous token, dictionary lookup (**PART_OF**) and word case (**WCASE**, **SHAPE**) features are the most significant. It is also noticeable that semantic

features for current and previous token show a presence of correlation, meanwhile such features for the parent token are less informative.

Table 3 Impact of exclusion of individual features on the F-score; parentheses indicate area, in which the feature is calculated: t. — current token, par. — token parent, prev. — previous token

Excluded features	F- score (devel.data)	F-score (test data)
—	91.76	87.54
All semantic (t., par.)	91.41	87.23
SURFSL, SYNTF (t., par.)	91.65	87.18
NOUN_TYPE (t., prev., including all combinations)	91.45	86.97
POS (t., prev.)	91.35	86.97
All semantic (par., prev.)	90.81	86.51
WCASE (t., prev.)	91.66	86.51
SHAPE ([−2..0] window)	90.69	85.66
PART_OF (t.)	90.34	85.47
All semantic (t., prev.)	88.28	83.79
All semantic (t., par., prev.)	88.18	83.79

7. Classifier error exploration

We have explored about 100 classifier errors which show a number of widespread special cases leading to both parser and CRF-based classifier errors. Since correct parse results are statistically more common, the trained model assigns great weight to "deep" (semantic) features computed by the parser. In consequence, classifier makes a misclassification given an incorrect parse result, but some of these errors can be corrected due to presence of surface features like SHAPE and WCASE. Here are some examples of detected errors:

- Corpus text headers are given in capital letters. The parser often errs in this case; it is especially obvious when *JAPAN* receives "black varnish" semantic class and *CHINA* becomes "*porcelain*". It's evident that the parser also uses word cases to solve ambiguities.
- Name «*Bitar*» causes an error in the analysis of composites. The parser splits the name into two words which together mean "bi-resin" and the resulting features make correct classification impossible.
- In a number of cases proper names in semantic hierarchy are described in several different branches, leading to parse ambiguity. Examples of such names are *IRA* as a person name *Ira*, *Tom* as a person and a river name, *Moody* as a surname and a city name.
- Several errors made by a classifier with an absolutely correct parse tree were also found. These errors require a deeper analysis of prevalence of corresponding feature values in the training data.

8. Conclusions

We have demonstrated that the classifier built upon a CRF model and a feature set provided by the Compreno parser allows to achieve results comparable to the recent researches which do not use external NE lists and non-local features. Out of all such systems only one where a large local feature set in a wide window was considered (Rosenfeld, Feldman, Fresko, 2006) shows results that overcome ours. It lets assume that adding features based upon external NE lists and non-local document- and collection-level features to our classifier can allow reaching highest at the current moment results in NER field. However, the results of (Rosenfeld, Feldman, Fresko, 2006) also mean that greater attention should be given to features in linear and tree contexts of tokens.

Exploration of individual feature contributions shows that features related to token semantic class play the greatest role in NE identification. This fact demonstrates that universal inter-language semantic hierarchy is a rich source of information for NER solutions. Accordingly, a research of interlingual portability of a NER system based on semantic features might be of a certain interest.

Since the choice of features played an important role, in the course of the research we used different methods of automatic feature selection: individual feature exclusion, "greedy" feature inclusion, methods based on mutual information. This part is not included in the paper due to size limitations, but it is also an interesting subject for future research.

References

1. *Anisimovich K. V., Druzhkin K. J., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii «Dialog»], (pp. 90–103). Bekasovo.
2. *Bogdanov A. V.* (2012). Description of gapping in a system of automatic translation. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii «Dialog»], (pp. 61–70). Bekasovo.
3. *Chieu H. L., Ng H. T.* (2003). Named Entity Recognition with a Maximum Entropy Approach. Proceedings of CoNLL-2003, (pp. 160–163). Edmonton, Canada.
4. *Finkel J. R., Manning C. D.* (2009). Joint parsing and named entity recognition. NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, (pp. 326–334). Stroudsburg, PA, USA.
5. *Florian R., Ittycheriah A., Jing H., Zhang T.* (2003). Named Entity Recognition through Classifier Combination. Proceedings of CoNLL-2003, (pp. 168–171). Edmonton, Canada.

6. *Klein D., Smarr J., Nguyen H., Manning C. D.* (2003). Named Entity Recognition with Character-Level Models. Proceedings of CoNLL-2003, (pp. 180–183). Edmonton, Canada.
7. *McCallum A. K.* (2002). MALLET: A Machine Learning for Language Toolkit. available at: http://mallet.cs.umass.edu
8. *Nadeau D., Sekine S.* (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes , 30 (1), pp. 3–26.
9. *Nekhay I. V.* (2012). Application of n-grams and other letter- and word-level statistics to semantic classification of unknown proper nouns. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue» [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii «Dialog»], (pp. 477–489). Bekasovo.
10. *Ratinov L., Roth D.* (2009). Design challenges and misconceptions in named entity recognition. CoNLL '09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning, (pp. 147–155). Stroudsburg, PA, USA.
11. *Rosenfeld B., Feldman R., Fresko M.* (2006). A Systematic Cross-Comparison of Sequence Classifiers. Proceedings of the Sixth SIAM International Conference on Data Mining. Bethesda, MD, USA: SIAM.
12. *Tjong Kim Sang E. F., De Meulder F.* (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. CONLL ,03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 4, pp. 142–147. Stroudsburg, PA, USA.
13. *Tkachenko M., Simanovsky A.* (2012). Named entity recognition: Exploring features. KONVENS 2012, (pp. 118–127). Vienna.

ЭГОЦЕНТРИЧЕСКИЕ ЕДИНИЦЫ ЯЗЫКА И РЕЖИМЫ ИНТЕРПРЕТАЦИИ[1]

Падучева Е. В. (elena.paducheva@yandex.ru)

ВИНИТИ РАН, Москва, Россия

Доклад посвящен ЭГОЦЕНТРИЧЕСКИМ ЕДИНИЦАМ ЯЗЫКА — словам, грамматическим категориям и конструкциям, значение которых предполагает, в качестве одного из участников описываемой ситуации, ГОВОРЯЩЕГО. Так, во фразе *Иван едва ли вернется* адвербиал *едва ли* предполагает говорящего в роли субъекта сомнения, являясь, тем самым, эгоцентрической единицей, иначе — ЭГОЦЕНТРИКОМ. В работе две части. Первая посвящена общим свойствам эгоцентриков: обосновывается их деление на ПЕРВИЧНЫЕ (жесткие) и ВТОРИЧНЫЕ (мягкие), которое объясняет различное поведение эгоцентриков в НЕКАНОНИЧЕСКИХ КОММУНИКАТИВНЫХ СИТУАЦИЯХ — в нарративе и в гипотаксисе. Мягкие эгоцентрики свободно употребляются в неканонических режимах, меняя лишь ориентацию — подразумеваемым субъектом у них может быть не говорящий, а какое-то другое лицо. А жесткие эгоцентрики в неканонических режимах не употребляются или меняют значение. Особый раздел посвящен ДИСКУРСИВНОМУ режиму интерпретации. Во второй части рассматривается пример — наречие *однажды*, которое ведет себя как жесткий эгоцентрик дискурсивного режима.

Ключевые слова: эгоцентрики, говорящий, неканонические коммуникативные ситуации, жесткие эгоцентрики, мягкие эгоцентрики

EGOCENTRICALS AND THEIR REGISTERS OF INTERPRETATION

Paducheva E. V. (elena.paducheva@yandex.ru)

VINITI RAS, Moscow, Russia

Linguistic entities (words, grammatical categories, syntactic constructions) are called EGOCENTRICALS, if their semantics presupposes the SPEAKER as one of the participants in the situation described, cf., for example, *sejčas*, as in *On sejčas doma* ['he's now at home', the speaker is the holder of the moment of speech], *edva li* 'unlikely', as in *On edva li pridet* ['he's unlikely to come', the speaker is the subject of doubt], subjunctive mood, as in *Byla*

[1] Данная работа была выполнена при финансовой поддержке РГНФ, грант № 11-04-00488а «Акциональные классы и актантные типы предикатных имен: семантика, грамматика, словарь».

by sejčas vesna! ['if it were spring now!', the speaker is the subject of volition]. Only CANONICAL communicative situations can afford a sterling, i. e. full value, speaker — with the synchronous addressee, with the field of vision common to the speaker and the addressee, etc. In NON-CANONICAL communicative situations, such as NARRATIVE or HYPOTAXIS, when the speaker is not accessible as a performer of his/her presupposed role, and some substitute of the speaker comes into play, different egocentricals behave differently. Two types of egocentricals are discerned — SHIFTABLE (i. e. secondary) egocentricals, which can be used in all types of communicative situations, and HARD (i. e. primary) egocentricals, which stick to the canonical communicative situation, thus belonging to the so called MAIN CLAUSE PHENOMENA. One egocentrical is discussed in detail: the adverb *odnaždy* 'once upon a time'.

Key words: egocentricals, speaker, non-canonical communicative situations, hard egocentricals, shiftable egocentricals

1. Эгоцентрики и их свойства

1.1. Эгоцентрики: примеры

Эгоцентрическими являются языковые единицы (слова, граммемы, синтаксические конструкции), семантика которых предполагает, в канонической коммуникативной ситуации, в качестве одного из участников описываемой ситуации, говорящего. Так, в семантике вводного слова *едва ли* (*Иван едва ли придет*) говорящий выступает в роли субъекта сомнения, в семантике сослагательного наклонения — в роли субъекта желания (*Была бы сейчас весна*!), в семантике изъявительного наклонения — в роли гаранта эпистемического обязательства (*Иван вернулся*). Пример синтаксической конструкции с эгоцентрической семантикой — генитив отрицания в контексте локативного *быть*, который выражает наблюдаемое отсутствие, см. Падучева 2006. Эгоцентрическими являются, в частности, дейктические слова — такие, которые обозначают лицо, место или время через отсылку к говорящему (как *ты, здесь, сейчас, завтра*), и, разумеется, такие, которые просто обозначают говорящего, как *я, мы*.

Эгоцентрические языковые единицы называют сокращенно эгоцентрики. Термины egocentrical и egocentric particulars принадлежат Б. Расселлу.

В неканонических коммуникативных ситуациях разные эгоцентрики ведут себя по-разному. Разным типам коммуникативных ситуаций (разным с точки зрения доступности говорящего как фигуры, на которую опирается интерпретация эгоцентрика) соответствуют разные режимы интерпретации эгоцентриков:

- диалогический режим (полноценный говорящий, у которого есть синхронный слушающий; у говорящего и слушающего общее поле зрения; говорящий имеет право на жесты, доступные восприятию слушающего; возможные отступления от каноничности — отсутствие у говорящего и слушающего общего места или времени);
- нарративный режим (нет полноценного говорящего; его замещает повествователь или персонаж); различается традиционный нарратив (в том

числе — нарратив от первого лица) и свободный косвенный дискурс (см. обзор в Падучева 1996: 206), в котором широко используется несобственная прямая речь.
- гипотаксический режим (говорящего замещает подлежащее матричного предложения).

Режим интерпретации — это то же, что контекст употребления. Приведем пример интерпретации эгоцентриков в нарративном режиме. Начало рассказа Чехова «Учитель словесности» (1889):

(1) Послышался стук лошадиных копыт о бревенчатый пол; вывели из конюшни сначала вороного Графа Нулина, потом белого Великана, потом сестру его Майку. Всё это были превосходные и дорогие лошади.

Читатель ощущает наличие в ситуации какого-то закадрового сознания, которое далеко не сразу воплощается в персонаже, а сначала проявляет себя как заместитель того говорящего, на которого в речевом режиме был бы ориентирован смысл таких эгоцентрических слов, как *послышался* или *превосходный*.

1.2. Параметры, характеризующие эгоцентрическую единицу

Как уже говорилось, эгоцентриками могут быть слова, граммемы и синтаксические конструкции.

Изначально семантической сферой эгоцентрического считался только дейксис. В классической работе Якобсона — Jakobson 1957 дейксис был объединен с модальностью. Третья и четвертая сфера эгоцентрического — оценка и эвиденциальность. Наконец, есть и пятая сфера — коммуникативная структура (тема-рематическое членение), которая имеет дело с оппозицией известное/неизвестное, т. е. апеллирует к говорящему и слушающему.

Говорящий может занимать в толковании эгоцентрика разные позиции. В Падучева 1992 различаются следующие семантические роли говорящего:
- говорящий как субъект речи (как в слове *кстати* = ‘кстати сказать’ или в семантике императива),
- говорящий как субъект дейксиса, т. е. точка для отсчета места и времени (как в словах *вчера, здесь*),
- говорящий как субъект восприятия (как в слове *послышаться*),
- говорящий как субъект сознания — в частности, оценки (как в *едва ли*, *превосходный*).

В связи с режимами интерпретации существенно деление эгоцентриков на первичные и вторичные.

вторичные (иначе — несобственные) эгоцентрики допускают употребление не только в диалогическом режиме, но и в гипотаксисе и в нарративе, свободно заменяя ориентацию на говорящего ориентацией на субъект подчиняющего предложения (в гипотаксисе) или просто салиентный (в нарративе). Например, в предложении *Маша сказала, что едва ли успеет к семи* субъект

сомнения не говорящий, а Маша. Про вторичные эгоцентрики можно сказать, что они свободно подвергаются проекции, не меняя значения. Различаются гипотаксическая и нарративная проекция.

первичные эгоцентрики — это такие, которые употребляются только в диалогическом режиме и ориентируются только на говорящего. В нарративе и гипотаксисе они не употребляются или меняют значение. Пример — прош. время несов. вида: в нарративе (традиционном) оно не только ориентируется на повествователя, а не на говорящего, но и обозначает синхронность наст. моменту повествователя, а не предшествование — т. е. меняет значение. Первичные эгоцентрики могут допускать ориентацию на персонажа в нарративе, но не в традиционном, а в свободном косвенном дискурсе.

Английские терминологические соответствия: вторичные эгоцентрики — shiftable indexicals, первичные эгоцентрики — pure indexicals; ср. также main clause phenomena.

И еще один параметр — внутрифразовая vs. дискурсивная интерпретация (Плунгян 2008). Одни эгоцентрики имеют обе интерпретации, например, вид, см. Падучева 2008). Другие образуют пары. Ср., например, *вчера* и *накануне*: слово *вчера* имеет дейктическое значение и не предполагает специального текстового контекста; а *накануне* — это анафор, который отсылает к предтексту, т. е. является эгоцентриком с исключительно дискурсивной интерпретацией. Есть, кроме того, эгоцентрики с интродуктивной, т. е. катафорической функцией (об интродуктивных употреблениях см. Арутюнова 1976: 221). Это тоже эгоцентрики с дискурсивной семантикой, только они отсылают к последующему тексту. Таково, в частности, *однажды*, см. раздел 3.

Принципиальный вклад в проблематику эгоцентрии внесла статья Апресян 1986. Был предложен синтаксический тест, который

- позволяет обнаружить присутствие подразумеваемого говорящего в семантике языковой единицы;
- позволяет различить первичную и вторичную эгоцентричность.

Так, в семантике слова *показался* имеется встроенный говорящий, т. е. это слово эгоцентрическое; и это порождает аномалию в предложении (1а), где говорящий является субъектом восприятия самого себя. При этом слово *показался* вторичный эгоцентрик — оно свободно поддается проекции, например, гипотаксической, см. (1б), где аномалия пропадает, поскольку субъектом восприятия оказывается не говорящий, а субъект подчиняющего предложения.

(1) а. *На дороге *показался* я (ср. На дороге *показался* всадник);
б. Иван говорит, что именно в этот момент на дороге *показался* я.

Подразумеваемый субъект вторичного эгоцентрического слова был назван наблюдателем. Возникло противопоставление говорящего и наблюдателя.

В статье Булыгина 1982: 15 про глаголы типа *белеть, чернеть* говорится, что они «могут функционировать только в предикациях, описывающих конкретную, «актуальную» ситуацию, в которой находится (или в которую

помещает себя) говорящий»; в них усматривается «эффект соприсутствия» говорящего. Между тем, в Апресян 1986 эти глаголы — один из примеров, демонстрирующих встроенного наблюдателя. Компонент восприятия отмечен у них в словаре Ушакова. Пример (2) подтверждает наличие подразумеваемого наблюдателя, поскольку наречие *отчетливо* осмысленно только в применении к перцептивному компоненту в семантике глагола:

(2) Было около восьми часов вечера. За домами башня собора *отчетливо чернела* на червонной полосе зари. (Набоков. Возвращение Чорба)

Слова, предполагающие наблюдателя по Ю. Д. Апресяну (Апресян 1986), — это вторичные эгоцентрики с имплицитным говорящим в роли субъекта дейксиса (*вдалеке*) или восприятия (*показался*).

Была попытка называть наблюдателем то лицо, на которое ориентирован любой эгоцентрик, если он вторичный, — независимо от роли этого лица как участника ситуации, описываемой эгоцентрическим словом. Однако такое употребление не привилось. Так, в Апресян 2002 говорится, что в толкование глаголов интерпретации (типа *ошибаться, выпендриваться*) входит говорящий, а не наблюдатель, — хотя они являются вторичными эгоцентриками, т. е. поддаются гипотаксической проекции (ср. **Я ошибаюсь*, **Я выпендриваюсь* и нормальное *Он считает, что я ошибаюсь, выпендриваюсь*), а роль субъекта сознания очень близка к роли субъекта восприятия.

В Левонтина 2004 для слов *неожиданно, вдруг* предлагается толкование: ‘имеет место Р; говорящий или наблюдатель не ожидал, что будет Р или что Р произойдет именно в данный момент’. Эту дизъюнкцию следует, видимо, понимать так, что эти слова предполагают говорящего либо в роли субъекта восприятия, и тогда это ‘наблюдатель не ожидал’, либо в роли субъекта сознания, и тогда ‘говорящий не ожидал’.

Говорящий может принимать участие в такой ситуации, само существование которой осознается по его в ней участию, а языком выражено неявно. Так, в примере (3) слова *кончился* <лес> и *начались* <болота> предполагают идущего (или едущего). Говорящий является не только наблюдателем, но и прямым участником ситуации; обычно это перемещение (см. о движущемся наблюдателе в Апресян 1974: 161):

(3) За озером хвойный лес кончился и начались болота.

Ориентация эгоцентрика зависит не только от режима интерпретации, но и от модальности. Так, известно, что вопрос во многом переносит ориентацию эгоцентриков с говорящего на слушающего. Это вопросительная проекция (Падучева 1996: 268):

(4) Ну что, он так и не показался? [= ‘в твоем поле зрения’];

(5) Вкусно? [= ‘тебе’].

1.3. Краткая история вопроса

В самую краткую историю эгоцентрии должны войти: статья Jakobson 1957 (которая начинается со ссылки на книгу Волошинов 1930 о несобственной прямой речи); книга Успенский 1970 о точке зрения; статья Апресян 1986 (со ссылкой на Fillmore 1982).

За узловым для этой проблематики понятием «режим интерпретации» стоит Э. Бенвенист (1974), который ввел различие между «планом речи» (plan de discourse) и «планом повествования» (plan de récit). То, что Бенвенист называет «планом» <высказывания>, и есть тип коммуникативной ситуации, или режим интерпретации.

Термин каноническая коммуникативная ситуация — из Lyons 1977. Там же приводятся примеры неканонических ситуаций. Важным вкладом в проблематику является книга Ковтунова 1986, где есть специальный раздел, посвященный (неканонической) коммуникативной ситуации лирического стихотворения.

В. Успенский 2011, где рассматриваются только дейктические эгоцентрики, не используется понятие режим интерпретации, т. е. контекст употребления: различается «первичный дейксис, или дейксис в собственном смысле, и вторичный дейксис, при котором соотнесение с речевым актом осуществляется непрямым (опосредованным) образом». Очевидно, первичный дейксис — это то же, что употребление дейктических эгоцентриков в диалогическом режиме, а вторичный — в нарративном. Однако, скажем, на сферу модальности это словоупотребление распространить нельзя.

Ю. Д. Апресян первоначально тоже употреблял термины первичный vs. вторичный дейксис в значении диалогический vs. нарративный способ интерпретации: «Различаются первичный и вторичный дейксис. Первичный дейксис — это дейксис диалога, дейксис нормальной ситуации общения. <…> Вторичный дейксис, называемый также нарративным, <…> не связан непосредственно с речевой ситуацией. Это дейксис пересказа, в том числе художественного повествования.», Апресян 1986. Однако в Апресян 2004/2009: 516 противопоставляется, на примере возможности двоякого употребления слова *сейчас*, уже не первичный и вторичный дейксис, а диалогический и нарративный режим интерпретации.

Чтобы избежать этой опасной неоднозначности терминологии, я иногда заменяю эпитет «первичный» на эпитет жесткий эгоцентрик (т. е. ориентированный только на говорящего, hard egocentrical), и эпитет «вторичный» на эпитет мягкий эгоцентрик (т. е. допускающий разные виды проекции, shiftable egocentrical), ср. термин *жесткие дезигнаторы* в модальной логике. Существенно, что вторичные эгоцентрики могут употребляться, не меняя значения по существу, как в диалогическом («первичном» по Б. А. Успенскому), так и в нарративном или гипотаксическом («вторичных») контекстах.

Далее в докладе рассматривается роль режима интерпретации в семантике эгоцентрических единиц на примере местоименного наречия *однажды*.

2. Наречие *однажды* как показатель слабой определенности

Слово *однажды* не было до последнего времени предметом специального исследования. Принципиально важный шаг на пути постижения его семантики, синтактики и референции сделан в статье Иорданская, Мельчук 2013 (далее — И&М 2013), где обращено внимание на то, что *однажды* связано — не только по форме, но и семантически, — со словом *один* в значении слабой определенности. (См. об *один* в этом не счетном значении в Падучева 1985: 212–214; там же и о других показателях слабой определенности — *кое-какой* и *некоторый*.) Однако кое-что еще предстоит уточнить.

Слабая определенность — это жемчужина русской картины мира: в русском языке неопределенность разработана как нигде, см. Падучева 1996а. В английском языке с трудом различается specific indefiniteness и non-specific indefiniteness, причем только на семантической основе. Между тем в русском это формально выраженная оппозиция — есть серия местоимений неизвестности (или нерелевантной идентификации), на *-то*, и серия нереферентных местоимений, на *-нибудь*. А кроме того, есть третья серия, на *кое-*, слабоопределенные местоимения; к местоимениям на *кое-* примыкают другие показатели слабой определенности — слова *некоторые* и *один*. Слабая определенность — это полу-определенность, т. е. определенность для говорящего и неопределенность для слушающего. Наблюдение, сделанное в И&М 2013, превращает *однажды* из изолята в слово, для которого заготовлено место в системе референциальных противопоставлений.

Как и *один*, *однажды* употребляется не только как референциальный показатель, но и в значении счета ‘<имело место> один раз’ (*Я видел его лишь однажды*); это значение здесь остается вне рассмотрения.

2.1. ОДНАЖДЫ: попытки истолкования

В И&М 2013 фразы (1а) и (1б) обе признаются недопустимыми. Можно думать, фраза (1б) признана недопустимой справедливо, а фраза (1а) — нет:

(1) а. После этого Изабелла Купер[2] поселилась в одном из городов Среднего Запада, а затем перебралась в Лос-Анджелес, где однажды покончила жизнь самоубийством;
б. *Однажды Изабелла Купер покончила жизнь самоубийством.

Многое прояснится, если рассмотреть употребление *однажды* в контексте соответствующей ему коммуникативной ситуации.

Слово *один* — это первичный, т. е. жесткий, эгоцентрик: оно предполагает каноническую коммуникативную ситуацию, поскольку его смысл требует

[2] Речь идет о возлюбленной генерала МакАртура.

обращения не только к говорящему, но и к адресату. Так, в смысл сочетания *один человек* входят компоненты (i) ‘я знаю этого человека’ и (ii) ‘я исхожу из того, что ты его не знаешь’. Если предположение (ii) не оправдывается, говорящий терпит коммуникативную неудачу; так, в примере (2) (ср. Падучева 1985: 155) говорящий не ожидал, что слушатель тоже знает этого человека и имеет представление о том, что с ним было:

(2) — И тут *один человек* выбежал из толпы …
— Он не выбежал, его выпихнули.

Как все эгоцентрики, смысл которых отсылает не только к говорящему, но и к адресату, *один* неподчинимо. В частности, нецитируемо: если Зина сказала мне (3а), я понимаю, что Зина знает об этой женщине больше, чем мне говорит. Фразой (3б) я не совсем точно передаю слова Зины, поскольку в гипотаксическом контексте *один* не сохраняет своего значения слабой определенности (‘я знаю, а тебе не говорю’):

(3) а. Вася женился на *одной* китаянке,
б. Зина сказала мне, что Вася женился на *одной* китаянке.

Впрочем, в примере (4а) говорящий остается субъектом слабой определенности: он дает понять, что знаком с юношей; а в примере (4б) обнаруживается нечто подобное гипотаксической проекции — субъектом слабой определенности является денотат подлежащего в матричном предложении:

(4) а. Сверх того, дошло до моего сведения, что *один проезжий москвич, добрейший, впрочем, юноша*, мимоходом отозвался обо мне на вечере у губернатора как о человеке выдохшемся и пустом. [И. С. Тургенев. Гамлет Щигровского уезда (1849)]

б. Но он недоверчиво посмотрел на меня и сказал, что *один* немец-книголюб, который у него бывает запросто, обещал ему свое покровительство и просил его ни о чем не беспокоиться. [Н. Н. Берберова. Курсив мой (1960–1966)]

Так что употребления *один*, в которых слабая определенность утрачивается и переходит в неопределенность, как в (3б), характеризуют, скорее, небрежную речь.

Что же касается *однажды*, то оно, как мы увидим, довольно свободно употребляется в выветренном значении: есть несколько контекстов, где слабая определенность переходит в обычную неопределенность, т. е. в несущественность временной локализации события для говорящего, или даже в нереферентность временного показателя.

В отличие от *один*, слово *однажды* в своем первичном употреблении, дискурсивное: предложение с *однажды* требует продолжающего текста — так же,

как местоимение требует антецедента. Интродуктивность — это то, в чем проявляется слабая определенность при интерпретации дискурсивного эгоцентрика в нарративном режиме. Однако *однажды* допускает и не интродуктивное употребление (как и *один*, см. пример (3а)). Интродуктивность *однажды* зависит от его места в коммуникативной структуре: тематическое *однажды* интродуктивно по преимуществу, а в других позициях это может быть не так.

В И&М 2013 дается следующее толкование *однажды*: *однажды* (Q) означает, что время ситуации Q и сама Q «неидентифицируема для Адресата». Далее неидентифицируемость ситуации Q сводится к трем свойствам ее референта:

1) необязательность (= факультативность; например, самоубийство — в отличие от смерти),
2) повторимость (например, инфаркт — в отличие от самоубийства),
3) ординарность (например, посещение кинотеатра — в отличие от инфаркта).

Утверждается, что если ситуация Q не обладает хотя бы одним из этих свойств, употребить *однажды* в соответствующей фразе нельзя. Это попытка выразить прагматический аспект семантики слабой определенности, «неизвестность Адресату», через «объективные» семантические свойства ситуации.

Рассмотрим каждое из этих свойств. Начнем со свойства 2) повторимость. Утверждается, что ситуация, вводимая словом *однажды*, должна быть повторимой, так что неповторимость — это источник неадекватности предложения (1б). Например, предложение *Однажды Изабелла Купер заболела* нормально, поскольку Изабелла Купер могла болеть много раз.

Однако ситуация во фразе с *однажды* в примере (5) тоже неповторима, а между тем, фраза нормальна — она естественно вписывается в свой контекст[3]:

(5) *Однажды* Никита Михалков дней так за двадцать снял тихий фильм «Пять вечеров». Станислав Любшин, Людмила Гурченко, оттенки серого, неброскость одежд, мягкость интонаций. <…> Первый канал *на этой неделе* лихо разделался с ароматом того старого, «нерейтингового» произведения. [Комсомольская правда, 2004.09.15]

Здесь обнаруживается отличие *однажды* от *один*: *один* возможно только в контексте общего имени объекта (например: *<один> мой одноклассник, <один> американский президент, <одна> китаянка*), а *однажды* возможно не только в контексте общего имени ситуации, но и в контексте неповторимой ситуации (которая задана определенной дескрипцией).

---

[3] А именно, заключительная фраза этой главы, *В самом деле, уж рассветало: молодые люди допили свои рюмки и разъехались*, служит основой для гениального наблюдения: «В конце первой главы происходит открытое сошествие автора в изображаемый им мир» (Виноградов 1936: 107). Действительно, *в самом деле* — это диалогическая реакция. Сейчас мы бы сказали, что тут повествователь вступает в диалог с героем.

Конечно, *однажды* чаще употребляется в контексте повторимой ситуации, однако пример (5) показывает, что неповторимость ситуации не исключена. Так что сама по себе она не может быть источником неадекватности предложения (1б).

Беда предложения (1б) в другом: слово *однажды* занимает в нем тематическую позицию, и потому имеет интродуктивное употребление. Фраза с *однажды* должна вводить в рассмотрение ситуацию — ее время, место и участников. А тут фраза гласит, что главный участник перестал существовать. Сцена остается пустой: слово-катафор не может реализовать своей обязательной валентности.

Еще один пример, который приводится в И&М 2013 в подтверждение неповторимости события как препятствия для *однажды*.

(6) Однажды он погиб в горах, сорвавшись со скалы.

По замечанию Б. А. Успенского, *однажды* в контексте предикатов *погиб* или *умер* все-таки возможно, если дальше последует рассказ о посмертной судьбе героя, т. е. о его жизни в потустороннем мире. Однако есть и другие возможности совместить *однажды* с глаголом *умереть* — развитие текста может идти по линии связи не с героем, а, например, с введенным ранее местом действия. Так, (7) — вполне адекватный текст:

(7) Изабелла Купер увлекалась архитектурой. Переехав в Лос-Анжелес, она построила себе дом необыкновенной красоты. Однажды Изабелла Купер умерла. Ее дом в Лос-Анжелесе стал музеем.

Надо сказать, что *однажды* в примере (5) так уместно потому, что момент создания фильма, введенный в рассмотрение повествователем с помощью *однажды*, взаимодействует с возникающим далее другим моментом, *на этой неделе*, так что *однажды* в конце концов заполняет свою катафорическую валентность.

Свойство 3) ординарность из И&М 2013 (посещение кинотеатра — в отличие от инфаркта) вообще нельзя признать релевантным. Так, текст (8) ничуть не аномален, хотя инфаркт — событие вполне неординарное:

(8) Однажды у него случился инфаркт. Вызвали скорую помощь.

Свойство 1) необязательность (которым обладает самоубийство — в отличие от смерти), демонстрируется, в частности, на примере (9). Утверждается, что *однажды* не сочетается с обозначением обязательной (иначе — неизбежной, предвидимой) ситуации:

(9) *Однажды наступила зима.

Казалось бы, для фразы *Однажды наступила зима* трудно найти уместный контекст. Однако Яндекс дает 115 примеров на это сочетание, например:

Купил он её <машину> летом, и кроме маленького багажника
ничего особо не напрягало. Но вот *однажды* наступила зима…

Итак, представляет интерес только одно из трех свойств — повторимость. Вернемся к примеру (6). Вот контекст, в котором такое неповторимое событие, как гибель человека в горах, является нормальной сферой действия для *однажды*:

(10) Он переехал в Южную Америку, где однажды
погиб в горах, сорвавшись со скалы.

Дело в том, что предложение (6) аномально как автономное высказывание, в котором *однажды* занимает тематическую позицию, что заставляет воспринять предложение как интродуктивное, т. е. требующее продолжения. Между тем в гипотаксическом контексте смысл *однажды* выветривается. Так, для (10) *однажды* (Q) = ‘ситуация Q имела место в некий, неважно какой момент’.

Пропадает компонент ‘повествователь имеет в виду некоторое конкретное событие, про которое он знает больше того, что вытекает из его дескрипции’. Рассказ на этом может закончиться, повествователь больше ничего не хочет сказать адресату. Слабая определенность низводится до простой неопределенности. Снимается и условие повторимости.

Другие примеры употребления *однажды* в гипотаксической позиции, где оно имеет значение простой неопределенности ‘в какой-то момент’:

(11) есть прогнозы, которые реализовались именно потому, что *однажды* были сделаны. [В. Н. Комаров. Тайны пространства и времени (1995–2000)];

(12) И добровольный работник думать не думал, что *однажды* всё
куда-то денется. [Алексей Варламов. Купавна, 2000];

(13) Это значит, что ей там, в кино, просто стало неуютно, как
однажды стало неуютно в Америке. [«Домовой», 2002.12.04] .

Теперь понятно, почему *однажды*, недопустимое в предложении (1б), допустимо в (1а): попав в гипотаксический контекст, *однажды* перестает быть тематическим, а значит и интродуктивным. Из показателя слабой определенности оно превращается в показатель неопределенности обычной.

В И&М 2013 убедительно продемонстрированы некоторые свойства *однажды*, прямо вытекающие из семантики слабой определенности. Так, слабая определенность исключает употребление *однажды* в вопросе и в побуждении (для *один* это описано в Падучева 1985: 214). В самом деле, слабоопределенное *однажды* локализует во времени событие, которое для говорящего является конкретно-референтным, т. е. определенным: говорящий имеет в виду конкретное событие, реально имевшее место. А в вопросе и побуждении *однажды* не может иметь индивидного референта, поскольку

речь идет о виртуальном событии. Примеры неправильных употреблений (из И&М 2012, с сохранением нумерации) — вместо *однажды* тут надо было сказать *когда-нибудь*:

(31) a. *А ты отдыхал *однажды* в Египте?
b. *Отдохни *однажды* в Египте!

Впрочем, как отмечено в И&М 2013 (номер сохранен), в будущем времени *однажды* может быть употреблено — в выветренном значении ‘когда-нибудь’, т. е. для обозначения нереферентного момента времени:

(23) *Однажды* президентом этой организации станет азиат.

Другой пример из И&М 2013 (номер сохранен) показывает, что нереферентное *однажды* допустимо и при глаголе прош. времени — в сфере действия квантора общности, где оно имеет дистрибутивное значение: момент произнесения для каждого слова свой. При этом *однажды* можно заменить не на *когда-нибудь*, а на *когда-то*:

(29) Все слова на свете были *однажды* сказаны.

Об употреблении *-то* в значении *-нибудь* см. Падучева 1985: 219-220. Но про допустимость слабоопределенных местоимений в дистрибутивном контексте до сих пор не было известно.

Итак, *однажды* выражает слабую определенность и тем самым является первичным эгоцентриком. Одновременно оно является дискурсивным словом, так что семантика слабой определенности (как эпистемического неравенства говорящего и адресата) проявляется у него в интродуктивности, т. е. в обязательности продолжения, иначе — катафоричности. Однако интродуктивность свойственна только тематическому *однажды*. В гипотаксическом контексте, а также в будущем времени и в сфере действия некоторых операторов, *однажды* становится показателем обычной неопределенности или даже нереферентности временного показателя. Пример (7) показывает, что интродуктивность *однажды* может сниматься также нарративным контекстом.

2.2. О первой фразе «Пиковой дамы»

Известно, что с начальной фразой пушкинской «Пиковой дамы» —

(1) Однажды играли в карты у конногвардейца Нарумова

— что-то не в порядке (Падучева 1995). Текстов с начальным предложением такой структуры не встречается (если не считать начала жуткого рассказа Шаламова «На представку», с явной аллюзией именно к этой пушкинской фразе).

В Грамматике 1954 эта фраза фигурирует как пример неопределенно-личного предложения (НЛП). Однако она демонстративно нарушает какие-то нормы НЛП. Чтобы сделать эту фразу обычным НЛП, начинающим текст, достаточно было бы, например, изменить порядок слов — передвинув глагол в рематическую позицию, как в (2а), или хотя бы убрать *однажды*, как в (2б):

(2) а. Однажды у конногвардейца Нарумова *играли* в карты.
б. Играли в карты у конногвардейца Нарумова[4].

Не пойдя ни по одному из этих простых путей, Пушкин задал нам загадку, которую попытался разгадать В. В. Виноградов (Виноградов 1936): он объясняет эту фразу (и следующий за ней текст) присутствием повествователя, который как бы причисляет себя к тому же кругу, что играющие, — иначе говоря, повествователя- рассказчика:

> Повествователь в «Пиковой Даме», сперва не обозначенный ни именем, ни местоимениями, вступает в круг игроков как один из представителей светского общества. Он погружен в мир своих героев. Уже начало повести: «Однажды играли в карты у конногвардейца Нарумова. Долгая зимняя ночь прошла незаметно; сели ужинать в пятом часу утра» — повторением неопределенно личных форм — и г р а л и , с е л и у ж и н а т ь — создает иллюзию включенности автора в это общество. К такому пониманию побуждает и порядок слов, в котором выражается не объективная отрешенность рассказчика от воспроизводимых событий, а его субъективное сопереживание их, активное в них участие. Повествовательный акцент на наречии — н е з а м е т н о , поставленном позади глагола («прошла незаметно» — в контраст с определениями ночи — «долгая зимняя»); выдвинутая к началу глагольная форма — и г р а л и («однажды играли в карты»; ср. объективное констатирование факта при такой расстановке слов: «однажды у конногвардейца Нарумова играли в карты»); отсутствие указания на «лицо», на субъект действия при переходе к новой повествовательной теме — «с е л и ужинать», внушающее мысль о слиянии автора с обществом (т. е. почти рождающее образ — м ы) — все это полно субъективной заинтересованности. Читатель настраивается рассматривать рассказчика как участника событий. <…> Эта близость повествователя к изображаемому миру, его «имманентность» воспроизводимой действительности легко допускают драматизацию действия. (Виноградов 1936: 106)

Эта трактовка фразы (1), при том, что она многократно повторялась на разные лады, см., например, Онипенко 2001, Сидорова 2011, Никитина 2012, лингвистически не обоснована. Что верно — это что повествователь «Пиковой дамы» дает о себе знать в конце главы I, и это открытие Виноградова — краеугольный

---

[4] В статье Разлогова 2012 исследуется восемнадцать переводов «Пиковой дамы» на французский язык: *однажды* в них к а к п р а в и л о остается без перевода.

камень, заложенный им в теорию нарратива[5]. Но этот повествователь не рассказчик: он не входит в мир, о котором повествует. Взять хотя бы то, что он называет игроков в карты «молодые люди».

Что же касается предложения (1), Виноградов прав в том смысле, что если подставить в нее *мы*, то оно станет синтаксически безупречным. Он предлагает, однако, считать местоимение 1 лица мн.числа *мы* подразумеваемым субъектом НЛП. Между тем референциальный анализ подразумеваемых субъектов НЛП в русском нарративе показывает, что это абсолютно исключено.

Во фразе (1) две загадки, одна связана с подразумеваемым субъектом бесподлежащного глагола 3 лица, другая — с *однажды*. Рассмотрим вначале фразу (1) без *однажды*, т. е. предложение (2б).

Чтобы говорить о референциальных характеристиках субъекта и порядке слов в НЛП, следует различить НЛП с конкретно-референтным подразумеваемым субъектом (и с актуальным значением вида глагола), как в *В большом доме напротив играли на рояле*, и с обобщенным, как в *Не очень-то нынче старших уважают*. В Грамматике 1954 даже относят эти последние не к НЛП, а к обобщенно-личным предложениям (ОЛП). На самом деле, они, конечно, тоже НЛП, но между этими двумя референциальными типами НЛП есть существенные различия.

Про НЛП с обобщенным субъектом (и узуальным значением вида глагола) еще можно обсуждать, включен ли говорящий в множество референции. Что же касается НЛП с актуальным видовым значением глагола, то у него лицо подразумеваемого субъекта т о л ь к о третье. Более того, статус этого субъекта однозначно неопределенный. Субъект действия может быть неизвестен говорящему (*У меня украли мобильник*) или говорящий не считает нужным его называть, так как он неважен или очевиден (*Тебя сегодня спрашивали?*). В Грамматике 1980: 356 говорится, что «в условиях конситуации неопределенность субъекта может сниматься и субъект может мыслиться говорящим как вполне определенный, данный (например, в сообщении о приходе того, кого ожидают: *Пришли*)». Едва ли, однако, это *Пришли* следует вообще трактовать как НЛП — скорее тут просто эллипсис. А в предложении *К вам пришли* подразумеваемый субъект является грамматически неопределенным, независимо от степени знакомства говорящего с пришедшим.

Подразумеваемый субъект может быть также слабоопределенным — как, например, в предложении *Не беспокойтесь, меня проводят*, когда участник ситуации известен говорящему, но не сообщается адресату (Падучева 2012). Именно этот вариант неопределенности представлен в НЛП из начального фрагмента «Доктора Живаго» (пример приводится в Сидорова 2011, с другим анализом):

---

[5] А именно, заключительная фраза этой главы, *В самом деле, уж рассветало: молодые люди допили свои рюмки и разъехались*, служит основой для гениального наблюдения: «В конце первой главы происходит открытое сошествие автора в изображаемый им мир» (Виноградов 1936: 107). Действительно, *в самом деле* — это диалогическая реакция. Сейчас мы бы сказали, что тут повествователь вступает в диалог с героем.

Шли и шли и пели «Вечную память», и когда *останавливались*, *казалось*, что ее по залаженному продолжают
петь ноги, лошади, дуновения ветра.

Здесь в первой части предложения субъект вводится в рассмотрение, а во второй бесподлежащные предикаты подразумевают тот же субъект: *останавливались* те же люди, что *шли и пели*; и *казалось*, скорее всего, им же — хотя возможен и внешний наблюдатель, особенно что касается дуновений ветра.

Дальше идет текст:

Прохожие пропускали шествие, считали венки, крестились.
Любопытные входили в процессию, спрашивали:
«Кого *хоронят*?» Им *отвечали*: «Живаго».

У *хоронят* субъект идентифицируется однозначно, и не упоминается потому, что очевиден — предложение можно даже трактовать как эллиптичное, а не неопределенно-личное; скорее, как неопределенно-личное, потому что субъект скорее неважен, чем опущен. А *отвечали*, скорее всего, те, кто *шли и пели*; точнее, кто-то из них.

Особенность употребления НЛ-конструкции в первой фразе «Доктора Живаго» — в том, что она не используется как средство выражения «отчуждения», как это свойственно НД-конструкции в других случаях, см. Булыгина, Шмелев 1997: 345–346. Напротив, ее подразумеваемый субъект является фокусом эмпатии повествователя и используется как антецедент для последующей нулевой анафоры.

По аналогии с этим примером можно трактовать и связи между подразумеваемыми субъектами в начальном фрагменте «Пиковой дамы». Повторим:

Однажды играли в карты у конногвардейца Нарумова. Долгая
зимняя ночь прошла *незаметно*; *сели ужинать* в пятом часу утра.

Первая фраза предполагает некую группу лиц, к которым отсылает подразумеваемый субъект НЛ-сказуемого *играли*. Далее имеется в виду: *незаметно* — для тех, которые играли; *сели ужинать* — они же. Никакого рассказчика не нужно для обеспечения нулевой анафоры, выражающей кореферентность подразумеваемых субъектов у *незаметно* и *сели*.

Итак, первые два аргумента в пользу *мы* (т. е. в пользу присутствия рассказчика среди играющих), которое Виноградов прочит в подразумеваемые субъекты для *незаметно* и *сели*, отпадают. Остается третий аргумент — порядок слов (напомним, что, по Виноградову, в порядке слов «…выражается не объективная отрешенность рассказчика от воспроизводимых событий, а его субъективное сопереживание их, активное в них участие»). Виноградов отмечает необычность выдвинутой к началу глагольной формы *играли*, справедливо полагая, что объективное констатирование факта дало бы другую расстановку слов: *Однажды у конногвардейца Нарумова играли в карты*. Этот аргумент

в пользу повествователя-участника более серьезный, поскольку, как уже говорилось, именно порядок слов делает фразу странным началом.

Как отмечено в Падучева 2012, НЛ-предложениям с конкретно-референтным субъектом свойственно следующее ограничение на коммуникативную структуру: глагол должен находиться в рематической позиции — что хорошо согласуется с неопределенностью подразумеваемого субъекта НЛП. Так, предложение (3) (которое в Грамматике 1980: 357 трактуется как НЛП), не может быть понято как конкретно-референтное НЛП:

(3) В Двине *купались* ночью. (Ю. Казаков)

Вот его контекст:

И вот несколько дней назад на пароходе «Юшар»
мы пришли в Мезень, и ходу было всего два дня от Архангельска
… Весь июль стояла на Севере противоестественная жара.
В Двине *купались* ночью. (Ю. Казаков. Северный дневник)

Предложение (3) может быть понято, в этом контексте, либо как ‘мы купались’, т. е. как неполное; либо как ‘люди на Севере купались в этом июле’, т. е. как НЛП с глаголом в узуальном значении, в котором ограничения на порядок слов нет. (Хотя предложение *В Двине ночью купались* само по себе может быть понято как ‘кто-то купался’.)

Тематическая позиция глагола в НЛП с конкретно-референтным (не обобщенным) значением, не полностью исключена, но она требует специального контекста; так, *Кричали далеко* (пример из Грамматики 1954) вполне допустимое неопределенно-личное предложение, но оно предполагает, что о крике уже шла речь.

Есть другая возможность оправдать порядок слов с тематическим глаголом в НЛ-конструкции — когда читатель предполагается уже осведомленным, о ком речь. Но нормально такая фраза может быть только продолжением истории, а не ее началом. К этой возможности мы еще вернемся. Пока важно, что а) отвергнуто *мы* как подразумеваемый субъект у *незаметно* и *сели* во второй фразе и б) показана необязательность *мы* для оправдания необычного порядка слов в первой.

Теперь обратимся к *однажды*. Как видно из сопоставления (1) и (2б), контекст *однажды* усугубляет требования к коммуникативной структуре неопределенно-личного предложения, которое начинает текст: НЛП без *однажды* может быть и нерасчлененной ремой, как (2).

Итак, виноградовский повествователь-рассказчик, принимающий участие в событиях, не подтверждается лингвистически. Он опровергается и текстологическим анализом. Согласно комментариям Б. В. Томашевского, «Пиковая дама» первоначально задумывалась как повествование от 1 лица (ср. начало чернового наброска к ранней редакции повести: *Года четыре тому назад собралось нас в Петербурге несколько молодых людей, связанных между собою обстоятельствами*). Оказывается, Пушкин в ходе работы над «Пиковой дамой» сознательно отказался от повествователя-рассказчика, т. е. от перволичного повествования,

и перешел к традиционному нарративу от третьего лица. (Как известно, Виноградов писал свою статью в ссылке и не имел доступа к черновикам.)

Против виноградовского повествователя-рассказчика свидетельствует также и то, что он полностью пропадает в последующей части повести. Что же касается диегетического повествователя, то он хоть и ненавязчиво, но присутствует — речь идет не только о *в самом деле*, но и о других вводных словах. Один раз он даже назван в первом лице — как бы в диалоге с читателем:

> <…> это случилось <…> за неделю перед той
> сценой, на которой *мы* остановились.

Отвергнув рассказчика, мы можем теперь предложить другую разгадку для порядка слов в первой фразе «Пиковой дамы». Эта фраза не является первой — она связана с эпиграфом (за эту идею — благодарность Л. Н. Иорданской; ср. также Онипенко 2001): *А в ненастные дни Собирались они Часто.* <…> Так что подразумеваемый субъект третьего лица у *играли* отсылает к этим ‘они’ и является данным. Это и предопределяет для первой фразы ее коммуникативную структуру с тематическим сказуемым. Но тогда перед нами специальный неканонический тип нарратива, в котором первая фраза синтаксически зависит от эпиграфа.

Заключение

Рассмотренные примеры показывают, что смысл эгоцентрических слов и категорий не исчерпывается толкованием. Обращение к режиму интерпретации открывает новые аспекты их семантики[6].

Литература

1. *Апресян Ю. Д.* (1986) Дейксис в лексике и грамматике и наивная модель мира // Семиотика и информатика. Вып. 28. М.
2. *Апресян Ю. Д.* (2004/2009) Понятийный аппарат системной лексикографии //Ю. Д. Апресян. Исследования по семантике и лексикографии. Т. I. М.
3. *Арутюнова Н. Д.* (1976) Предложение и его смысл. М.: Наука.
4. *Бенвенист Э.* (1974) Общая лингвистика. М.: Прогресс.
5. *Булыгина Т. В.* (1982) К построению типологии предикатов в русском языке //О. Н. Селиверстова (отв. ред.). Семантические типы предикатов. М.: Наука.
6. *Булыгина Т. В., Шмелев А. Д.* (1997) Языковая концептуализация мира (на материале русской грамматики). М.: Языки русской культуры.
7. *Виноградов В. В.* (1936) Стиль «Пиковой дамы» //Временник Пушкинской комиссии. 2. М.-Л.

[6] Автор благодарен анонимным рецензентам Диалога-2013 за проницательные замечания.

8. *Волошинов В. Н.* (1930) Марксизм и философия языка. Л.: Прибой.
9. *Иорданская Л. Н., Мельчук И. А.* (2013) Наречие *однажды*: неопределенный временной спецификатор. Вопросы языкознания, № 1, 22–37.
10. *Ковтунова И. И.* (1986) Поэтический синтаксис. М.: Наука.
11. *Левонтина И. Б.* (2004) неожиданно, вдруг //Новый объяснительный словарь синонимов русского языка / Под общ. рук. акад. Ю.Д.Апресяна. 2-е изд. М. — Вена.
12. *Никитина Е. Н.* (2012) Еще раз о деепричастиях в неопределенно-личных предложениях //Русский язык в научном освещении. № 24, 23–41.
13. *Онипенко Н. К.* (2001) Теория коммуникативной грамматики и проблема системного описания русского синтаксиса //Русский язык в научном освещении. № 2, 107–121.
14. *Падучева Е. В.* (1985) Высказывание и его соотнесенность с действительностью. М.: Наука.
15. *Падучева Е. В.* (1995) В.В.Виноградов и наука о языке художественной прозы. Известия ОЛЯ. Серия литературы и языка, т. 54, № 3, с. 39–48.
16. *Падучева Е. В.* (1996) Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. М.: Языки русской культуры.
17. *Падучева* 1996а — Неопределенность как семантическая доминанта русской языковой картины мира. Problemi di morphosintassi delle lingue slave, v. 5. Determinatezza e indeterminatezza nelle lingue slave. Padova: Unipress, 1996, 163–186. http://lexicograph.ruslang.ru/TextPdf1/dominanta1_1996.pdf
18. *Падучева Е. В.* (2006) Родительный отрицания и проблема единства дейктического центра высказывания. Известия РАН. Серия литературы и языка, № 4, 3–10.
19. *Падучева Е. В.* (2008) Дискурсивные слова и категории: режимы интерпретации. //Исследования по теории грамматики, вып. 4. Грамматические категории в дискурсе. М.: 56–86.
20. *Падучева Е. В.* (2012) Неопределенно-личное предложение и его подразумеваемый субъект. Вопросы языкознания, №1, 27–41.
21. *Плунгян В. А.* (2008) Дискурс и грамматика. //Исследования по теории грамматики, вып. 4. Грамматические категории в дискурсе. М., 7–34.
22. *Разлогова Е. Э.* (2012) «Пиковая дама» в зеркале французских переводов. Вопросы языкознания, № 6, 66–92.
23. *Успенский Б. А.* (1970) Поэтика композиции. М.: Искусство.
24. *Успенский Б. А.* (2011) Дейксис и вторичный семиозис в языке. Вопросы языкознания, № 2, 4–30.
25. *Fillmore Ch. J.* (1982) Towards a descriptive framework for spatial deixis. — In: Speech, place and action. Ed. R. J. Jarvel la, W. Klein. Chichester etc.: J. Wiley and sons, 31–60.
26. *Jakobson R.* (1957) Shifters, verbal categories and the Russian verb. Cambridge: Mass. (Русский перевод: Р. Якобсон. Шифтеры, глагольные категории и русский глагол //О. Г. Ревзина (отв. ред.). Принципы типологического анализа языков различного строя. М.: Наука, 1972).
27. *Lyons J.* (1977) Semantics. V. 1–2. Cambridge: Univ. Press.

АВТОМАТИЧЕСКОЕ ИСПРАВЛЕНИЕ ОПЕЧАТОК В ПОИСКОВЫХ ЗАПРОСАХ БЕЗ УЧЕТА КОНТЕКСТА

Панина М. Ф. (mar-fed@yandex-team.ru),
Байтин А. В. (baytin@yandex-team.ru),
Галинская И. Е. (galinskaya@yandex-team.ru)

Яндекс, Москва, Россия

Анализируя ошибки в поисковых запросах нетрудно заметить, что большая часть из них имеет однозначное исправление, не зависящее от словарного окружения, и может быть исправлена в автоматическом режиме. В данной работе мы попытались выделить классы ошибок, которые можно исправлять автоматически, определить долю контекстно-независимых исправлений, и для выбранного множества ошибок разработать классификатор, позволяющий разделить исправления на надежные (пригодные для автоматической замены) и малонадежные (пригодные только для подсказки).

В качестве кандидатов для исправлений были использованы подсказки поискового спелл-чекера, знакомые пользователям поисковых систем по сообщению «Возможно, вы имели в виду...». Для обучения классификатора были использованы лексические и статистические признаки словарного (бесконтекстного) уровня.

Проведенные эксперименты показали высокую эффективность признаков и возможность настройки классификатора на заданный уровень точности. Применение предложенного метода автокоррекции тривиальных опечаток может значительно повысить качество исправления ошибок в поисковых запросах.

Ключевые слова: исправление опечаток в поисковых запросах, автоисправление опечаток, машинное обучение, оценка надежности, контекстно-независимые ошибки

CONTEXT-INDEPENDENT AUTOCORRECTION OF QUERY SPELLING ERRORS

Panina M. F. (mar-fed@yandex-team.ru),
Baytin A. V. (baytin@yandex-team.ru),
Galinskaya I. E. (galinskaya@yandex-team.ru)

Yandex, Moscow, Russia

While analyzing errors in the search queries, it is easy to notice that the most part of query spelling errors are trivial typos. Such errors usually do not depend on the surrounding words and their correction can be done in the automatic mode. In this work we tried to define a class of query spelling errors that can be corrected automatically. For the selected class we developed a classifier dividing corrections into reliable (suitable for automatic query spelling correction) and low-reliable (suitable only for the query spelling suggestion). As candidates for autocorrections we used query speller suggestions familiar to the users of search engines by «Did you mean...» function. For the classifier training we used typical lexical and statistical features. The experiments showed high performance of the word-level features and the ability to configure the classifier for a given level of accuracy. The application of the proposed method of trivial typo correction can significantly improve the quality of the query spelling errors correction.

Key words: query spelling correction, autocorrection, machine learning, confidence estimation, spellchecker

1. Введение

Анализ логов современных поисковых систем показывает, что пользователи делают ошибки в 10–15% поисковых запросов [8,5]. Поскольку поиск по искаженному запросу обычно приводит к нерелевантной выдаче и негативно влияет на общее качестве поиска, поисковые системы уделяют большое внимание проблеме исправления опечаток в запросе и стараются исправлять их на всех этапах поискового процесса — начиная с корректировки запроса во время набора и заканчивая переформулировками во время показа результатов поиска.

Все способы исправления запросов в той или иной форме используют поисковый спелл-чекер — программу, которая находит в запросе ошибки и предлагает для них исправления. В данной работе мы рассмотрим две наиболее распространенные функции коррекции запросов — подсказку и автозамену.

Подсказка (Рис.1) представляет собой сообщение *(«Быть может, вы искали: <…>»)* и ссылку (*«качок»*), по которой можно перейти на страницу с поисковой выдачей по исправленному запросу. Подсказка полезна в тех случаях, когда оригинальное написание (*качек)* является допустимым, но в то же время совпадает с опечаткой другого, более употребимого в данном контексте слова (*качок*); или когда исправление неоднозначно (для *альтар вики* исправлением может быть и *алтарь вики,* и *альтаир вики*). Подсказка работает в интерактивном режиме и является консервативным способом коррекции запроса (показывается вместе с выдачей по оригинальному запросу и не искажает результатов поиска). Недостатком подсказки является необходимость совершать дополнительные действия (кликать по ссылке). Кроме того, пользователи часто не замечают подсказку или не доверяют ей, упуская возможность увидеть более релевантную поисковую выдачу.

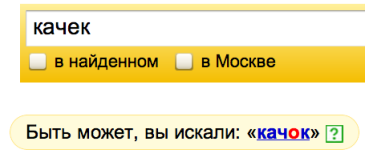


Рис. 1: Пример *подсказки*

В случае автозамены (Рис.2) пользователю показывается поисковая выдача сразу по исправленному запросу. Под поисковой строкой помещается сообщение об автоисправлении («*В запросе <…> была исправлена опечатка*»). Функция работает в автоматическом режиме и экономит время, которое приходилось бы тратить на ручное исправление ошибки или клик на подсказку (см. выше). На случай неверного автоисправления предусмотрена ссылка на страницу с результатами поиска по оригинальному написанию («*диаллог 2013*»). Следует заметить, что каждое неверное автоматическое изменение запроса сильно раздражает пользователя [15], поэтому требование к точности автозамен намного выше, чем к точности подсказок.

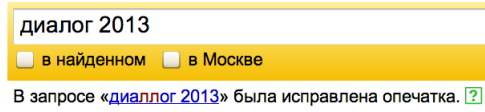


Рис. 2: Пример *автозамены*

Обнаружив в запросе ошибку, поисковая система оказывается перед выбором — какую из функций коррекции использовать. Обе функции оптимизируют полноту исправлений, однако пороги точности у подсказок и автозамен существенно различаются. Поэтому проблему выбора обычно решает не спелл-чекер, а специальный модуль (классификатор надежности), разделяющий исправления на надежные (автозамены) и ненадежные (подсказки).

Анализируя ошибки в запросах, можно заметить, что в основном они являются тривиальными опечатками и имеют очевидное исправление, не зависящее от словарного окружения. Это наблюдение приводит к идее об исправлении тривиальных ошибок в автоматическом режиме без учета информации о соседних словах. Цель данной работы состоит в оценке возможностей и разработке метода, позволяющего повысить эффективность автоматического исправления для этого класса словарных ошибок.

На фоне большого количества работ, посвященных исправлению ошибок, работ по автоисправлению относительно немного [2, 3, 9, 14]. Наиболее интересной является работа [14], посвященная построению системы для автокоррекции обычных текстов. Автоматическое исправление ошибок работает по трехуровневой модели: 1) для каждого слова спелл-чекер генерирует отранжированный список подсказок; 2) классификатор надежности

подсказки определяет является ли слово ошибкой; 3) классификатор надежности автоисправлений решает, можно ли исправить ошибку автоматически. Мы применили аналогичный подход и большую часть признаков, использованных в [14], однако надо отметить, что исправление ошибок в запросах и текстах имеют ряд существенных отличий: а) частота опечаток и степень искажения запросов гораздо выше; б) значительную долю слов в запросах составляют новые слова с неустоявшимся написанием; в) запросы в среднем состоят из 2–3 слов, поэтому для их исправления невозможно использовать широкий контекст.

Несмотря на то, что все современные поисковые системы используют функцию автозамены, нам не удалось найти ни одной статьи, посвященной непосредственно автоматическому исправлению поисковых запросов. В то же время, существуют интересные работы в области машинного перевода, где есть похожая на автозамену задача оценки качества перевода. Для того, чтобы определить степень готовности машинного перевода к ручному постредактированию или к публикации, необходимо автоматически оценивать качество перевода, имея в наличии только предложение и его машинный перевод. В работах [12,13] для решения этой задачи применяются методы машинного обучения, используются разнообразные лексические и статистические признаки. В ранних работах [12] использовался бинарный классификатор, в дальнейшем авторы перешли на пятибалльную шкалу оценки качества [13]. Важно отметить, что задача оценки качества решается независимо от задачи машинного перевода.

В данной работе для решения задачи разделения исправлений на «надежные» (автозамены) и «ненадежные» (подсказки) мы используем бинарный классификатор, подбираем для него словарные (бесконтекстные) признаки и проводим эксперимент на случайной выборке запросов.

2. Ошибки в запросах пользователей

Прежде чем приступить к разработке метода оценки надежности исправлений, необходимо произвести типизацию ошибок и их исправлений в поисковых запросах [1, 7, 8]. Поскольку ошибки каждого типа обладают особенностями и исправляются с неодинаковым качеством, для классификации надежности каждого типа нужен свой набор признаков и свои настройки классификатора.

2.1. Данные для исследования

Из дневного лога поисковой системы Яндекс случайным образом были выбраны 10 000 запросов. Из них 9242 оказались уникальными. Двум асессорам была поставлена задача найти среди них запросы с ошибками, предложить вариант исправления и определить язык запроса. После

завершения разметки каждому ошибочному запросу Q и его исправлению C автоматически, с помощью алгоритма выравнивания Левенштейна [11], был сопоставлен набор пар $\{q_i \rightarrow c_i\}$, $i = 1...m$, где $q_i \subset Q$, $c_i \subset C$ — слова запроса и исправления, m — количество опечаток в запросе. Например, для пары запрос-исправление *rfr сделатьсуши дома → как сделать суши дома* количество опечаток m=2, массив пар слово-исправление: *{rfr →как; сделатьсуши → сделать суши}*.

В наборе было найдено 1132 запроса с ошибками (12,2 %). Распределения запросов по количеству ошибок и по языку запросов представлены в табл. 1,2:

Таблица 1. Распределение количества ошибок в запросах с ошибками

Кол-во ошибок	%
1	83,6
2	11,7
>2	4,8

Таблица 2. Распределение запросов по языкам

Язык запроса	%
Русский	80,8
Английский	7,2
Русский + Английский	8,2
другие	3,8

Из Таблиц 1,2 видно, что большая часть запросов с ошибками (83,6%) содержит всего одну ошибку и почти все запросы задаются на русском или английском языках (96,2%).

2.2. Типы ошибок

Ошибки, встречающиеся в поисковых запросах, можно отнести к одному из четырех классов:

- Ошибки в отдельных словах — пропуск, вставка, перестановка и замена букв в словах (напр. *сборик → сбор**н**ик, стать**т**я → статья, ко**пм**ьютер → компьютер, модер**м** → модерн)*
- Ошибки слитно-раздельного написания — пропуск или вставка пробела между словами (напр. *томхарди → том харди, т акси → такси*)
- Неверная раскладка клавиатуры — использование английской раскладки для набора русских слов или наоборот (напр. *сфеы → cats, rjns → коты*)
- Транслитерация — использование латинского алфавита для набора русских слов или наоборот (напр. *uragan →ураган, констракшн → construction*)

Стоит заметить, что, хотя неверная раскладка и транслитерация формально ошибками правописания не являются, мы относим их к искаженным написаниям, негативно влияющим на качество поисковой выдачи.

Таблица 3. Типы ошибок в запросах

Тип ошибки	%
Ошибки в отдельных словах	63,7
Ошибки слитно-раздельного написания	16,9
Неверная раскладка клавиатуры	9,7
Транслитерация	1,3
Смешанные ошибки	8,3

В данной работе мы будем рассматривать самый большой класс ошибок (63,7%) — «ошибки в отдельных словах». Для простоты изложения в дальнейшем будем называть такие ошибки словарными или просто опечатками.

2.3. Зависимость исправлений от контекста

Для того чтобы проверить предположение о том, что большая часть словарных ошибок в запросах являются тривиальными, т. е. имеющими однозначное исправление независимо от контекста, мы попытались оценить влияние словарного окружения на исправление ошибок и провели следующий эксперимент. Из набора (гл. 2.1) взяли все запросы со словарными ошибками (620 запросов), выделили из них все пары слово-исправление (754 пары), отфильтровали пары-орфоварианты[1] и пары-словоформы[2], из оставшихся 714 пар взяли слова и поручили аналитику исправить в них ошибки, не глядя на контекст этих слов (текст запроса). Полученные результаты сравнили с исправлениями асессоров, сделанными с учетом контекста запроса (гл. 2.1). Если исправления совпадали, мы считали пару ошибка-исправление контекстно-независимой (например, аналитик: *рюкзк* → ***рюкзак;*** асессор*: рюкзк* → ***рюкзак***). Если исправления не совпадали, мы считали пару ошибка-исправление контекстно-зависимой (например, аналитик: *крепк* → ***скрепка;*** асессор: *крепк* → ***крепко***).

Контекстно-независимых исправлений оказалось 74%, что подтвердило наше предположение об однозначности исправлений для большинства словарных ошибок. Примеры исправлений тривиальных ошибок: *актриссa* → *актриса, спа**ис**бо* → *спасибо, бук**ен**естический* → *букинистический*.

---

[1] Орфоварианты — слова, имеющие одно и то же значение, но различное, как правило, очень близкое, написание (напр. *кэйтлин* → *кейтлин*).

[2] Неправильная словоформа — слово с опечаткой и его исправление являются разными формами одного и того же слова (напр. *лето* → *летом*).

Среди контекстно-зависимых исправлений большую часть составляют ошибки в коротких словах, допускающие разные, зависящие от контекста, исправления. Например ***сво*** *законов* → ***свод*** *законов,* и ***сво*** *игра* → ***своя*** *игра.* Примеры других контекстно-зависимых ошибок представлены в Таблице 4.

Таблица 4. Примеры контекстно-зависимых опечаток

Опечатка	**Исправление без контекста**	**Исправление в контексте запроса**
Крепк	**С**крепка	*крепк**о*** заваренный чай
Скчать	Ска**ч**ать	как не *ск**у**чать* в отпуске
Моне	Моне	эдуард *м**а**не*

Таким образом, простой эксперимент показал, что для большей части словарных ошибок (74%) исправления в контексте и без контекста совпадают. Это позволяет сделать оценку доли словарных ошибок, которые можно исправлять без учета контекста.

3. Описание подхода

Выбрав класс словарных ошибок в качестве целевого и убедившись, что большая часть словарных ошибок не зависят от контекста, мы можем перейти собственно к оценке надежности исправлений ошибок.

В данной работе мы решаем задачу оценки надежности исправления словарных ошибок как задачу бинарной классификации. Все пары «опечатка-исправление» *(q→c)* должны быть отнесены к одному из классов: «надежная» или «ненадежная». Пара *(q→c)* считается надежной, если *c* является правильным исправлением *q*.

Задача оценки надежности состоит из 4-х этапов:

1. Исправление. Для поиска и исправления опечаток в запросах применяется поисковый спелл-чекер Яндекса, определяющий лучший вариант исправления *C'* запроса *Q* с помощью модели канала с ошибками [3, 5, 6]:

$$C' = argmax\, \mathrm{P}(Q|C)\cdot \mathrm{P}(C),$$

где *P(Q|C)* — вероятность трансформации запроса *C* в запрос *Q* (модель ошибок), *P(C)* — вероятность запроса *C* (языковая модель).

2. Выравнивание. Для выделения слов с ошибками пара запрос-исправление *(Q→C)* выравнивается по словам (см. гл. 2.1). В результате получаем набор пар опечатка-исправление $\{q_i \rightarrow c_i\}$, $i = 1...m$.

3. Фильтрация. Для выделения ошибок типа “ошибки в отдельных словах” применяются простые фильтры:

- Слово с опечаткой и его исправление не содержат пробелов.
- Слово с опечаткой и его исправление принадлежат одному алфавиту.

Дополнительно отфильтровываются ошибки типа «неправильная словоформа» и ошибки в коротких словах (<4 букв).

4. Классификация. Для каждой пары опечатка-исправление ($q \to c$) решается задача классификации “надежное/ненадежное” исправление. В качестве метода машинного обучения используется логистическая регрессия.

4. Метрики качества

Эффективность метода определения надежности исправлений оценивается по полноте и точности классификатора:

- Полнота — отношение количества верно классифицированных надежных исправлений ко всем надежным исправлениям.
- Точность — отношение количества верно классифицированных надежных исправлений ко всем исправлениям, которые классификатор отнес к надежным.

5. Признаки

В этой работе при построении классификатора надежности опечаточных исправлений мы использовали признаки только словарного уровня. Были использованы статистические и лексические признаки, широко применяемые в задачах компьютерной лингвистики [14, 12].

1. Вес по словарной языковой модели для *q* и *c*. Для вычисления признака использовалась 3-грамная языковая модель, собранная из запросов к поисковой системе Яндекс за полгода [4].
2. Вес по буквенной языковой модели для *q* и *c*. Признак вычислялся по 3-грамной буквенной языковой модели, построенной по тем же данным, что и словарная модель.
3. Длины слов *q* и *c* в символах.
4. Присутствие *q* и *c* в словарных источниках. Слова проверялись по русскому и английскому морфологическим словарям, используемым поисковой системой Яндекс.
5. Язык запроса. Так как в рассматриваемых запросах преобладают русский и английский языки (см. табл. 2), в данной работе использовался бинарный признак en/ru. Язык слова определялся по принадлежности к кириллическому/латинскому алфавиту. Напомним, что *q* и *c* принадлежат одному алфавиту (см. гл. 3)

6. Вероятность написания q и c с заглавной буквы. Как показывает опыт, значительную часть плохих подсказок составляют имена собственные, например, редкие фамилии, названия небольших фирм и т. п. Поскольку имена собственные чаще пишутся с заглавной буквы, вероятность написания слова с заглавной буквы можно использовать в качестве признака. Для вычисления этого признака мы использовали корпус, состоящий из 100 миллионов веб-документов.
7. Взвешенная дистанция редактирования для пары ($q \to c$). Под дистанцией редактирования понимается вероятность трансформации $q \to c$, вычисляемая с использованием модели ошибок, описанной в [8].
8. Взаимный словарный контекст q и c. Для q и c строятся вектора слов, с которыми q и c встречаются в 3-грамной словарной языковой модели на расстоянии одного слова. Полученные вектора используются для оценки меры схожести словарных контекстов q и c. В данной работе мы использовали меру схожести контекстов двух слов w_1 и w_2, предложенную в [10]:

$$P_c(\mathrm{w}_2|\mathrm{w}_1) = \sum (\mathrm{P}(\mathrm{w}|\mathrm{w}_1) \cdot P(\mathrm{w}|\mathrm{w}_2) \cdot P(\mathrm{w}_2)) \,/\, P(\mathrm{w})$$

где w принадлежит множеству слов, которые встречаются как со словом w_1, так и со словом w_2. Признак вычислялся по словарной языковой модели.

6. Эксперименты

6.1. Тестовый и обучающий наборы

Для настройки и тестирования предложенного метода были подготовлены обучающий и тестовый наборы. Из дневного лога поисковой системы Яндекс случайным образом были выбраны 30 000 запросов и проверены поисковым спелл-чекером. Спелл-чекер обнаружил и исправил опечатки в 2134 запросах. Для обучения словарного классификатора мы извлекли из пар запрос-исправление все пары слово-исправление. Таких пар получилось 2545. Из них мы удалили пары, состоящие из коротких слов (354 пары) и орфовариантов (41 пара), после чего оставшиеся 2150 пар были переданы на разметку аналитику. Разметка состояла в том, чтобы разделить пары слово-исправление на правильные и неправильные. Правильных оказалось 77 %. Размеченный набор пар слово-исправление был разделен на две выборки (обучающую и тестовую), равные по размеру и по содержанию правильных и неправильных исправлений.

6.2. Базовый набор признаков

За базовый уровень мы приняли показатели классификатора, построенного на признаках, используемых спелл-чекером для выявления и исправления ошибок в запросах. Это веса слова и его исправления по языковой модели, а также взвешенная дистанция редактирования слова и исправления (см. признаки 1,7 в гл. 5).

6.3. Результаты

Учитывая тот факт, что разные поисковые задачи предъявляют разные требования к точности классификации надежности исправлений, для оценки возможности настройки классификатора мы проводили эксперименты с двумя порогами точности (0,9 и 0,95). Результаты, полученные на тестовой выборке, представлены в Таблице 5:

Таблица 5. Качество классификатора на тестовой выборке

Набор признаков	**Полнота** *(порог точности 0,9)*	**Полнота** *(порог точности 0,95)*
Базовый	0,364	0,214
Полный	0,773	0,549

По результатам эксперимента можно сделать следующие выводы:

Полный набор обеспечил прирост полноты автозамен на 0,3÷0,4, что свидетельствует о высокой эффективности признаков словарного уровня.

Если учесть, что опечатки составляют 63,7% всех ошибок (табл. 3), полнота классификатора 0,549 означает возможность автоматически исправлять 35% словарных ошибок в запросах с точностью не ниже 0,95. Это весьма высокий показатель для функции автоматической коррекции. Предложенный метод можно признать пригодным и рекомендовать для автоисправления запросов в поисковой системе.

Из анализа результатов эксперимента следует, что наиболее эффективным признаком оказался «взаимный словарный контекст». Особенность этого признака заключается в том, что он работает не с контекстом данного запроса, а агрегирует информацию о взаимозаменяемости слов в одних и тех же контекстах по разным запросам, собранным за большой период времени. Эта информация позволяет с высокой точностью установить, является ли одно слово опечаткой другого. К сожалению, этот признак не работает для редких слов (из-за отсутствия статистических данных об их взаимных контекстах).

Среди ошибок первого рода следует указать неверное отнесение редких слов к классу опечаток. Проблема заключается в нехватке статистических данных для построения значимых признаков (взаимный контекст, вероятность написания с заглавной буквы) для редких слов (фамилии, названия небольших компаний, новых сайтов и т. п.). Здесь необходимо увеличивать объемы исходных данных, а также добавлять новые признаки, например, присутствие слова в специализированных словарях (товаров, топонимов и т. п.).

Ошибки второго рода связаны с признаком присутствия слов в морфологическом словаре. Для исправления контекстно-зависимых ошибок у классификатора словарного уровня недостаточно данных. В этих случаях, очевидно, необходимы контекстные признаки, не используемые в данной работе.

7. Заключение

В данной работе решалась проблема повышения эффективности автоматического исправления ошибок в поисковых запросах. В качестве целевого класса были выбраны словарные ошибки (пропуск/вставка/замена/перестановка букв в словах), составляющие две трети всех ошибок в запросах. Было показано, что значительная часть ошибок являются тривиальными (исправление очевидно и однозначно), не зависят от словарного окружения и могут быть исправлены в автоматическом режиме.

Для принятия решения о возможности автоматического исправления использовался бинарный классификатор, разделяющий исправления на надежные, пригодные для автозамен, и ненадежные, пригодные только для подсказок. Поскольку тривиальные ошибки не зависят от контекста, для определения надежности исправлений достаточно признаков словарного (бесконтекстного) уровня, что значительно упростило задачу подбора признаков. В работе были использованы наиболее распространенные лексические и статистические признаки, применяемые при решении поисковых и лингвистических задач. Построенный на их базе классификатор показал приемлемое качество и возможность регулирования баланса полнота/точность. С помощью предложенного в работе метода можно с высокой точностью автоматически исправлять больше половины словарных опечаток, т. е. почти треть всех ошибок в поисковых запросах.

Несмотря на положительные результаты, работу по повышению качества автоматического исправления ошибок можно продолжить по многим направлениям. Помимо добавления новых признаков и исправления «несловарных» типов ошибок, в качестве перспективных задач можно назвать добавление новых классов надежности подсказки и использование данных о пользовательских кликах.

References

1. *Baba Y., Suzuki H.* (2012) How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 373–377, Jeju, Republic of Korea, 8–14 July 2012.
2. *Baldwin T., Chai J. Y.* Autonomous Self-Assessment of Autocorrections: Exploring Text Message Dialogues. In: 2012 Conference of the North American Chapter of the Associational for Computational Linguistic: Human Language Technologies, pages 710–719, Montreal, Canada, 2012.
3. *Bajtin A.* (2008) [Ispravlenie poiskovyh zaprosov v Yandekse. Verojatnostnaja jazykovaja model] Rossijskie internet-tehnologii 2008.
4. *Brants T., Popat A. C., Xu P., Och F. J., Dean J.* (2007). Large Language Models in Machine Translation. In: EMNLP'07. pp. 858–867
5. *Brill E., Moore R. C.* (2000). An Improved Error Model for Noise Channel Spelling Correction. In: ACL'00. pp. 286–293
6. *Cucerzan S., Brill E.* (2004). Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users . In: EMNLP'04 pages 293–300
7. *Erehinskja T. N., Titova A. S., Okat'ev V. V.* (2011) Syntax Parsing For Texts With Misspellings In Dictascope Syntax [Sintaksicheskij analiz teksta s orfograficheskimi oshibkami v sisteme Dictascope Syntax] Trudy Mezhdunarodnoj Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual ɜTechnologies: Proceedings of the International Conference "Dialog 2011"], Bekasovo.
8. *Karpenko M. P., Protasov S. V.* (2011) Some Methods for Language Model Pruning [Nekotorye metody ochistki slovarja zaprosov poiska] Trudy Mezhdunarodnoj Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"], Bekasovo.
9. *Kukich K.* (1992) Techniques for Automatically Correcting Words in Text. In: ACM Computing Surveys, Vol. 24, No. 4.
10. *Li M., Zhu M., Zhang Y., Zhou M.* (2006). Exploring Distributional Similarity Based Models for Query Spelling Correction . In: ACL'06. pp. 1025–1032.
11. *Rabiner, L. R.* (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE, 77 (2), pp. 257–286
12. *Specia L., Cancedda N., Dymetman M., Turchi M., Cristianini, N.* (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In: EAMT'09. pp. 28–35
13. *Specia L., Hajlaoui N., Hallett C., Aziz W.* (2011). Predicting Machine Translation Adequacy. In: MTSummit'11. pp. 513–520
14. *Whitelaw C., Hutchinson B., Chung G. Y., Ellis G.* (2009). Using the Web for Language Independent Spellchecking and Autocorrection. In: EMNLP'09. pp. 890–899
15. http://productforums.google.com/forum/#!topic/websearch/g_XQupJ8Pvgcc

СОВМЕСТНАЯ ВСТРЕЧАЕМОСТЬ СЛОВ: ОПЫТ КЛАССИФИКАЦИИ

Паперно Д. А. (denis.paperno@unitn.it)
Università degli studi di Trento, Тренто, Италия

Ройтберг А. М. (cvi@yandex.ru),
Хачко Д. В. (mordol@lpm.org.ru),
Ройтберг М. А. (mroytberg@lpm.org.ru),
ИМПБ РАН, Пущино, Россия;
НИУ Высшая школа экономики, Москва, Россия

Ключевые слова: сочетаемость, коллокации, тематические модели, повторы

BREEDS OF COOCCURRENCE: AN ATTEMPT AT CLASSIFICATION

Paperno D. A. (denis.paperno@unitn.it)
Università degli studi di Trento, Trento, Italy

Roytberg A. M. (cvi@yandex.ru),
Khachko D. V. (mordol@lpm.org.ru),
Roytberg M. A. (mroytberg@lpm.org.ru)
IMPB RAS, Pushchino, Russia; RSU HSE, Moscow, Russia

The paper proposes a substantial classification of collocates (pairs of words that tend to cooccur) along with heuristics that can help to attibute a word pair to a proper type automatically.

The best studied type is frequent phrases, which includes idioms, lexicographic collocations, and syntactic selection. Pairs of this type are known to occur at a short distance and can be singled out by choosing a narrow window for collecting cooccurrence data.

The next most salient type is topically related pairs. These can be identified by considering word frequencies in individual documents, as in the well-known distributional topic models.

The third type is pairs that occur in repeated text fragments such as popular quotes of standard legal formulae. The characteristic feature of these is that the fragment contains several aligned words that are repeated in the same sequence. Such pairs are normally filtered out for most practical purposes, but filtering is usually applied only to exact repeats; we propose a method of capturing inexact repetition.

Hypothetically one could also expect to find a forth type, collocate pairs linked by an intrinsic semantic relation or a long-distance syntactic relation; such a link would guarantee co-occurrence at a certain relatively restricted range of distances, a range narrower than in case of a purely topical connection, but not so narrow as in repeats. However we do not find many cases of this sort in the preliminary empirical study.

Key words: cooccurrence, collocations, topic models, repeats

1. Introduction

Word cooccurrence has innumerous applications in computational linguistics. Much of the early research on co-occurrence focused on lexicographic tasks, using association measures to form lists of candidate multiword expressions to be included in dictionaries (e.g. [Smadja 1993]). Cooccurrence data have further usages in improving parsing algorithms (as in e.g. Yoon et al. [2001]) or as cue on a word's semantics; practical uses of such cues lead to the development of the distributional semantic models (DSMs). Applications of DSMs, range from semantic similarity recognition to word sense induction [Tamir and Rapp 2003], to word sense disambiguation [Mitrofanova et al. 2008], to entailment, to predicting association norms etc. A better understanding of the nature of co-occurrence, to which we aim to contribute here, could help improve many of these computational linguistic models.

Standardly, collocation extraction is based on corpus statistics. Collocations (in the broad, non-lexicographic sense) are word pairs that co-occur more often than expected from the frequencies of individual words and the statistical model adopted [Herbst 1996, Nesselhauf, 2004]. This approach was pioneered by [Firth 1957] and his followers [Halliday 1961] and [Sinclair 1991, Sinclair, Carter 2004]. See [Evert 2004] for methods of collocation extraction and an extensive literature overview.

The linguistic nature of statistically associated collocates varies, and our paper attempts at an exhaustive if coarse-grained classification. One of the best studied collocation types is idiomatic expressions: phrases whose meaning is not reduced to the meanings of constituent parts, such as *set forth*or *real estate*. Another subtype is lexicographic collocations, i.e. phrases with a more or less compositional meaning that are the default expression of a certain complex idea established in usage (e.g. *hard rain*, *strong tea*). Unlike idioms, lexicographic collocations allow for certain variation of phrase components while keeping the meaning largely intact (*strong sweet tea*vs. **real expensive estate*). approaches to lexicographic collocations include Mel'cuk [1998], Fillmore and Key [1988, 1992], and others.

There are cases of statistical association beyond multiword expressions such as idioms discussed above. For example, it has been noticed that members of a semantic field tend to co-occur (this is in fact a subclass of the "clustering" type, see below). However, no systematic classification of collocations by linguistic nature has been proposed to date. This paper fills this gap, proposing such a classification along with heuristics that allow for automatic attribution of collocations to one or another class, which we apply to the Brown corpus [Francis and Kucera, 1964].

This paper reports work in progress; further directions are outlined in section 4.

2. Materials and methods

2.1. Corpus

We conducted a preliminary quantitative analysis of the small but well-balanced Brown corpus. The Brown corpus contains 500 text fragments of approximately 2,000 words each. We lemmatized and retagged the corpus using FreeLing software [http://nlp.lsi.upc.edu/freeling/], which attributed the corpus's 1,010,058 words to 48,153 distinct lemmas.

2.2. Collocation extraction

As a measure of association significance we use the standardized deviation of pair frequency from the expected mean, known as the Z score:

$$Z(w_1, w_2, d) = \frac{f_C - E}{\sqrt{E}}$$

where f_c is the absolute cooccurrence frequency of w_1, w_2, at a given distance d. $E = f_1 \times (f_2/N)$ is the maximum-likelihood estimate of the mean and dispersion for co-occurrence frequency of w_1, w_2 at distance d, assuming the independence of occurrence of w_1 and w_2, where f_1 is the frequency of w_1, f_2 is the frequency of w_2 , N is the corpus size. Although the choice of association measure does affect the ranking of pairs, we note that the set of top pairs remains comparable when switching between measures. For example, among the word pairs that occur at least 3 times in the Brown corpus, in the lists of top 10K pairs the Z and the t scores share 71.8% of pairs, the Z score and PMI 95.8%; t score and PMI 77.5% of pairs. We use Z as our primary measure.

We identify associated pairs, or collocations in the broad sense, as pairs with Z score over 5. The first word in each pair was selected among the top 5K most frequent content words (verbs, nouns, adjectives, adverbs, or numerals). In addition, we used frequency thresholds; the quantitative results below include pairs that occur at least 3, 5, or 7 times. We also discuss the data obtained at the threshold of 5 in more detail.

3. Results

3.1. Classification of collocations

3.1.1. General

There are three groups of collocations with different linguistic nature and distribution.

- «phrases» are multiword expressions with an immediate syntactic relation between words;
- «repeats» are members of (nearly) identical text fragments such as legal formulae, see 3.1.3;
- «clustering» collocations are conditioned by corpus heterogeneity, see 3.1.4.

We conjecture that there is no substantial class of word association beyond these three.

Repeat-based collocations are filtered by entropy (3.1.3), clustering-based collocations are those whose Z score falls below 5 when calculated as described in 3.1.4. The heuristics proposed here are preliminary. For example, one could use methods other than entropy to identify repeats, or rely on paragraphs or other units rather than documents for detecting clustering effects.

The residual collocations are mostly phrasal collocations, which lie within the distance range of 3–5. A small number of statistically associated pairs do not seem to stand in a meaningful relation, and are not attributed to any of the three groups by heuristics. We believe that these data are mostly explained as noise, see discussion below.

3.1.2. Phrasal collocations

This class includes syntactically related associated words, in particular, the idioms and lexicographic collocations mentioned above. Table 1 contains some examples.

Table 1. Examples of phrasal collocations

№	Word 1	Word 2	Distance	Frequency	Z score
1	real	estate	1	24	231.30
2	urethane	foam	1	12	374.66
3	arc	voltage	2	5	160.49
4	great	deal	2	43	138.93
5	play	role	3	10	45.45
6	write	letter	3	12	33.03

Quite a few lexical collocations appear at a range of distances rather than a fixed distance, cf.:

Table 2. *Make clear* at distances 1–5

Distance	Word1	Word 2	Frequency	Z score
1	make	clear	13	17.459
2	make	clear	16	21.653
3	make	clear	6	7.673
4	make	clear	1	0.683
5	make	clear	0	−0.725

3.1.3. Repeats

Repetition of text fragments or clichés is quite common in corpora of naturally occurring text. This repetition raises association scores between all words in the repeat regardless of the distance or syntactic relation between them. Repeats can result from direct copying, as it often happens when one electronic document is created on the basis of another, but they can also stem from natural formulaic expressions. For example, Document H08 (Rhode Island Governor's Proclamations) of the Brown Corpus contains seven proclamations, each ending with the following: «In testimony whereof I have hereunto set my hand and caused the seal of the State to be affixed this 17th day of May in the year of Our Lord one thousand nine hundred and sixty-one». The formula repeats 7 times almost exactly (only the dates vary), boosting the association between all words in it. For instance, the pair *affix this* gets the Z score of 25.16. Another inexact repeat with more variable elements is *between N p.m. and K a.m.*, raising the association between *p.m.* and *a.m.* to 134.2 (Z score), with 4 occurrences at distance 3.

To identify a repeat, we take all occurrences of a pair at a given distance, and calculate the entropy of each position in those contexts. We rely on the heuristic that a small average entropy for all positions for the window containing the given word pair, all positions between them, and 10 positions on each side, indicate a repeat. For each position we calculate entropy as

$$E = -\sum_i \left(P(i) \times \ln(P(i))\right)$$

where i ranges over words, and P(i) is the probability of word i in the given position.

We take the threshold for average entropy across all positions to be 0.8. In contrast to existing approaches, the entropy-based method allows us to identify even inexact repetition.

3.1.4. Clustering

Corpus structure affects statistical word association. If some part of a corpus has higher frequencies of words *x* and *y* than the rest, we also expect it to contain more pairs of *x* and *y*. So even independent occurrence of *x* and *y* in the subcorpus may lead to statistical association given the overall frequencies of *x* and *y*. Let's show the role of corpus heterogeneity by an example.

Verbs *tell* and *think* are quite frequent, these lemmas occur 766 and 1,044 times respectively in the Brown corpus, with relative frequencies $p_1 = 766 \div 1{,}010{,}058 = 0.076\%$ and $p_2 = 1{,}044 \div 1{,}010{,}058 = 0.104\%$. Assuming that these verbs are distributed independently (the null hypothesis for any word pair), the probability of finding *tell* and *think* at any given fixed distance is $p_1 \times p_2$, with $p_1 \times p_2 \times C$ expected occurrences of the pair, where C is corpus size. So we can expect the pair *tell*, *think* to appear roughly once at each distance ($0.076\% \times 0.104\% \times 1{,}010{,}058 = 0.798$).

Now imagine that both *tell* and *think* occur exclusively in fiction, which contributes about a quarter of the Brown corpus, and are not attested elsewhere. In this case the fiction corpus should contain all the pairs of these words, and the expected number can be obtained by multiplying the frequencies of *tell* and *think* in the fiction subcorpus by its size (about 252K words), i.e. $(766 \div 252{,}000) \times (1{,}044 \div 252{,}000) \times 252{,}000 = 3.17$. In fact, the lemmas *tell* and *think* are attested 7 times at distance 10, which corresponds to the Z score of 6.98 (assuming independence of occurrence and expected frequency of 0.798) or 2.15 (assuming that both lemmas occur only in fiction). As one can see, taking into account corpus heterogeneity can lead to significantly different association measures.

(Of course, for *tell* and *think* both models are crude idealizations. The truth is in the middle: for the 7 occurrences of *tell* and *think*, the Z score based on actual corpus heterogeneity is 4.69, almost exactly between 6.98 and 2.15.)

In practice, for almost all lemmas w_1, w_2 there are several rather than two subcorpora characterized by different frequencies of w_1 and w_2. For the purpose of this paper, we treated each document as a potentially distinct thematic subcorpus. This assumption is harmless: if in fact documents form blocks characterized by even word frequency distributions, the sum of expected frequencies for all documents will give, on average, a correct estimate of the expected frequency for the whole block. As the expected overall frequency of a pair in the corpus, we take the sum of expected frequencies for all documents:

$$E = -\sum_D \left(f_{1(D)} \times (f_{2(D)} / N_D)\right)$$

where D ranges over all documents, N_D is the size of D, $f_{1(D)}$, $f_{2(D)}$ are the frequencies of w_1, w_2 in D. This calculation is valid regardless of how diverse document sizes are. Adjusted association scores such as Z can then be calculated on the basis of this corrected E. If such an adjusted score of a collocation is low, then the collocation owes its initially high association measure solely to clustering of both words in the same documents.

Of course we do not imply that the property of two lexemes to occur in the same texts is irrelevant. To the contrary, it reveals an association through features of genre, style, or topic; this includes sameness of semantic field. What we want to emphasize is that association by clustering is substantially different from other types of collocation, and should be separated for practical applications. For instance, extraction of lexicographic collocations might be improved by disregarding the effects of clustering, while for the study of topic structure only clustering effects are relevant, but not other types of co-occurrence.

We also note that for clustering collocations, the specific numeric value of association is an artifact of corpus composition. Indeed, it was quite an arbitrary decision on behalf of the creators of the Brown corpus to dedicate just a quarter of its size to fiction, as opposed to a bigger or a smaller part. But it is the fraction of texts of each topic and genre that determines how high the association measure will be for words characteristic of that topic or genre.

Corpus heterogeneity creates quite many associated pairs. Examples include pairs *member—church*, *student—college*, *state—federal*, which are not spread across the whole corpus but are clustered in a small set of documents. For *member—church* the basic Z score is 14.35, but it drops to 4.19 when adjusted word frequencies in individual documents; *student—college* has Z scores of 18.05 and 3.26, *state—federal* 15.73 and 4.44, respectively. All of these pairs illustrate a primarily topical relation between words.

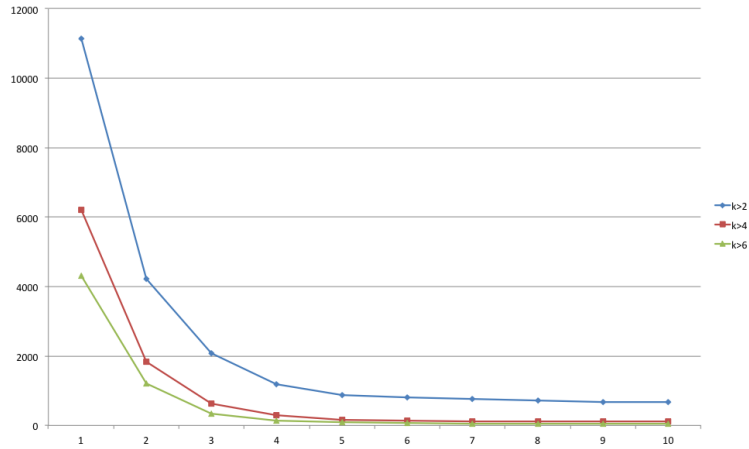
3.2. Quantitative observations

3.2.1. Distance and collocation type.

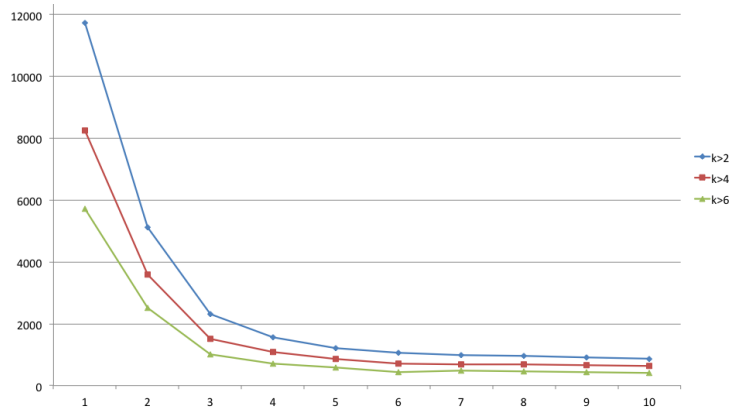
Figure 1.a–c shows the dependence of the number of collocations found on distance, cf. 2.2. Both content and function words are included. Distribution shape is stable across association measures (a–c).

As the graphs show, at distance of 5 the number of collocations stabilizes, while short distances (1 and 2) contribute many more pairs. Between 3 and 5 the number of collocations decreases relatively slowly. This pattern agrees with the standard assumption that collocations are mostly found at distances up to 3–5 words. Our own informal observations on the lists of collocations agree with this assumption.

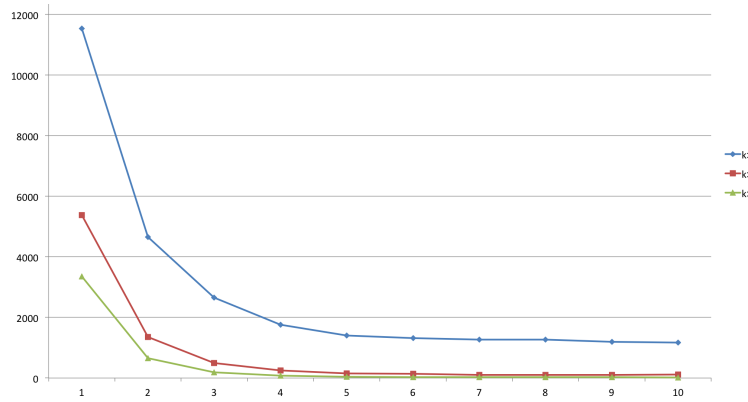
The pattern is the same for all three thresholds. In what follows we use only the 5 threshold.



a) $Z_{score} > 5$



b) $t_{score} > 1.7$



c) PMI > 0.72

Fig. 1. Number of collocations by distance and frequency threshold

3.2.2. Distribution of individual types of collocations

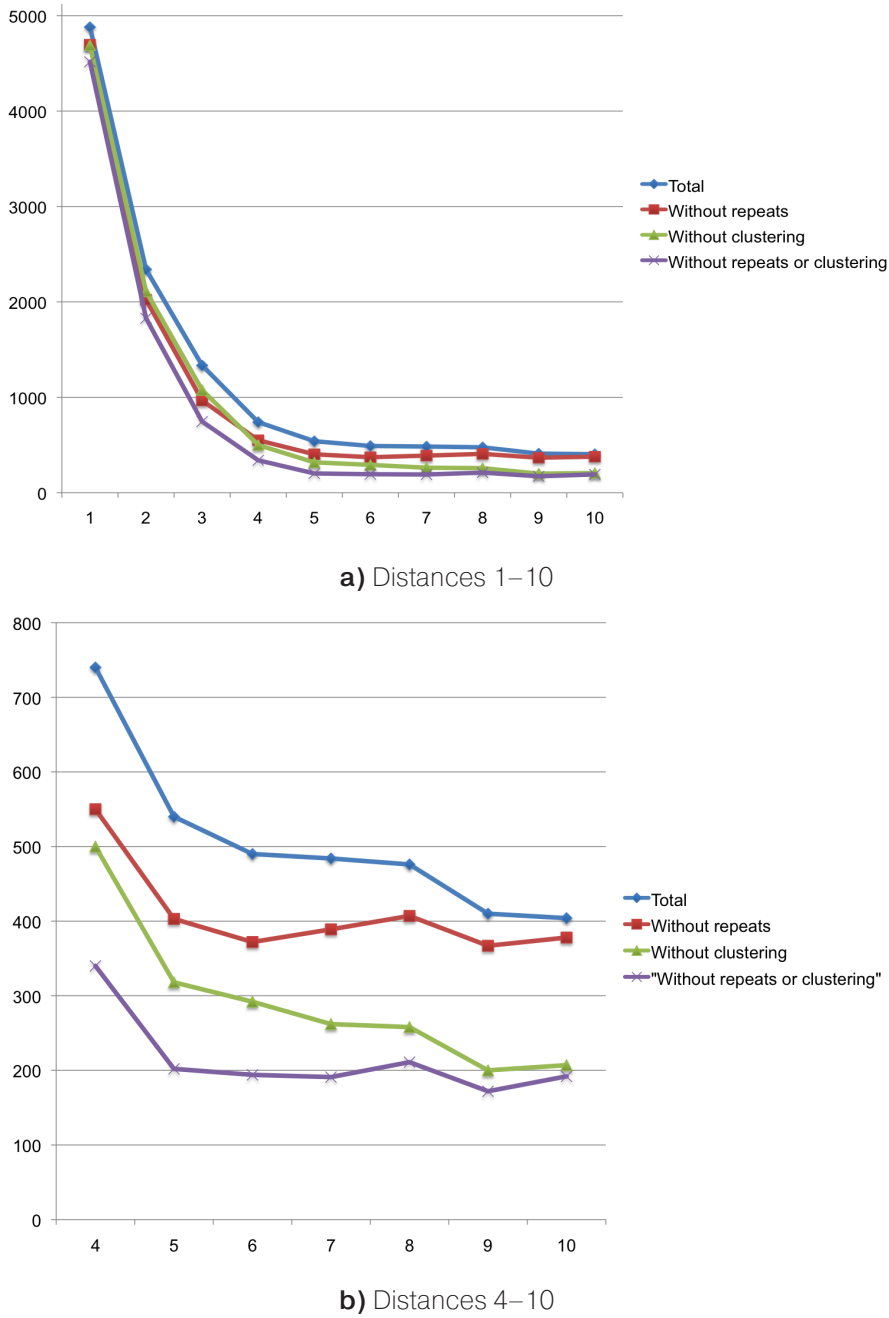


Fig. 2. Number of collocations of different classes; only content words considered

As we see from graph 2, on distances (greater than 5) the majority of pairs belong to the "clustering" collocations and repeats. Most of the rest, by our qualitative observations, is statistical noise. We deal with pairs of words which exhibit no meaningful relation(*say — New York*, *number — eye, so — work*). Raising the frequency threshold essentially eliminates those "remote" collocations. Some of the "remote" collocations are mostly due to the clustering effect, but were not shelled out by the formal criterion because of random statistical fluctuation on top of the clustering effect. One such example is the pair *optimal — state*: its Z score at distance 8 drops from 32.6 to 6.7 when taking corpus heterogeneity into account, and the pair also occurs at distances of 6 and 9 (Z drops from 6.39 to 0.8) and 7 (from 19.48 to 3.75).

A priori, there could be more types of remote collocations: two words could be related by a relation of syntactic or discourse nature at a greater distance. We could expect that a pronoun is anteceded by a coreferent name (anaphora), that after mentioning the evil an author is likely to talk about the good (associative relation), or that discourse markers would tend to occur in a certain order (*although…still*, *on the one hand…on the other hand*), one sentence or in different sentences.All of these cases are genuine word-word relations at remote distances, as opposed to links mediated by a popular quote or by the text topic. , the analysis of the Brown data reveals almost no examples of this kind. The only exceptions are pairs of markers *first…then*, *only… also*.

4. Discussion

While individual factors of statistical association have been noticed previously (cf. [Evert 2004]), this paper is the first attempt at a substantial classification of associated pairs by the main underlying factor. The novel tentative conclusion of this paper is that the three types discussed here exhaust all the statistical collocations. One practical consequence, briefly discussed below, is that the number of different types of collocations could be a useful characteristic of a corpus.

The classification proposed here can help improve any practical applications of lexical association measures, from collocation extraction to refining distributional semantic models that build semantic vectors based on association measures [Turney, Pantel 2010]. For small distances, it will be interesting to evaluate how much filtering repeats and clustering helps identify true lexicographic collocations.

Automatic classification of collocations that we implemented can also serve as a basis for qualitative assessment of natural language corpora. In particular, a corpus of identical size should be more valuable for most applications if it has fewer repeats. Perhaps even more significant could be the number of "clustering" collocations. Indeed, if each of those pairs points to a particular topic or genre represented by a distinct subcorpus, then abundance of such topical pairs, other things being equal, tells us that the corpus is diverse and balanced. The balancing effect arises because if a certain topic takes up a disproportionately large part of the corpus, the word pairs that correspond to the topic get lesser weight. In an analogous but more balanced corpus the same topic will contribute more statistical collocations thanks to a greater degree of clustering. We leave the development of a specific procedure of corpus evaluation to future research.

References

1. *Francis W. N., Kucera H.* (1964), Department of Linguistics, Brown University, Providence, Rhode Island, USA. http://icame.uib.no/brown/bcm.html.
2. *Tamir R., Rapp R.* (2003), Mining the Web to Discover the Meanings of an Ambiguous Word. IEEE International Conference on Data Mining — ICDM, pp. 645–648.
3. *Bell E. J. L.* (2007), Collocation Statistical Analysis Tool: An evaluation of the effectiveness of extracting domain phrases via collocation. B.Sc. Dissertation, Lancaster University.
4. *Evert S.* (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgar.
5. *Fillmore Ch., Kay P., O'Connor M.* (1988), Regularity and idiomaticity in grammatical constructions: The case of LET ALONE, Language, Vol. 64, pp. 501–518.
6. *Firth J. R.* (1957) Modes of Meaning, Papers in Linguistics 1934–51, pp. 190–215, Oxford University Press.
7. *Halliday M. A. K.* (1961), Categories of the Theory of Grammar, Word 17, pp. 241–92.
8. *Herbst T.* (1996) What Are Collocations: Sandy Beaches or False Teeth? English Studies No. 4, pp. 379–393.
9. *Manning C., Schutze H.* (1999) Foundations of Statistical Natural Language Processing, MIT Press. Cambridge.
10. *Mel'cuk, I.* (1998). Collocations and Lexical Functions. Cowie, A. P. (ed.), Phraseology. Theory, Practice and Applications, Oxford University Press, pp. 23–53, Oxford.
11. *Mitrofanova O. A., Belik V. V., Kadina V. V.* (2008), Corpus Analysis of Selectional Preferences of Frequent words in Russian [Korpusnoe issledovanie sochetaemostnyh predpochtenij chastotnyh leksem russkogo jazyka], Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog 2008". Moscow.
12. *Nesselhauf N.* (2004) Collocations in a Learner Corpus, Amsterdam/Philadelphia, Benjamins.
13. *Padró L., Stanilovsky E.* (2012) FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey.
14. *Sinclair J.* (1991), Corpus, Concordance, Collocation, Oxford University Press, Oxford.
15. *Sinclair, J., Carter, R.* (2004) Trust the Text. Language, Corpus and Discourse, Routledge, London/New York.
16. *Zaharov V. P., Hohlova M. V.* (2010) Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts, Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialog 2010.

ИНКОРПОРАЦИЯ В ГЛАГОЛЬНЫХ ФОРМАХ В РУССКОМ ЯЗЫКЕ

Пазельская А. Г. (avis39@mail.ru)

ЗАО «Ай-Теко», Москва, Россия

Ключевые слова: глагольные композиты, именные композиты, инкорпорация, глагольное словообразование, именное словообразование

INCORPORATION IN VERB FORMS IN RUSSIAN

Pazelskaya A. G. (avis39@mail.ru)

I-Teco JSC, Moscow, Russia

This paper investigates constraints on incorporation of nominal roots into compound verbs in Russian. This type of incorporation is generally impossible. The author examines several apparent exceptions from this generalization and proposes an explanation to the constraint itself as well as to the exceptions. A special attention is paid to the relation between (non-existing) compound verbs and compound nominals corresponding to the same nominal+verbal complex.

Exceptions from the general constraint “no nominal roots within a compound verb” include deverbal adjectives which are formally equivalent to participles, verbs with reflexive and reciprocal “pronominal” components, verbs derived from compound nominals and compound verbs that have lost their semantic interpretability as complex verb.

This interpretability is postulated to be the crucial feature correlating with the constraint on verbal compounds with nominal component, for the reason that this interpretability indicates the presence of two independent nodes (V and NP) in the structure of the compound. If such two node structure becomes a verb, the inner NP node receives case from higher structure levels and cannot incorporate into compound verb.

Key words: compound verbs, compound nouns, incorporation, verbal derivation, nominal derivation

1. Инкорпорация именных основ в глаголах и именах

В заголовок этой работы вынесено невозможное в русском языке явление, и задача данной статьи — исследовать границы этой невозможности с целью попытаться дать ей объяснение и вскрыть лежащие в её основе механизмы.

Простое описание исследуемой проблемы состоит в том, что инкорпорация именной основы в глагольную основу доступна для отглагольных имён, но не для глаголов, к которым эти имена восходят (здесь и далее примеры с указанием источника в квадратных скобках взяты из НКРЯ, см. ruscorpora.ru)[1]:

(1) *Детей учили рисованию, технологии* ***деревообработки*** *и резьбы по дереву.* [Л. Резанов. Богородская игрушка // «Наука и жизнь», 2008]

(2) **Детей учили* ***деревообрабатывать****.*

Это ограничение распространяется на все сложные слова вида «именная основа + глагольная основа», независимо от семантической роли участника, названного именным корнем, в ситуации, описываемой глаголом[2]. Так, в (1) представлен наиболее распространённый случай, когда именная основа обозначает Пациенс, в (3) — Агенс, в (4)–(5) — косвенного участника:

(3) *А дальше — снова* ***бурелом****. У Алексея Петровича свой мир — в голове, настоящий.* [Татьяна Толстая. Ночь (1983)] — **А дальше — снова* ***буреломит****.*

(4) *Как раз в 1964 году открыли действующий и сейчас* **путепровод**. [коллективный. История БМО (2006–2008)] — **Как раз в 1964 году стали* ***путепроводить****.*

(5) *Море снабжало пищей местных рыбаков, отсюда уходили в дальние вояжи* ***мореплаватели****, открывшие, между прочим, Канаду и далёкие Фолклендские острова.* [Владимир Гаков. Во Францию — на машине времени (2001) // «Туризм и образование», 2001.03.15] — **Отсюда уходили* ***мореплавать****.*

---

[1] Слово «инкорпорация» традиционно используется для обозначения включения одной основы в состав другой в словоизменении (в полисинтетических языках). Мы будем трактовать этот термин несколько расширительно, подразумевая под ним словообразовательные отношения в русском языке. Синонимы для понятия «глагол с инкорпорированной именной основой» — «сложный глагол с первым компонентом — основой существительного» (Шведова, 1980; Янко-Триницкая, 2001), глагольный композит на базе имени и глагола (ср., например, Озерова, 1998; Петров, 2003).

[2] Мы здесь не касаемся вопроса интерпретации существительных и глаголов (если таковые возможны) с инкорпорированными именными основами: для русского этот вопрос был частично освещён в Богданов, 2007; ср. тж. Bagasheva, 2011; Rice, Prideaux, 2012 для английского; Booij, 2009 для голландского и японского; а также конструкционный подход на базе английского в Tuggy, 2005.

Как можно видеть выше, возможность инкорпорации именной основы в отглагольное существительное и её невозможность для глагола не зависит ни от суффикса отглагольного существительного, ни от того, является ли оно именем ситуации (1), результата (3), инструмента (4) или деятеля (5).

2. Исключения и пограничные явления

2.1. Именные основы в причастиях

Можно привести случаи, когда именная основа инкорпорируется в формально глагольную форму — причастие:

(6) *Кроме того, возможны поставки **деревообрабатывающих** станков, машин для упаковки, холодильных установок, разливочных линий, небольших пекарен и сушилок.* [Польские продовольственные товары на российском рынке (2002) // «Внешняя торговля», 2002.03.29]

Однако несложно показать, что это не причастие, а отглагольное прилагательное, поскольку оно не допускает при себе никаких приглагольных наречий, ср.:

(7) **быстро/*хорошо/*с трудом **деревообрабатывающий** станок*

(8) *быстро/хорошо/с трудом **пишущий** человек*

2.2. Глаголы с местоименными компонентами *само-* и *взаимо-*

Отглагольное существительное может включать в себя не именной, а «местоименный» компонент *само-*:

(9) *Но вся книга и есть раскаяние, **самобичевание**, непроходящая боль и мука, вечная мука.* [Виктор Астафьев. Затеси (1999) // «Новый Мир», 2000].

(10) *С применением «пантер» вообще произошел конфуз: машины прибыли на фронт настолько несовершенными, что большинство просто-напросто сломалось — чего только стоит **самовозгорание** двигателей!* [Илья Бояшов. Танкист, или «Белый тигр» (2008)]

Компонент *само-* является местоименным в том смысле, что в отглагольных существительных он привносит практически весь спектр производных значений, что постфикс *-ся* в глаголах: в (9) он представлен в возвратном значении, в (10) — в декаузативном.

Интересным образом, вне зависимости от значения *само-*, оно, в отличие от именных корней, возможно и в глаголах, с обязательным дублированием постфиксом *-ся*:

(11) *Для умилостивления гнева Господня совершались покаянные шествия с заженными свечами, тихим пением и громкими воплями* ***самобичующихся***. [Д. С. Мережковский. Петр и Алексей (1905)]

(12) *Теплолюбивые микробы, живущие в торфе, в жаркое лето могут так его нагреть, что он* ***самовозгорается***. [Б. Сергеев. Печь и холодильник // «Юный натуралист», 1975]

Показательно, что Е. А. Земская при разборе новых слов, зафиксированных в 1981 году по сборнику НС-81, обнаружила лишь шесть глаголов, образованных от двух основ (=словосложением), из них пять с *само- (самосохраняться, самоудобряться, самоускоряться, самоограничиваться, самоорганизоваться)* и один с *вибро-* — *виброизолировать*. Для сравнения, только имён лиц с инкорпорацией именной основы нашлось 22. Слово *виброизолировать* будет обсуждено ниже наряду с рядом других исключений. Отметим лишь, что глагол *виброизолировать* образует также нормальное причастие, а не просто отглагольное прилагательное, ср. (13) и (7) выше:

(13) *хорошо* ***виброизолирующие*** *материалы*

Наряду с само встречаются также регулярно образующиеся глаголы с инкорпорированным «местоименным» компонентом *взаимо-*: *взаимодействовать*, *взаимообогащать(ся)*, *взаимодополнять(ся)* и т. п., что примерно соответствует взаимно-возвратному залогу.

2.3. Глаголы, образованные от существительных с инкорпорированной именной основой

Однако именные основы встречаются в составе не только маргинальных сложных глаголов, при условии что эти глаголы образованы от имён, которые, в свою очередь, содержат инкорпорированную именную основу:

(14) *«Бог в помощь, — говорю, — бабушка. Червей копаешь,* ***рыболовствуешь****?» Разумеется, в шутку.* [Б. Л. Пастернак. Доктор Живаго (1945–1955)]

(15) *Сейчас я буду страшно рефлексировать и* ***женоненавистничать***. *http://www.rusrep.ru/article/2011/07/12/man_end*

Так, в (14) глагол *рыболовствовать* образован от существительного *рыболовство*, а оно — от *рыболов*, а *женоненавистничать* в (15), не зафиксиро-

ванное в Корпусе, но встречающееся в Яндексе, — от *женоненавистник*. Глаголы с инкорпорацией именной основы непосредственно в глагольную выглядели бы как **рыболовить* и **женоненавидеть*, и таких глаголов в русском языке закономерно нет.

2.4. «Обратная деривация» от существительных с инкорпорированной именной основой

В книге Янко-Триницкая, 2001 также приводятся примеры глаголов, образованных и от других существительных, включающих в себя именную и глагольную основы — необязательно имен ситуаций: например, *сенокóсить*, *пылесосить* от *сенокос* и *пылесос* (глаголы, образованные присоединением именной основы непосредственно к глаголу, звучали бы как **сенокоси́ть* и **пылесосать*, соответственно). Такой же деривационный путь Янко-Триницкая предполагает и для глаголов *психоанализировать* (от *психоанализ*), *местожительствовать* (от *местожительство*) и даже *мелодекламировать* (от *мелодекламация*), для которых подобного доказательства привести уже нельзя.

Более того, Янко-Триницкая также отмечает, что «очень часто существительное <…> представляет собой обычное, широкоупотребительное слово, а глагол — необычное и даже окказиональное. В этих случаях происходит так называемое “обратное словообразование”, когда к имеющемуся сложному имени подбирается по аналогии возможный базовый глагол, например: бракосочетание — бракосочетаться, радиовещание — радиовещать, звукоподражание — звукоподражать» и т. п. (всего 16 примеров; см. Янко-Триницкая, 2001, с. 271).

Особенно важно для нас здесь то, что автор постулирует деривацию от существительного к глаголу даже несмотря на внешне противоположное направление формальной деривации и на противоположное же направление деривации в той же основе без инкорпорации: от *вещать* к *вещание*, но от *радиовещание* к *радиовещать*. Это, однако, согласуется и с интуицией автора данной статьи, и с данными словоупотребления: в НКРЯ в основном корпусе глагол радиовещать представлен одним вхождением в виде причастия, если вообще не отглагольного прилагательного:

(16) *Совещание весьма нужное: за два года своего существования О-во «Радиопередача» проделало значительный путь: от неизвестного общества<…>— к Всесоюзному О-ву «Радиопередача» с семнадцатью* ***радиовещающими*** *станциями по всему СССР.* [Маро. Совещание уполномоченных «Радиопередачи» // «Радио Всем», 1927]

Слово радиовещание на том же корпусе имеет 536 вхождений.

В Грамматике 1980 приводится несколько другой список «отдельных глаголов, не образующих словообразовательных типов, в которых в качестве первого компонента выступает также основа существительного или прилагательного:

видоизменить, *злоупотребить*, *кровохаркать*, *мелодекламировать* (декламировать под мелодию) (с усечением первого компонента), *плодоносить* (приносить плоды) (с отсечением префикса мотивирующего глагола), *трудоустроить* (спец.)»

Про мелодекламировать см. выше; нетрудно также показать, что по крайней мере некоторые из эти глаголов являются кальками из других языков, где ограничения на сложные глаголы с именными основами, очевидно, нет. Или же кальками из других языков являются существительные, с которыми эти глаголы соотносятся. Вот, например, что пишет В. В. Виноградов про слово *видоизменение*: «Слово *видоизменение* появилось в русском научном языке первой половины XIX в. Оно представляет собой калькированный перевод латинского *modificatio*, французского *modification*, немецкого *Modifikation*» (Виноградов 1999: 87).

Аналогично, глагол *кровохаркать* соотносится с существительным *кровохаркание* — калькой латинского *haemoptoe*, *плодоносить* — с *плодоношение* (лат. *frūctificātio*), *злоупотреблять* — с лат. *abutor* и т. п. (напрямую или через другие европейские языки). Зачастую побочным действием калькирования оказывается утрата семантической прозрачности результата: *трудоустроить* — это не *устроить (кому-либо) труд*, не *устроить (кого-либо) на труд* и т. п. Часть подобных слов относится к устаревшим, ср. упомянутые в Словаре русского языка XVIII века *мореплавать* и *мореплавствовать*.

Сюда же примыкает и группа глаголов с именным компонентом, обозначенных в Шведова, 1980 как результат продуктивного словообразовательного типа на фицировать: *радиофицировать*, *электрифицировать*, *газифицировать*. Дело в том, что второй компонент этих глаголов связанный и вне сложного слова невозможен: в русском языке нет глагола **фицировать*. К этому же типу можно отнести и глагол *виброизолировать*, у которого, наоборот, связанным оказывается первый (именной) компонент.

Таким образом, в русском языке все такие глаголы утратили (или никогда и не имели) членимость на именную и глагольную основы, и их нельзя считать результатом действия живого словообразовательного процесса, присоединяющего именные основы к глагольным для образования сложных глаголов.

3. Анализ

Таким образом, для объяснения вышеизложенного нам нужна теория, которая позволяла бы ответить на вопросы, почему в русском языке:

1) в сложных словах инкорпорация именной основы в глагольную возможна в именных формах (существительных и прилагательных) и невозможна в глагольных (даже если речь идёт о причастиях, совпадающих по форме с отглагольными прилагательными);
2) существительное, образованное от глагольной основы, способно образовать новый глагол, и инкорпорированная внутри существительного именная основа этому не препятствует;

3) возможно образование глаголов от основ, содержащих в своём составе именной и глагольный компоненты, но утративших членимость на синхронном уровне;
4) инкорпорация в глаголы «местоимений» *само-* и *взаимо-* не имеет таких ограничений, как инкорпорация именных основ.

3.1. Деривационная история существительных с инкорпорированной именной основой

Попытаемся ответить на вопрос о деривационной истории слов типа деревообработка. Традиционный взгляд на отглагольные имена состоит в том, что они образуются от глагольной основы (17):

(17) *обработ-(ать) + -ка = обработ-ка*

Слово деревообработка теоретически могло бы появиться двумя путями: путём словосложения из существительных *дерево + обработка* (18а) или путём номинализации из глагола **деревообработать* (18б):

(18) а. *дерево + обработка = деревообработка*
б. *дерево + обработать = *деревообработать*
**деревообработ-(ать) + -ка = деревообработка*

Отсутствие в русском языке глаголов вида **деревообработать* заставляет в качестве нулевой гипотезы выбрать (18а) — по крайней мере, для слова *деревообработка*. Но даже в тех случаях, когда соответствующий глагол в русском языке есть, исследователи словообразования стремятся, как мы видели выше, возводить не существительное к глаголу, а наоборот — зачастую вопреки формальному направлению деривации.

На схеме (19) представлена предполагаемая деривация для глагола *местожительтвовать*, а (20) — *видоизменять*:

(19) *жить + тель = житель*
житель + ство = жительство
место + жительство = местожительство
местожительство + овать = местожительствовать

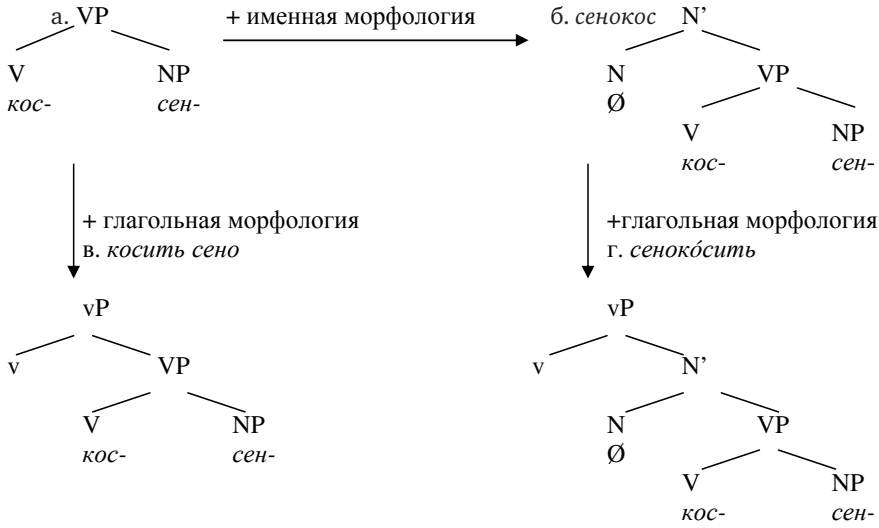
(20) *лат. modi-ficatio = рус. видо-изменение*
видоизменение = видоизменяться

Однако уже с глаголами *сенокóсить* и *пылесосить* такая трактовка вызывает определённые трудности, поскольку их происхождение нельзя объяснить ни по схеме (18а), ни по схеме (18б). Действительно, в русском языке нет существительных **кос* и **сос*, от которых по схеме (18а) могли бы получиться имена

сенокос и пылесос, от которых, в свою очередь, могли бы появиться сенокóсить и пылесосить. С другой стороны, для постулирования деривации по первой части схемы (18б) нам бы понадобилось, во-первых, объяснить, почему в данном случае оказывается возможным присоединение имени к глаголу для образования сложного глагола, и, во-вторых, допустить существование глаголов **кóсить* и **сосить*.

Наблюдаемые формы сложных слов указывают на то, что они появляются «целиком», а не путем сложения двух имеющихся по отдельности слов. Таким образом, можно предположить такой механизм деривации сложных слов, в состав которых входят именная и глагольная основы[3]:

(21) *сенокос, сенокóсить и косить сено*



Рассмотрим образование существительного сенокос, глагола *сенокóсить*, а также группы косить сено, представленное на схеме (21). Все они включают в себя элементарную структуру из VP, состоящей из именной основы *сен-* и глагольной основы *кос* (21а). Дальнейшая судьба этой VP может быть двоякой.

1) Во-первых, она может присоединить именную морфологию и стать существительным *сенокос* (21б). При этом она претерпевает два изменения. Первое — это присоединение суффикса отглагольного имени (в данном случае нулевого, но ср., например, *-ка* в слове *деревообработка*). Второе — объединение вершин снизу вверх (предложенное в Baker 1988, 2011, Baker et al. 2005, Barrie 2012 head-to-head movement; ср. тж.

[3] Дерево для номинализованной структуры с инкорпорацией (21б) взято из Богданов, 2007, где оно было построено по мотивам Baker, 1988).

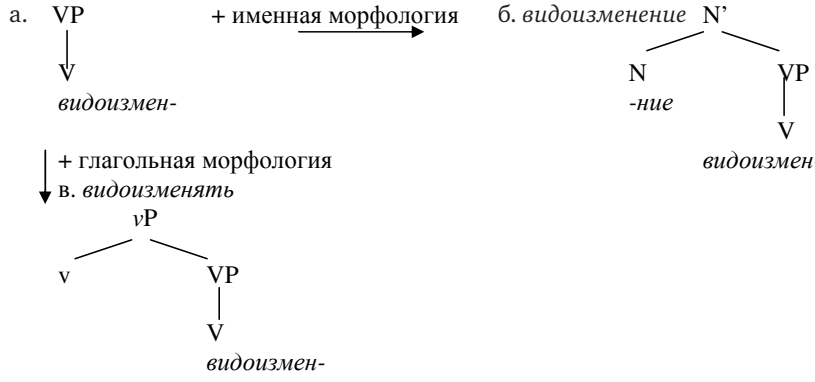
синтаксическое решение без передвижения в Muro, 2009). При этом NP *сен-* (с соединительным гласным между объединяемыми основами) объединяется с V *кос-* и с вышеупомянутым суффиксом отглагольного имени. Существительное *сенокос*, в свою очередь, поверх именной может получить глагольную морфологию — как и многие другие имена существительные в русском языке, ср. *магнит — магнитить*, *пудра — пудрить*. Так получается глагол *сенокóсить* (21г). Важно, что после номинализации структура VP *кос- сен-* теряет прозрачность извне, и получившийся из номинализованной VP глагол уже не знает о наличии внутри прямого дополнения *сен(о)*.

2) Во-вторых, глагольная группа *кос- сен-* может непосредственно присоединить глагольную морфологию с дальнейшими глагольными проекциями (vP, AspP, TP, IP…). При этом vP приписывает прямому дополнению винительный падеж — следовательно, инкорпорироваться в глагольную основу оно не может. Именно поэтому мы не наблюдаем глагола *сенокоси́ть*.[4]

3.2. Кальки и обратная деривация

Обратимся к глаголам с именной основой в составе, образованным «обратной деривацией» от имён, калькированных из других языков, как, например, *видоизменять* от *видоизменение*. В их внутренней структуре отсутствует членение на отдельные V и NP, они представляют собой единую основу:

(22) *видоизменять и видоизменение*



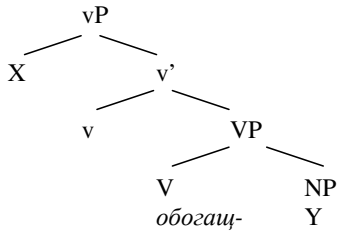
[4] Мы здесь не касаемся вопроса интерпретации существительных и глаголов (если таковые возможны) с инкорпорированными именными основами: для русского этот вопрос был частично освещён в Богданов 2007, ср. тж. *Bagasheva* 2011, *Rice*, *Prideaux* 2012 для английского, *Booij* 2009 для голландского и японского, а также конструкционный подход на базе английского в *Tuggy* 2005.

Соответственно, в случае присоединения такой основой глагольной морфологии винительный падеж приписывать нечему, и основа остаётся единой. Это же верно и для других случаев, когда основа из глагольного и именного компонента не членится в современном русском языке: глаголов на *-фицировать*, *плодоносить*, *радиовещать* и т. п.

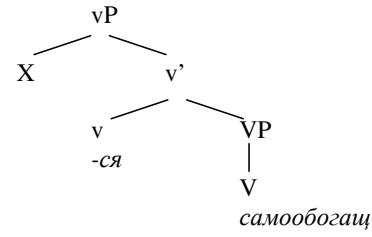
3.3. *Само-* и *взаимо-*

Рассмотрим присоединение к глагольным основам компонента *само-* (те же рассуждения применимы и к *взаимо-*). По своей семантике само работает как показатель рефлексивизации, т. е. преобразования глагольной/именной основы из активного залога в возвратный (с сопутствующими чисто возвратными, декаузативными и т. п. значениями). Поэтому логично рассматривать его наряду с другими рефлексивизаторами в русском и других языках. Как показано в классических работах Reinhart&Siloni 2004, 2005, показатель рефлексива является поверхностным выражением оператора, который удаляет (reduce) из глагольной структуры семантическую роль, ассоциированную с внутренним аргументом, если она референциально совпадает с внешним аргументом:

(23) а. *X обогащает Y*



б. *X самообогащается*



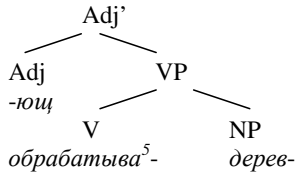
Всё это происходит существенно раньше приписывания падежей (в некоторых языках, к которым, по-видимому, относится и русский — в словаре), поэтому никаких препятствий для существования глаголов с *само-* (и, аналогично, *взаимо-*) в русском языке нет.

3.4. Отглагольные причастия и отглагольные прилагательные

Остаётся объяснить, почему по-разному ведут себя отглагольные прилагательные и совпадающие с ними по форме причастия: первые могут инкорпорировать глагольную основу, вторые нет. Для этого нам придётся отказаться от лежащей на поверхности идеи о «конверсии» причастия в отглагольное прилагательное: в случае прилагательного *деревообрабатывающий* (и аналогичных отглагольных по форме прилагательных) в силу отсутствия причастия конвертировать просто нечего. Остаётся лишь постулировать, что

в прилагательном *деревообрабатывающий* появляется адъективирующий суффикс, совпадающий с причастием лишь по форме, причём присоединяется он достаточно низко, т. е. до приписывания vP винительного падежа прямому дополнению:

(24) *деревообрабатывающий*



Поскольку адъективизатор *-ющ* присоединяется к VP раньше приписывания винительного падежа прямого дополнения, отглагольное прилагательное с инкорпорированной именной основой оказывается возможно.

«Отглагольные» прилагательные с инкорпорированной именной основой, подобно другим прилагательным в русском языке, могут субстантивироваться:

(25) *А еще в одной из торговых точек на Набережной мы нашли замечательную фигурку кальмарчика, вырезанную из дерева. Стоит деревянное* ***морепла-вающее*** *180 рублей.* [Ксения ВОРОНЕЖЦЕВА. Туристы вывозят из Владивостока воздух и морскую соль // Комсомольская правда, 2009.09.05]

Литература

1. *Богданов А. В.* (2007) Генерическая vs. эпизодическая интерпретация номинализаций-компзитов в русском языке. Структуры и интерпретации: Работы молодых исследователей по теоретической и прикладной лингвистике. М.: Изд-во Моск. ун-та. С. 33–57.
2. *Виноградов В. В.* (1999) История слов. М.: ИРЯ РАН.
3. *Земская Е. А.* (2005) Словообразование как деятельность. М.: КомКнига.
4. *НС-81* (1986) Новое в русской лексике. Словарные материалы — 81. М.
5. *Озерова Е. Г.* (1998) Сложные слова в детской речи. Дисс…. кандидата филологических наук. Белгород: Белгородский гос. ун-т.
6. *Петров А. В.* (2003) Гнездовой толково-словообразовательный словарь композитов, Симферополь.
7. *Словарь* русского языка XVIII века (1984). Вып. 13. Молдавский — Напрокудить. Л.: Наука. Эл. публикация: http://feb-web.ru/feb/sl18/slov-abc/13/sld03306.htm
8. *Шведова Н. Ю.* (отв. ред.) (1980) Грамматика современного русского литературного языка. Т. I–II, М.: Наука.
9. *Янко-Триницкая Н. А.* (2001) Словообразование в современном русском языке. М.: Индрик.

10. *Bagasheva A.* (2011). Compound verbs in English revisited. Bucharest Working Papers in Linguistics, Vol. 1, pp. 125151.
11. *Baker M.* (1988) Incorporation: A theory of grammatical function changing, Chicago, University of Chicago Press.
12. *Baker M.* (2011). On the syntax of surface-adjacency: The case of pseudo noun incorporation. Manuscript, Rutgers University. Available at http://www.rci.rutgers.edu/Bmabaker/PNI-adjacency.pdf.
13. *Baker M., Aranovich R., & Golluscio L. A.* (2005). Two types of syntactic noun incorporation: Noun incorporation in Mapudungun and its typological implications. Language, 138–176.
14. *Barrie M.* (2012). Noun Incorporation and the Lexicalist Hypothesis. Studies in Generative Grammar, 22, 235–61.
15. *Booij G.* (2009). A constructional analysis of quasi-incorporation in Dutch. Gengo Kenkyu, 135, 5–27.
16. *Muro A.* (2009) "Noun Incorporation: A New Theoretical Perspective." PhD Diss, Universita degli studi di Padova, Padua.
17. *Reinhart T., Siloni T.* (2004). Against the unaccusative analysis of reflexives. The unaccusativity puzzle, Oxford University Press, pp. 288–331.
18. *Reinhart T., Siloni T.* (2005). The lexicon-syntax parameter: Reflexivization and other arity operations. Linguistic inquiry, 36(3), pp. 389–436.
19. *Rice S., & Prideaux G.* (2012). Event-packing: the case of object incorporation in English. In Proceedings of the Annual Meeting of the Berkeley Linguistics Society (Vol. 17).
20. *Tuggy D.* (2005). Cognitive approach to word-formation. In Handbook of word-formation. Springer Netherlands. Pp. 233–265.

References

1. *Bagasheva A.* (2011). Compound verbs in English revisited. Bucharest Working Papers in Linguistics, Vol. 1, pp. 125–151.
2. *Baker M.* (1988) Incorporation: A theory of grammatical function changing, Chicago, University of Chicago Press.
3. *Baker M.* (2011). On the syntax of surface-adjacency: The case of pseudo noun incorporation. Manuscript, Rutgers University. Available at http://www.rci.rutgers.edu/Bmabaker/PNI-adjacency.pdf.
4. *Baker M., Aranovich R., & Golluscio L. A.* (2005). Two types of syntactic noun incorporation: Noun incorporation in Mapudungun and its typological implications. Language, 138–176.
5. *Barrie M.* (2012). Noun Incorporation and the Lexicalist Hypothesis. Studies in Generative Grammar, 22, 235–61.

6. *Bogdanov A. V.* (2007) Generic vs. episodic interpretation of Russian nominalized compounds [Genericheskaya vs. epizodicheskaya interpretatsija nominalizatsij-kompozitov v russkom jazyke]. Structures and interpretations. Papers on theoretical ad applied linguistics by young researchers [Struktury i interpretatsii. Raboty molodyh issledovatelej po teoreticheskoy I prikladnoy lingvistike]. Moscow, pp.33–57.
7. *Booij G.* (2009). A constructional analysis of quasi-incorporation in Dutch. Gengo Kenkyu, 135, 5–27.
8. *Dictionary* of XVIII century Russian [Slovar' russkogo jazyka XVIII veka] (1984). Issue 13. Moldavskij—Naprokudit'. Leningrad: Nauka. Available at http://feb-web.ru/feb/sl18/slov-abc/13/sld03306.htm
9. *Janko-Trinitskaja N. A.* (2001) Word formation in contemporary Russian [Slovoobrazovanie v sovremennom russkom jazyke]. Moscow: Indrik.
10. *Muro A.* (2009) "Noun Incorporation: A New Theoretical Perspective." PhD Diss, Universita degli studi di Padova, Padua.
11. *NS-81* (1986) New items of Russian lexicon. Dictionary materials — 81 [Novoe v russkoj leksike. Slovarnye materially — 81]. Moscow.
12. *Ozerova E. G.* (1998) Compound words in children's speech. PhD Diss. [Slozhnye slova v detskoj rechi. Diss. … kandidata filologicheskih nauk]. Belgorod: Belgorod State University.
13. *Petrov A. V.* (2003) Explanatory and derivational dictionary of compound nests [Gnezdovoj tolkovo-slovoobrazovatel'nyj slovar' kompozitov]. Simferopol.
14. *Reinhart T., Siloni T.* (2004). Against the unaccusative analysis of reflexives. The unaccusativity puzzle, Oxford University Press, pp. 288–331.
15. *Reinhart T., Siloni T.* (2005). The lexicon-syntax parameter: Reflexivization and other arity operations. Linguistic inquiry, 36(3), pp. 389–436.
16. *Rice S., & Prideaux G.* (2012). Event-packing: the case of object incorporation in English. In Proceedings of the Annual Meeting of the Berkeley Linguistics Society (Vol. 17).
17. *Shvedova N. Ju.* (ed. in chief) (1980) Grammar of contemporary literary Russian [Grammatika sovremennogo russkogo literaturnogo jazyka]. Vol. I–II, Moscow: Nauka.
18. *Tuggy D.* (2005). Cognitive approach to word-formation. In Handbook of word-formation. Springer Netherlands. Pp. 233–265.
19. *Vinogradov V. V.* (1999) History of words [Istorija slov]. Moscow: IRJa RAN.
20. *Zemskaya E. A.* (2005) Word formation as activity [Slovoobrazovanie kak dejatel'nost']. Moscow: KomKniga.

СЕМАНТИЧЕСКИЕ ФАКТОРЫ ИЗМЕНЕНИЯ УПРАВЛЕНИЯ СУЩЕСТВИТЕЛЬНЫХ В СОВРЕМЕННОМ РУССКОМ ЯЗЫКЕ

Пестова А. Р. (pestova2012@gmail.com)

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

В статье рассматриваются новые варианты управления существительных, возникшие в современном русском языке под влиянием семантических факторов. Во-первых, это развитие лексического значения: у слов *авария*, *пируэт* и *концепция* оно способствовало появлению вариантов управления *авария с чем-л.*, *пируэт с чем-л. / вокруг чего-л.* и *концепция по чему-л.* Во-вторых, это действие семантической аналогии: слова *бум*, *фобия* и *востребованность* заимствуют синтаксическое поведение синонимов (*бум на что-л.* — по аналогии с *мода*, *спрос*; *фобия перед чем-л.* — по аналогии со *страх*, *фобия к чему-л.* — по аналогии со словами семантического класса ‘расположенность — нерасположенность к кому-чему-л.’ (*неприязнь*, *уважение* и т. п.), *востребованность в чём-л.* — по аналогии с *потребность*). Приведены результата опроса образованных носителей русского языка с их оценкой изучаемых вариантов управления.

Ключевые слова: именное управление, вариативность, норма, семантическое развитие, изменение по аналогии

SEMANTIC FACTORS OF THE NOUN GOVERNMENT CHANGES IN THE MODERN RUSSIAN LANGUAGE

Pestova A. R. (pestova2012@gmail.com)

V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper deals with the new variants of noun government in the modern Russian language. These variants are accounted for by certain semantic factors, such as development of meaning and semantic analogy. Due to the development of meaning the nouns *avarija* 'accident', *piruèt* 'pirouette' and *kontseptsija* 'conception' get new variants of government (*avarija s* + instrumental, *piruèty s* 'with'+ instrumental / *vokrug* 'around' + genitive, *kontseptsija po* 'on'+ dative). By semantic analogy the nouns *bum* 'boom', *fobija* 'phobia' and *vostrebovannost'* 'demand' adopt syntactic features of their synonyms. *Bum* 'boom' accepts a PP *na* 'on'+ accusative (by analogy with words *moda* 'fashion' or, *spros* 'demand'). *Fobija* 'phobia' governs either *pered* 'before'+ instrumental (by analogy with the noun *strakh* 'fear'), or *k* + dative (by analogy with words belonging to the semantic group 'attitude (positive or negative) toward smb, or smth', e.g. *neprijazn'* 'dislike', *uvazhenije* 'respect'). *Vostrebovannost'* 'demand' governs *v* 'in'+ prepositional case by analogy with the semantically similar word *potrebnost'* 'need'. Well-educated native speakers were asked to fill in questionnaires containing phrases with these variants. Their answers are presented.

Keywords: noun government, variation, norm, semantic development, analogical change

Предметом статьи являются новые варианты управления существительных, появившиеся в современном русском языке под действием семантических факторов.

О связи семантики и синтаксиса, об унификации синтаксиса под давлением семантики, о давлении синтаксиса на семантику см. [Апресян 1967: 23–36]. В этой работе выделено большое количество семантических разрядов глаголов с однотипными синтаксическими свойствами [118 и сл.]; см. также [Ицкович 1968: 20]. Тем не менее, для большого числа слов управление определяется индивидуально, *в словарном порядке*[1]. Однако полного словаря управления современного русского языка: ни прескриптивного, ни дескриптивного — не существует. Именно поэтому столь частотны случаи нарушения норм управления и связанные с этим колебания и вариативность норм.

Семантические факторы можно условно разделить на два типа: с одной стороны, это **развитие лексического значения** управляющего слова, с другой, — действие **семантической аналогии**. В результате семантической аналогии слово в одном из своих значений, *не меняя его*, приобретает синтаксическое оформление либо под влиянием другого своего значения, либо под влиянием синонима. Во многих случаях эти факторы действуют совокупно.

Материалом исследования послужили примеры из Национального корпуса русского языка (далее — НКРЯ), печатных и электронных СМИ, из блогов и устной речи (публичной и бытовой).

---

[1] Л. В. Щерба замечал: «Следует предостеречь от общераспространенного предрассудка, будто управление слов определяется грамматикой; на самом деле оно чаще всего оказывается принадлежностью каждого слова, а поэтому является фактом словаря» [Щерба 1947: 96].

1. Изменения в семантической структуре главного слова

Рассмотрим действие этого семантического фактора на примере слов *авария*, *пируэт* и *концепция*.

АВАРИЯ

Существительное *авария* толкуется как «повреждение, выход из строя какого-л. механизма, машины, транспортного средства и т. п. во время действия, движения» [МАС[2]]. Согласно иллюстративным примерам в словарях, оно управляет родительным падежом: *а. судна* [СУш; Тихонов 2001]. Встречается, кроме того, не зафиксированное словарями управление *с чем-л.* Поиск по НКРЯ даёт следующие результаты: 216 вхождений *а. чего-л.*, 22 употребления *а. с чем-л*[3].

Как показывает наблюдение над современным узусом, в указанном словарями значении возможны оба варианта управления: *а. поезда / с поездом, судна / с судном*:

авария чего-л.	***авария с чем-л.***
(1) *Железнодорожное сообщение на месте* ***аварии поезда*** *в Бурятии частично восстановлено*[3].	(2) *Причины* ***аварии с поездом*** *«Невский экспресс» … пока не известны*[4].
(3) *При* ***аварии судна*** *в море вылилась нефть*[5].	(4) *31 июля 2009 близ берегов Швеции произошла очень серьезная* ***авария с судном*** *«Langeland»*[6].

В разговорной речи у этого существительного развивается переносное метафорическое значение. *Авария* может означать не только повреждение механизмов и машин, но и разного рода другие повреждения (например, телесные: *а. с ногой*, *с носом*; предметов одежды: *а. со штанами*, *с платьем*). Это значение отмечено в «Толковом словаре русской разговорной речи», работа над первым томом которого ведётся в Отделе современного русского языка Института русского языка им. В. В. Виноградова РАН. В рукописи этого словаря *авария* толкуется как «непредвиденная бытовая неприятность» [ТСРР]. В этом значении

---

[2] Здесь и далее расшифровку аббревиатур см. в приложении «Словари».

[3] Поиск осуществлялся в апреле 2013 г. Было произведено отсеивание неадекватных запросу примеров (например, *авария такого масштаба; аварии, жертвы; авария двадцатилетней давности*). Первое вхождение беспредложного управления Р. п. — 1903 г. (*а. судов*), предложного *с чем-л.* — 1907 г. (*а. с летательным аппаратом*).

[4] «Вести.Ru», 03.03.2013.

[5] Новостной сайт, 2009.

[6] НКРЯ, 2003.

[7] Новостной сайт, 2009.

возможно управление **только** конструкцией *с чем-л.* Так, встречаются *а. с носом, с пальцем* и *а. со штанами, с одеждой,* но не встречаются **а. носа, пальца* и **а. штанов, одежды*:

(5) ***Авария с носом.*** *Разбили его, да ещё и синяк на полщеки*[8].

(6) *Что ж делается, сама с сыном вчера из больницы… А тут вот и у вас читаю…* ***авария с пальцем***[9].

(7) *А сейчас позвонил Женечка и сказал, что у него* ***авария со штанами*** *— молния совсем сломалась*[10].

(8) *Яркий пример того, как с помощью запланированной* ***аварии с одеждой*** *можно восстановить увядающую карьеру. Не особо популярная в последнее время модель Мария Фаулер выходила с подругой из ночного клуба в Лондоне, когда у нее вдруг внезапно лопнуло платье*[11].

ПИРУЭТ

Пируэт, согласно словарям, — это «полный круговой поворот всем телом на носке одной ноги (в танце)» [МАС]. В этом значении оно употребляется без управления: *Балерина сделала п.* В переносном значении, которое в словарях ещё не зафиксировано, но уже активно употребляется, это слово (чаще в форме множественного числа) управляет конструкцией *с чем-л.*:

(9) *При таком драконовском законе* ***пируэты с*** *муниципальными* ***рынками*** *станут невозможны*[12].

(10) ***Пируэты с евро*** [о выходе Греции из зоны евро] *означают, что другие страны рискуют оказаться в водовороте дальнейшего кризиса*[13].

В этом значении под пируэтами подразумеваются некоторые, обычно замысловатые, затейливые действия, совершаемые с определённой, часто неблаговидной целью. В этом значении *пируэт* сближается с переносным значением слова *манёвр*: «действие, приём с целью обмануть, перехитрить кого-л., добиться чего-л.» [БТС].

---

[8] Блоги, 2009.

[9] Блоги, 2008.

[10] Блоги, 2009.

[11] Блоги, 2012.

[12] НКРЯ, 2003. Это первое вхождение в НКРЯ данного управления.

[13] Русская служба Би-би-си, 15.05.2012.

Возможно, кроме того, управление рассматриваемого слова (обычно также в форме множественного числа) предложно-падежным сочетанием *вокруг чего-л.*[14]:

(11) *Сама Америка, как известно, несмотря на все угрожающие* ***пируэты вокруг*** *ядерной* ***программы*** *Тегерана, нападать на Иран не решилась*[15].

(12) *Как … объяснить* ***пируэты вокруг тюрьмы*** *Гуантанамо, закрыть которую обещал во время выборной гонки Обама?*[16].

Конструкция *вокруг чего-л.* вносит значение широкого обсуждения, общественного резонанса ввиду скандальности описываемых событий или ситуаций. Это согласуется с контекстом фразы. В примере (11) это слова *как известно* и *угрожающие пируэты*, в (12) — *предвыборная гонка*, которая обычно сопровождается агитацией и предполагает дискуссии, споры и освещение в СМИ.

КОНЦЕПЦИЯ

У существительного *концепция* словари указывают два значения: «система связанных между собой и вытекающих один из другого взглядов на то или иное явление» и «основная мысль, идея произведения, сочинения и т. п.» [БТС]. В этих значениях оно, согласно иллюстративным примерам в словарях, может управлять только беспредложным Р. п.: *к. общественного развития, к. романа* [СШ, БТС, ИТСИС]. В первом значении *концепция* употребляется, кроме того, с не зафиксированными словарями вариантами *о чём-л.* и *по чему-л.*[17]:

(13) ***Концепция о формировании*** *новой исторической общности людей — советского народа действительно имела под собой серьезные основания*[18].

(14) *Нужна была и более продуманная и развернутая* ***концепция по истории*** *России и Русской революции 1917 года*[19].

---

[14] В НКРЯ это управление не зафиксировано.

[15] Электронная газета Forum.msk.ru, 01.11.2007.

[16] Новостной сайт, 05.06.2009.

[17] Первые вхождения этих вариантов в НКРЯ — *о чём-л.* — 1928 г. (*к. о ней* [о поэзии]), *по чему-л.* — 1948 г. (*концепции по гносеологии, этике, эстетике*).

[18] «Российская газета», 15.05.2003.

[19] НКРЯ, 2002.

В современном употреблении у этого слова развивается метонимическое значение «официальный документ, содержащий систему взглядов на что-л.». В этом значении концепция управляет конструкцией *по чему-л.*[20]:

(15) *На рассмотрение кабинета поступила* ***концепция по*** *корпоративному* ***управлению*** *и поправки в Административно-процессуальный кодекс РФ*[21].

(16) *Разработаны такие документы, как* ***концепция по снижению*** *злоупотребления алкоголем*[22].

Конструкция с предлогом *по* при словах, обозначающих письменные документы, начала активизироваться в середине XX в.: *листок по учету кадров, проект по реке* [Шведова 1966: 43]. В конце XX столетия размах её употребления значительно расширился [Гловинская 1996: 248–252].

В новом значении встречается также управление конструкцией *о чём-л.*:

(17) *…законопроект в существующем виде не отвечает требованиям, провозглашенным в президентской* ***концепции о реформировании*** *госслужбы*[23].

2. Действие семантической аналогии

Действие семантической аналогии на синтаксическое поведение слова в современном русском языке можно проиллюстрировать существительными *бум*, *фобия* и *востребованность*. Подчеркнём, что, в отличие от рассмотренных выше случаев, изменений в семантической структуре этих слов не происходит.

БУМ

Согласно словарям, *бум* имеет два значения: «кратковременный подъём, резкое оживление в промышленности, торговле и других сферах хозяйства» и «шумиха, искусственное оживление вокруг какого-н. события» [ИТСИС]. В первом значении оно традиционно употребляется без управляемых слов, обычно с согласуемым прилагательным: *нефтяной б., книжный б.*

---

[20] Первое вхождение управления существительного *концепция* в новом значении предложно-падежной конструкцией *по чему-л.* в НКРЯ датируется 1996 г. (*к. по предотвращению и урегулированию конфликтов*).

[21] «Независимая газета», 14.02.2006.

[22] «Российская газета», 06.10.2010.

[23] НКРЯ, 2003. Это первое вхождение в НКРЯ указанного варианта управления. Отметим, что в официальном названии концепции употребляется беспредложное управление родительным падежом: «**Концепция реформирования** системы государственной службы РФ».

В современном употреблении встречаются, однако, примеры с конструкцией *на что-л.*, по аналогии с синонимичными словами *спрос, мода* (*на что-л.*)[24]:

(18) *К 2050 году* ***бум на*** *английский* ***язык*** *окончится, и индустрия обучения английскому языку станет жертвой собственного успеха*[25].

(19) *Как говорят сотрудники загса,* ***бум на*** *цветные* ***платья*** *пришелся на 90-е годы*[26].

Во втором, переносном значении слово *бум* испытывает также влияние синонимичных существительных *шум*, *шумиха*, *скандал* и управляет предложно-падежным сочетанием *вокруг чего-л.*[27]:

(20) *Некоторые предсказывают, что в России не будет такого* ***бума вокруг книг*** *о Гарри Поттере, как в Англии, Америке*[28].

(21) *Для болельщиков* ***бум вокруг*** *фигурного катания начался с чемпионата Европы 1965 года во Дворце спорта в Лужниках*[29].

ФОБИЯ

Существительное *фобия* ‘навязчивое состояние страха’, согласно словарям, употребляется без управляемых слов. В современной речи оно приобретает управляющие свойства, притом в управлении этого слова из-за действия разных семантических аналогий наблюдаются колебания.

С одной стороны, под влиянием слов семантического класса ‘расположенность–нерасположенность к кому-чему-л.’ (*любовь, уважение, неприязнь* и т. п.) [Апресян 1995: 545], *фобия* управляет предложно-падежной конструкцией *к кому-чему-л.*:

(22) *У премьера скоро, наверное, появится стойкая* ***фобия к*** *подобным* ***мероприятиям*** *и многотысячным* ***залам*** *и* ***трибунам***[30].

---

[24] Первое вхождение этого управления в НКРЯ — 2000 г. (*б. на высшее образование*).

[25] «Известия», 24.12.2007.

[26] «Комсомольская правда», 20.12.2010.

[27] Первое вхождение в НКРЯ — 1976 г. (*б. вокруг сверхлёгких самолётов*).

[28] «Комсомольская правда», 06.12.2001.

[29] «Труд-7», 24.10.2003.

[30] «Эхо Москвы», 26.11.2011.

(23) *У меня развилась* ***фобия к кино*** *на русском, потому что если это не сериал, то непременно поросший мхом фильм из 70-х*[31].

С другой стороны, действует аналогия со словом *страх*, которое управляет конструкцией перед *кем-чем-л.*[32]:

(24) *Причины различных слухов и* ***фобий перед запуском*** *коллайдера — это вполне естественный страх человека перед неизвестным*[33].

(25) *Врачи диагностировали британке* ***фобию перед овощами***[34].

ВОСТРЕБОВАННОСТЬ

В современном употреблении смешивается управление семантически близких существительных *востребованность* и *потребность*. Толкование слова *востребованность* приводится только в БТС «необходимость, потребность в ком-, чём-л.» и иллюстрируется примером с беспредложным родительным падежом: *в. новых технологий*. В текстах встречается и вариант управления предложно-падежной конструкцией *в чём-л.*[35]:

(26) *МАРХИ — вуз прекрасный и знаковый, и выпускники его известны не только в архитектуре, но и в литературе и музыке, и — что самое главное —* ***востребованность в специальности*** *вряд ли может вызывать сомнения*[36].

(27) *По итогам 2011 года в столице отмечено увеличение* ***востребованности*** *в профессиональных* ***бухгалтерах***[37].

При этом далее в тексте (27) употребляется управление, представленное в БТС: *За июль 2001 г. наивысшим показателем по* ***востребованности*** *профессиональных* ***бухгалтеров*** *в Москве наблюдается в сфере торговли.*

Видимо, вариант *в чём-л.* появился под влиянием синонима *потребность*. Как показывают современные примеры, влияние это взаимное: у слова *потребность,* в свою очередь, появился вариант управления беспредложным Р. п.:

---

[31] «Комсомольская правда», 14.01.2011.

[32] Оба рассматриваемых варианта управления в НКРЯ не зафиксированы.

[33] Сайт радио «Эхо Москвы», 10.09.2008.

[34] «Аргументы и факты», 01.08.2012.

[35] В НКРЯ эти варианты датируются 2000 (*в чём-л.*) и 2002 г. (*чего-л.*). Зафиксированного БТС варианта управления беспредложным Р. п. — 64 вхождения (при ручном отсеивании неадекватных запросу примеров), предложным *в чём-л.* — 6 вхождений.

[36] «Известия», 21.11.2012.

[37] Новостной сайт, 2012.

(28) *Госстрой выпустил свое первое постановление с «ориентирами» для регионов, исходя из которых рассчитывались* ***потребности помощи*** *каждой территории из федерального бюджета*[38].

(29) *Рассчитана* ***потребность воды*** *для населения в случае отключения водозаборов*[39].

3. Новые варианты управления в оценках носителей языка

Мы решили проверить, что из приведенных употреблений является случайной ошибкой, а что можно интерпретировать как росток новой языковой нормы. Следует отметить, что все рассматриваемые словосочетания достаточно распространены в современном русском языке. Многочисленные примеры их употребления можно найти в НКРЯ, в газетах, на новостных сайтах, услышать в телевизионных передачах. Согласно утверждению В. А. Успенского, такие варианты можно было бы отнести к новой норме: «…понятие **нормы** имеет в своей основе статистику: если “так говорит” или “так понимает” абсолютное большинство носителей языка, то это и есть **норма**» [Успенский 2006: 539]. Значение фактора распространённости, однако, нельзя преувеличивать: как отмечают многие исследователи нормы, распространённой может быть и явная ошибка (см., например, [Ицкович 1968: 40; Граудина 1980: 69; Крысин 2006: 303–304]). Помимо него нужно учитывать и другие критерии нормы: **системный** (непротиворечие явления системе языка) и **эстетический** (общественное одобрение и признание явления нормативным) [Семенюк 1990: 338]. Исследуемые варианты не противоречат системе русского языка, следовательно, **системному** критерию они соответствуют. Таким образом, открытым остаётся вопрос о признании данного явления нормативным образованными носителями русского языка, т. е. об **эстетическом** критерии.

Чтобы выяснить, насколько далеко отстоят от нормы изучаемые словосочетания, был проведён опрос. На материале собранных примеров была составлена анкета. Образованным носителям русского языка предлагалось оценить правильность предложений, содержащих словосочетания с новыми вариантами управления. Было опрошено 110 человек в возрасте 18–80 лет: студенты, аспиранты, специалисты, работающие в разных областях, пенсионеры. Анкета содержала следующие предложения:

1. ***Бум на*** *второе высшее* ***образование*** *начался в нашей стране в 1997 году.*
2. *У меня произошла* ***авария с компьютером****, и я потерял адрес этого человека.*
3. *При таком драконовском законе* ***пируэты с*** *муниципальными* ***рынками*** *станут невозможны.*

---

[38] «Российская газета», 28.08.2003.

[39] «Комсомольская правда», 09.08.2010.

4. *Эксперты объясняют апатию* ***фобией перед*** *радикальными* ***реформами*** *и* ***потрясениями***.
5. *По итогам 2011 года в столице отмечено увеличение* ***востребованности в*** *профессиональных* ***бухгалтерах***.
6. *Грузия два месяца назад приняла новую* ***концепцию по обороне*** *и* ***безопасности***.

Самым «правильным» по результатам опроса оказалось словосочетание ***бум на образование***: его не исправил 61% информантов. Другие варианты управления предлагают 18% опрошенных[40]: *бум образования / бум получения образования* (16%), *бум вокруг образования* (2%). Остальные (17%) заменили *бум на образование* на синонимичные словосочетания: *мода, спрос на образование, популярность образования, ажиотаж вокруг образования / на образование, шумиха вокруг образования, интерес к образованию, подъём образования; образовательный бум; образование стало популярным*. Несколько носителей русского языка (примерно 6%) не смогли исправить это предложение, оставив комментарии: «“Бум на” — непонятное управление какое-то, разговорно» (Ж, 25, переводчица); «Бум случился. И “бум на” не нравится» (Ж, 19, студентка 2 курса филфака МГУ).

Самым «неправильным» носители русского языка считают словосочетание ***авария с компьютером***: всего 35% опрошенных оставили такое управление без изменения. Остальные (55%) предлагали заменить это словосочетание на синонимичные выражения: *поломка компьютера, компьютерный сбой / сбой в компьютере, проблемы с компьютером, неполадки с компьютером; компьютер сломался / вышел из строя / испортился / полетел / сгорел*. Некоторые опрошенные (6%) предлагали другие варианты управления: *авария на компьютере / в компьютере / компьютера*.

Примерно посередине шкалы нормативности располагаются *концепция по обороне, фобия перед реформами* и *востребованность в бухгалтерах*. Словосочетание ***концепция по обороне*** является правильным по оценке 46% носителей русского языка. Столько же опрошенных предлагают другие варианты управления: 37% — беспредложным родительным (*концепция обороны*), остальные 9% — различными предложно-падежными конструкциями (*концепция об обороне / по вопросам обороны / в сфере обороны*). Около 5% предлагают заменить исходное словосочетание на синонимичное: *доктрина / конвенция об обороне, стратегия / программа по обороне, политика в сфере обороны, военная доктрина*.

47% информантов не исправили предложение со словосочетанием ***фобия перед реформами***. Остальные предлагали либо вариант управления беспредложным родительным: *фобия реформ* (18%), либо синонимичное словосочетание: *страх / боязнь перед реформами, страх / боязнь реформ* (33%).

Правильность словосочетания ***востребованность в бухгалтерах*** не вызвала сомнений у 47% носителей русского языка. 37% заменили предложно-падежное управление на беспредложное: *востребованность бухгалтеров*.

---

[40] Сумма процентов иногда составляет больше 100, т. к. учитывались все предложенные одним информантом варианты.

Остальные (20%) предложили синонимичные замены: *потребность в бухгалтерах, спрос на бухгалтеров, профессия бухгалтеров стала более востребованной*.

Больше всего затруднений вызвало у информантов словосочетание ***пируэты с рынками***: его не могли исправить, признавая при этом ненормативным, 18% опрошенных. Процент тех, кто счёл его правильным, довольно высок: 59%, однако этот показатель, вероятно, не вполне корректен, учитывая множество комментариев типа: «Словосочетание “пируэты с рынками” кажется не совсем грамотным, но я не знаю, как было бы правильно, поэтому +» (М, 28, юрист); «+. Наверчено, но как лучше сказать, не знаю (Ж, 26, экономист)» и мн. др. В качестве исправлений носители предлагали синонимичные словосочетания (21%): *манипуляции с рынками / рынками / над рынками, мошенничество с рынками, махинации с рынками, фокусы с рынками*, а также управление другими предложно-падежными конструкциями (6%): *пируэты на рынках / в рынках / над рынками*.

Этот небольшой опрос показывает незавершенность данного процесса даже у слов, где воздействие оказывают семантические факторы. Все они являются точками языкового напряжения, нормы управления которых расшатаны. Пока мы можем говорить только о некоторых тенденциях развития, но не о смене норм. Носители, сталкиваясь с этими точками напряжения, действуют по-разному: либо склоняются к новому варианту (*бум на что-л.*), либо колеблются между традицией и инновацией (*концепция чего-л. / по чему-л., фобия / страх перед чем-л., востребованность /потребность в чём-л.*), либо пытаются найти третий, «более удачный» вариант (*авария на компьютере, пируэты над рынками*). Многие комментарии информантов демонстрируют их отношение к разговорной речи, канцеляриту и языку СМИ. Обращение к таким предварительным установкам при оценке нормативности подчёркивает незавершённый характер становления новых вариантов управления.

Литература

1. *Апресян Ю. Д.* Избранные труды, том II. Интегральное описание языка и системная лексикография. — М.: Языки русской культуры, 1995. — 767 с.
2. *Апресян Ю. Д.* Экспериментальное исследование семантики русского глагола. — М.: Наука, 1967. — 252 с.
3. *Гловинская М. Я.* Активные процессы в грамматике (на материале инноваций и массовых языковых ошибок) // Русский язык конца XX столетия (1985–1995). — М.: Языки русской культуры, 1996. — С. 237–304.
4. *Граудина Л. К.* Вопросы нормализации русского языка. Грамматика и варианты. — М.: Наука, 1980. — 288 с.
5. *Ицкович В. А.* Языковая норма. — М.: Просвещение, 1968. — 92 с.
6. *Крысин Л. П.* Языковая норма в проекции на современную речевую практику // Русский язык сегодня. Вып. 4. Проблемы языковой нормы. — М.: Ин-т рус. яз. им. В. В. Виноградова РАН, 2006. — С. 294–311.

7. *Семенюк Н. Н.* Норма языковая // Лингвистический энциклопедический словарь. — М.: Советская энциклопедия, 1990. — С. 337–338.
8. *Успенский В. А.* Субъективные заметки о неправильной норме // Русский язык сегодня. Вып. 4. Проблемы языковой нормы. — М.: Ин-т рус. яз им. В. В. Виноградова РАН, 2006. — С. 537–571.
9. *Шведова Н. Ю.* Активные процессы в современном русском синтаксисе (словосочетание). — М.: Просвещение, 1966. — 156 с.
10. *Щерба Л. В.* Преподавание иностранных языков в средней школе. Общие вопросы методики. — М.: АПН РСФСР, 1947. — 304 с.

Словари

1. *БТС* — Большой толковый словарь русского языка / Под ред. С. А. Кузнецова. Электронный ресурс. Режим доступа: http://www.gramota.ru/slovari/info/bts/, свободный. Загл. с экрана. Данные соответствуют 28.01.2013. Словарь опубликован в авторской редакции 2009 года.
2. *ИТСИС — Крысин Л. П.* Иллюстрированный толковый словарь иностранных слов. — М.: Эксмо, 2011. 864 с.
3. *МАС* — Словарь русского языка: В 4-х т. / Под ред. А. П. Евгеньевой. — М.: Рус. яз., 1999.
4. *СУш* — Толковый словарь русского языка: В 4 т. / Под ред. Д. Н. Ушакова. — М.: Гос. ин-т «Сов. энцикл.»; ОГИЗ; Гос. изд-во иностр. и нац. слов, 1935–1940.
5. *СШ* — Толковый словарь русского языка с включением сведений о происхождении слов / Под ред. Н. Ю. Шведовой. — М.: Азбуковник, 2008.
6. *Тихонов 2001* — Комплексный словарь русского языка / Под ред. А. Н. Тихонова. — М.: Рус. яз., 2001.
7. *ТСРР* — Толковый словарь русской разговорной речи / Под ред. Л. П. Крысина. — М.: Институт рус. яз. им. В. В. Виноградова, 2010. — 346 с.

References

1. *Apresjan Ju. D.* (1995), Izbrannyje trudy, tom II. Integral'noje opisanije jazyka i sistemnaja leksikografija [Selected Works, Volume II. Integral Description of Language and System Lexicography]. Jazyki russkoj kul'tury, Moscow.
2. *Apresjan Ju. D.* (1967), Èksperimental'noje issledovanije semantiki russkogo glagola [Experimental Study of Russian Verb Semantics]. Nauka, Moscow.
3. *Glovinskaja M. Ja.* (1996), Aktivnyje protsessy v grammatike (na materiale innovatsij i massovyh jazykovyh oshibok) [Active Processes in Grammar (Based on Innovations and Linguistic Deviations)], in Russkij jazyk kontsa XX stoletija (1985–1995) [The Russian Language in the End of the XXth Century (1985–1995)]. Jazyki russkoj kul'tury, Moscow, pp. 237–304.
4. *Graudina L. K.* (1980), Voprosy normalizatsii russkogo jazyka. Grammatika i varianty [Problems of the Russian Language Normalization. Grammar and Variants]. Nauka, Moscow.
5. *Itskovich V. A.* (1968), Jazykovaja norma [Language Norm]. Prosveshchenije, Moscow.
6. *Krysin L. P.* (2006), Jazykovaja norma v projektsii na sovremennuju rechevuju praktiku [Language Norm and Its Reflection in Modern Usage], in Russkij jazyk segodnja [The Russian Language Today], no. 4, pp. 294–311.
7. *Semenjuk N. N.* (1990), Norma jazykovaja [Language Norm], in Lingvisticheskij Èntsiklopedicheskij Slovar' [Linguistic Encyclopaedia]. Sovetskaja èntsiklopedija, Moscow, pp. 337–338.
8. *Uspenskij V. A.* (2006), Sub"ektivnyje zametki o nepravil'noj norme [Subjective Notes on Wrong Norm], in Russkij jazyk segodnja [The Russian Language Today], no. 4, pp. 537–571.
9. *Shcherba L. V.* Prepodavanie inostrannyh jazykov v srednej shkole. Obshchije voprosy metodiki [Teaching Foreign Languages at Secondary School. Basic Methodological Problems]. APN RSFSR, Moscow.
10. *Shvedova N. Ju.* (1966), Aktivnyje protsessy v sovremennom russkom sintaksise (slovosochetanije) [Active Processes in the Modern Russian Syntax (Word-Combination)]. Prosveshchenije, Moscow.

ЛИТУРАТИВЫ В РУССКОМ ИНТЕРНЕТЕ: СЕМАНТИКА, СИНТАКСИС И ТЕХНИЧЕСКИЕ ОСОБЕННОСТИ БЫТОВАНИЯ[1]

Пиперски А. Ч. (apiperski@gmail.com),
Сомин А. А. (somin@tut.by)

Российский государственный гуманитарный университет, Москва, Россия

Ключевые слова: зачёркивание, литуратив, семантика, прагматика, коммуникативные постулаты, синтаксис, классификация, блоги, социальные сети

STRIKETHROUGH ON THE RUSSIAN WEB: SEMANTICS, SYNTAX AND TECHNICAL ISSUES

Piperski A. Ch. (apiperski@gmail.com),
Somin A. A. (somin@tut.by)

Russian State University for the Humanities, Moscow, Russia

This article deals with the use of strikethrough (also known as liturative) on the Russian Web. We summarize two previous attempts to classify the instances of liturative and propose a new classification based on three binary syntactic and semantic features. This classification allows distinguishing six main types of lituratives (two other theoretically possible types are not attested). The features in question are [± substitution] (whether or not the stikethrough text serves as a substitute for the normal text), [± violation of conversational maxims] and [± negative attitude towards the speaker] (whether or not the strikethrough text could possibly cause a negative attitude towards its author). All findings are illustrated with real examples extracted from Russian blogs. In the last section of the paper, we discuss technical issues of using strikethrough on the Web and its implementation on various websites (LiveJournal, Mail.ru, Yandex, Gmail, Facebook, VKontakte). We attempt to explain why the popularity of strikethrough is gradually decreasing.

Keywords: strikethrough, liturative, semantics, pragmatics, conversational maxims, syntax, classification, blogs, social networking services

[1] Работа выполнена при поддержке Программы стратегического развития РГГУ. Авторы выражают благодарность В. И. Матиссен-Рожковой за предоставление примеров (12), (14), (18), (21), (22).

1. Введение

В ситуации Интернет-коммуникации возник ряд новых языковых средств, одним из которым стало зачёркивание текста, например:

(1) *Друзья, товарищи, любители стихов, поклонники всего, что звучно, стройно, нежно, поделитесь секретнейшей информацией — как вы ~~дошли до жизни такой, что надобны вам только эти ваши~~ выбирали свои благозвучные ~~кликухи позорные~~ никнеймы?*[2] (из блога, 2005)

(2) *Тут Феликс слегка просветлел лицом, впервые с начала ~~экзекуции~~ чествования* (из блога, 2006)

Этот приём, получивший название «литуратив» (термин введён в работе [Гусейнов 2008]) заслуживает тщательнейшего лингвистического изучения. Цель настоящей статьи состоит в том, чтобы, во-первых, предложить новый подход к классификации литуративов, а во-вторых — исследовать связь между распространением литуративов и распространённостью тех или иных средств Интернет-коммуникации.

2. Обзор литературы

На данный момент имеется две основных работы, посвящённых использованию современных русских литуративов: пионерская статья Г. Ч. Гусейнова «Неполная коммуникация в блогосфере: эрративы и литуративы» [Гусейнов 2008] и статья Н. Н. Занегиной «~~Я этого не говорил~~: о литуративах, зачеркиваниях или мнимых текстах» [Занегина 2009], дополняющая и уточняющая классификацию литуративов, предложенную в [Гусейнов 2008].

В [Гусейнов 2008] описание семантики литуративов выполняется с использованием самого явления, а в [Занегина 2009] предлагается более традиционная с точки зрения способа описания классификация (или, точнее, две: по прагматическим и синтаксическим параметрам). Г. Ч. Гусейнов выделяет такие семантические типы литуративов:

- «~~Всё-то вам, дуракам, приходится объяснять, а то ведь сами бы не поняли, ага!~~»;
- «~~На самом деле, я хотел сказать вот это!~~»;
- «~~Всё дозволено~~»;
- «~~Хоть я и понимаю, что так говорить не принято, современные технологии позволяют мне ненадолго и почти безболезненно обойти требования приличий (политической корректности, логики и т. п.), но потом вернуться к привычной повестке дня~~»;
- «~~Хоть я и понимаю, что кому-нибудь хотелось бы сказать *это*, по здравом размышлении нельзя не признать вот *это*~~».

[2] В примерах сохранена авторская орфография и пунктуация.

В свою очередь, Н. Н. Занегина в работе [Занегина 2009] выделяет пять основных значений литуративов:

- «не буду говорить непринятое»;
- «не буду говорить неправду»;
- «не буду говорить правду»;
- «не буду говорить банальности»;
- «не буду говорить».

Ограниченный объём данной работы не позволяет подробно остановиться на достоинствах и недостатках рассматриваемых статей. Однако существенно следующее: в обеих статьях предложены в целом удачные классификации, основным достоинством которых является «прозрачное» описание семантики литуратива с помощью лишь немного формализованного языка — как в классических толковых словарях. Но в то же время это достоинство является и недостатком: с помощью простых языковых описаний можно выделить сколь угодно много классов без обобщения; кроме того, в обоих случаях отсутствует какое-либо общее основание классификации и понятные противопоставления пунктов друг другу. Безусловно, такой подход к описанию семантики лингвистических явлений имеет свою историю и традицию (ср., к примеру, классические описания семантики латинских падежей, выделяющие для каждого падежа список не связанных между собой функций — genitivus partitivus, genitivus subjectivus, genitivus objectivus, genitivus pretii и т. д.) и, в принципе, достаточно успешно применяется и в современной лингвистике. Если провести аналогию с фонетикой, предложенные классификации, по сути, являются простым перечислением звуков. Однако более системный подход требует более чёткого описания явления в виде системы признаков и оппозиций — как в таблице с артикуляционной классификацией. Необходимо найти инвариант, объединяющий в себе семантику литуративов всех типов, и проанализировать параметры варьирования.

3. Семантика литуративов: инвариант и его разновидности

Семантическим и прагматическим инвариантом литуратива, на наш взгляд, является нарушение норм коммуникации: зачёркнутые фрагменты в большинстве случаев так или иначе нарушают общепринятые правила, создавая тем самым стилистический эффект. Мы выделяем два основных типа таких нарушений: нарушение постулатов коммуникации и нарушение стандартных стратегий формирования образа автора.

3.1. Литуративы и нарушение постулатов коммуникации

Наиболее известным сводом правил, выполнение которых необходимо для эффективной коммуникации, являются так называемые постулаты Грайса [Grice 1975, Грайс 1985]. Современная прагматика во многом уточнила

наблюдения Грайса, однако они до сих пор лежат в основе всех списков коммуникативных правил (ср., напр., [Mey 2001: 67–91]).

Впрочем, буквальное следование постулатам Грайса в реальной коммуникации невозможно, и примеры нарушения этих постулатов с теми или иными коммуникативными целями приводятся ещё в [Grice 1975: 52–56, Грайс 1985: 229–234]. Давно известно, что абсолютное большинство проявлений юмора основаны на нарушении этих постулатов [Attardo 1994: 271]. Литуративы оказываются средством, которое позволяет одновременно нарушить тот или иной коммуникативный постулат, и сымитировать его ненарушение. При помощи литуративов автор делает вид, что необходимое для его коммуникативной задачи нарушение постулатов не было совершено, т. е. зачёркнутый текст якобы не был написан (произнесён). Таким образом, незачёркнутый текст приближается к «грайсовому идеалу». На связь литуративов с постулатами Грайса впервые было указано в [Занегина 2009], однако там эта идея не получила развития.

Один из основных принципов коммуникации, который часто нарушают литуративы, является постулат истинности, предложенный Г. П. Грайсом: «Старайся, чтобы твоё высказывание было истинным» (а точнее, «Не говори того, что ты считаешь ложным» и «Не говори того, для чего у тебя нет достаточных оснований») [Grice 1975: 46, Грайс 1985: 222–223]. Здесь, однако, часто в противоречие вступают «абсолютная истина» (описание с точки зрения объективной реальности) и «авторская истина» (описание с точки зрения автора): во многих случаях автор, безусловно, зная «абсолютную истину», более подходящей в данной ситуации считает собственную переносную характеристику события, объекта и т. п.. В ситуации выбора из двух истин той, которую необходимо вербализовать согласно постулату истинности, возникает конфликт, и это приводит к тому, что некоторые авторы зачёркивают «абсолютную истину», другие же — «авторскую», ср.:

(3) *В задачи трамбовщика якобы входит помогать пассажирам ~~втиснуться~~ сесть в поезд в часы-пик.*. (из блога, 2007; зачёркнута «авторская истина»)

(4) *Раньше я писала в жежешечку после работы — едешь себе по ~~Ленинградке~~ пробке и набираешь на телефон буковки* (из блога, 2012; зачёркнута «абсолютная истина»).

Отметим, что в обоих случаях зачёркнутой могла оказаться как одна, так и другая истина.

Нечасто, но всё же встречаются случаи, когда внешний наблюдатель не может однозначно определить какая из двух истин — литуратив или его субститут — является «авторской», а какая — «абсолютной»: это многообразие значений, среди прочих особенностей, создаёт возможность двоякого понимания литуратива:

(5) *Взял телефончег, созвонюсь на неделе на другую съемку, уже не под заказ а под свои ~~сексуальные~~ твор<ч>еские нужды* (из блога, 2007).

В достаточно редких случаях зачёркиванию подлежит заведомо ложная пропозиция, нарушающая постулат истинности и описывающая лишь желания и намерения говорящего, но не реальный факт:

(6) *Ученик после занятия спросил, используют ли на самом деле англичане артикли, или они есть только в учебниках. <…> Я ~~грязно выматерилась~~ пообещала в следующий раз показать ему отрывок из какого-нибудь фильма, где бы употребляли артикль the* (из блога, 2010).

Впрочем, постулат истинности — не абсолютная аксиома. Более того, в естественной коммуникации люди постоянно нарушают этот постулат ради художественного эффекта. Фактически, все тропы (метафора, метонимия, гипербола, ирония и др.) основаны именно на нарушении постулата истинности, о чём упоминает и сам Грайс [Grice 1975: 53, Грайс 1985: 230–231]. Поэтому неудивительно, что литуратив часто служит для маркирования тропов, ср. метафору в примере (7):

(7) *Между тем, уже конец января, а значит, на нас надвигается ~~цунами, ураган, потоп, извержение~~ очередной детский день рождения* (из блога, 2010).

При использовании тропа может быть зачёркнуто как «исходное» языковое выражение, так и результат его преобразования, полученный с применением тропа. Так, в примере (7) «исходное» словосочетание *(очередной детский день рождения)* не зачёркнуто, а литуративом становятся его замены, полученные после применения метафоры. Обратная ситуация тоже была бы возможна, ср. (7’):

(7’) *Между тем, уже конец января, а значит, на нас надвигается ~~очередной детский день рождения~~ цунами, ураган, потоп, извержение.*

Не будет преувеличением сказать, что нарушением постулатов Грайса (чаще всего — постулата истинности) являются практически все случаи языковой игры, которые маркируются литуративами. В целом, литуратив как средство маркирования языковой игры, тропов и юмора выполняет ту функцию, которая в устной речи выпадает на долю интонации, мимики, жестикуляции и т. п.

Ещё одним классическим случаем нарушения постулата истинности является употребление устойчивых выражений[3]. Так как устойчивое выражение по своей природе не может соответствовать истине (истине соответствует не совокупность смыслов всех членов выражения, а общий переносный смысл), автор заменяет некоторую его часть своим словом, добиваясь соблюдения истинности — ср. (8), (9). В некоторых же случаях выражение зачёркивается целиком, при этом его значение должно обязательно соответствовать или быть хотя бы как-то связанным со значением незачёркнутой части (10):

---

[3] Под устойчивым выражением мы понимаем фразеологизмы, поговорки, крылатые фразы, в том числе из рекламы, мемы и просто коллокации.

(8) *То есть улицы и тротуары там настолько ~~суровы~~*[4] *узкие, что иной раз надо пешеходу надо прижаться к стене дома, чтобы машина могла проехать* (из блога, 2010)

(9) *Вот сломанная детская игрушка, <...>, окаменевшие остатки пищи на столе, покрытым толстым — толстым слоем ~~шоколада~~*[5] *пыли...* (из блога, 2005)

(10) *~~Хозяин подарил Добби носок!~~*[6] *Забрала из службы свой загранпаспорт* (из блога, 2012).

Для «ухода» от нарушения остальных постулатов Грайса литуративы используются значительно реже. При этом заметно, что употребление литуративов для избегания нарушения некоторого конкретного постулата является характерной чертой авторского стиля, в отличие от «истинностных» литуративов, частотных у всех авторов. Приведём несколько примеров.

Одним из не очень частых семантических типов литуративов является нарушение второго постулата количества («высказывание должно содержать не больше информации, чем требуется»). Литуратив является очень удобным способом выразить ту информацию, которую автор хочет упомянуть в своём тексте, однако которая по той или иной причине может оцениваться как лишняя (постулат количества) или как отклонение от темы (постулат релевантности):

(11) *Сегодня ко-пайлоту было скучно, и он развлекался тем, что показывал мне своё рабочее место. ~~Нет, порулить не дали :)~~* (из блога, 2013)

(12) *Завтра будет отличный рабочий день ~~и не менее прекрасные пирожки от Светланы~~, так что набираемся сил!)* (из блога, 2010)

Ещё один интересный тип нарушения — несоответствие одному из постулатов ясности («избегай непонятных выражений»). В корпусе встречаются примеры литуративов, где пишущий употребляет профессионализм, зачёркивает его и заменяет на общепонятный термин:

(13) *А вот есть буквы И и Ы, которые различаются на письме, но в ~~устном дискурсе~~ речи никогда не создадут пары ~~минимальной~~ типа полка-палка, когда два слова различаются только этой буквой* (из блога, 2010).

Любопытно, однако, что встречается и ровно противоположная стратегия намеренного нарушения этого постулата, что отсылает нас к следующему разделу.

---

[4] Распространённый в Интернете мем, получивший свою популярность благодаря программе «Наша Russia».

[5] Фраза из рекламы шоколадных батончиков Mars.

[6] Фраза отсылает к циклу романов Дж. К. Роулинг о Гарри Поттере; её смысл сводится к обретению свободы персонажем.

3.2. Литуративы и формирование впечатления о говорящем

В норме участники коммуникации стремятся создать о себе положительное впечатление. Для создания такого впечатления, безусловно, необходимо соблюдать постулаты эффективной коммуникации Грайса, однако не только их. Сам Грайс в своей статье отвлекается от рассмотрения дополнительных факторов, хотя и упоминает их существование: «Конечно, существуют постулаты и иной природы (эстетические, социальные или моральные) — такие, как, например, “Будь вежлив”» [Grice 1975: 47, Грайс 1985: 223]. Несмотря на это, при описании литуративов такие параметры не учитывать нельзя, хотя их очень сложно, а то и вовсе невозможно формализовать (это может стать темой для отдельного исследования).

На впечатление от автора текста влияет очень многое: например, он должен вести себя в соответствии с системой ценностей адресатов (ср. постулат такта в классической работе [Leech 1983: 108]: «не делай того (в том числе не говори того), чего не хочет твой собеседник»). К примеру, в обществе, в котором принято уважительно относиться к творчеству известных режиссёров, но не к порнографическим фильмам, человек, демонстрирующий интерес к аниме Хаяо Миядзаки, вызывает положительное отношение, а человек, проявляющий интерес к хентаю[7] — отрицательное, ср.:

(14) *Раньше я думала, что всё хорошее аниме ограничивается ~~хентаем~~ Хаяо Миядзаки и его студией Ghibli* (из блога, 2009).

Автор в норме старается показать себя с хорошей стороны: в глазах читателей он должен быть умным, добрым (или, по крайней мере, не злым), воспитанным, вежливым, скромным и т. п.

Приведём несколько примеров:

(15) *Теперь кошка наконец-то лежит где положено ~~еще бы гвоздиками ее прибить для верности~~* (из блога, 2012; нельзя показаться злым).

(16) *<…> на 2-й* (синей) *линии метрополитена г. Санкт-Петербурга <…> сложные слова произносятся диктором с побочным ударением на первом слоге: Пѐтрогра́дская, Э́лектроси́ла. ~~Жалко, что сейчас придёт [никнейм] и скажет, что всё это чушь~~* (из блог, 2010; нельзя показаться не умным).

(17) *Последовал смс диалог, потом звонки с надеждой ~~трахнуть~~ подружиться с ней* (блог, 2008; нельзя показаться коварным).

(18) *Приятель мне много рассказывал об этой стране победившего ~~пофигизма~~ социализма* (блог, 2010) (нельзя показаться невежливым).

---

[7] Аниме эротического или порнографического характера.

Сюда же относятся упомянутые выше случаи избыточного использования профессионализмов (ср. (13)), употребления «внутреннего» названия вместо «общепринятого» и т. п. — автор полагает, что всё это может показаться читателю неприятным:

(19) *9 августа в 12 часов к/з «Минск» <…> осчастливят своим присутствием [имя, фамилия] с музом и соавтором ~~Лёвой~~ Львом [фамилия] :)* (из блога, 2009).

Наконец, устойчиво выделяющимся случаем можно считать зачёркивание бранных и нецензурных слов или их аналогов:

(20) (20) *А тем временем из тьмы и тумана начали появляться скалистые выходы Басегов. ~~Бля!~~ Мама дорогая, как же туда добраться?* (из блога, 2007).

(21) (21) *А всё же есть во мне что-то мазохистское! ~~*очень нехорошо выругалась*~~* (из блога, 2010).

4. Синтаксис литуративов

Литуративы делятся на две группы в зависимости от того, есть ли в тексте замена для зачёркнутого фрагмента или нет [Занегина 2009]. Интересно количественное распределение: в пилотном корпусе из 150 случайно выбранных литуративов двадцати пяти авторов количество литуративов с заменой оказалось практически равным количеству литуративов без замен, а именно 74 и 71 случай соответственно (5 случаев оказалось невозможно отнести с уверенностью в одну из групп). Литуративы без замены — это нередко фрагменты, нарушающие постулат количества (ср. (11), (12)), однако в абсолютном большинстве случаев они не нарушают постулаты Грайса, но так или иначе формируют отрицательное впечатление об авторе (ср. (15), (16), (20)). Подробная классификация зачёркнутых фрагментов на основании объёма зачёркнутого текста представлена в работе [Занегина 2009].

Интересно отметить, что в ситуации замены литуратив всегда ставится перед своим незачёркнутым коррелятом: ср. выше примеры (7) и (7’), где при перестановке зачёркивания пришлось переставить и члены конструкции.

5. Классификация литуративов

Обобщая всё вышесказанное, можно составить семантико-синтаксическую классификацию литуративов, которая будет включать в себя три бинарных признака:

[± замена]: есть ли в тексте незачёркнутая замена для литуратива или нет?
[± нарушение постулатов]: нарушает ли зачёркнутый текст постулаты коммуникации или нет?
[± отрицательное впечатление]: формирует ли зачёркнутый текст отрицательное впечатление о своём авторе или нет?

Таким образом, существует $2 \times 2 \times 2 = 8$ теоретически возможных типов литуративов. Но на самом деле их число меньше, а именно $2 \times (2 \times 2 - 1) = 6$. Дело в том, что не встречаются литуративы с комбинацией признаков [– нарушение постулатов][– отрицательное впечатление]: это объясняется тем, что для появления литуратива необходимо хотя бы одно прагматическое основание, то есть хотя бы один из этих признаков должен иметь положительное значение.

1. [+ замена][+ нарушение постулатов][+ отрицательное впечатление]: ср. (5) (*под свои ~~сексуальные~~ твор<ч>еские нужды*);
2. [+ замена][+ нарушение постулатов][– отрицательное впечатление]: ср. (11) (*~~Нет, порулить не дали :)~~*);
3. [+ замена][– нарушение постулатов][+ отрицательное впечатление]: ср. (19) (*с музом и соавтором ~~Лёвой~~ Львом*);
4. [– замена][+ нарушение постулатов][+ отрицательное впечатление]:

(22) *<…> я проснулась с чувством глубокого разочарования в себе, в который раз выслушала от бабушки, что я ~~говно, потому что~~ нихрена не учусь, ничего не делаю, <…>* (из блога, 2009; зачёркнутые слова нарушают постулат истинности и содержат отрицательную оценку автора; кроме того, употребление бранного слова говно также характеризовало бы автора отрицательно).

5. [– замена][+ нарушение постулатов][– отрицательное впечатление]: ср. (12) (*отличный рабочий день ~~и не менее прекрасные пирожки от Светланы~~*)
6. [– замена][– нарушение постулатов][+ отрицательное впечатление]: ср. (20) (*~~Бля!~~*)

Впрочем, далеко не каждому примеру можно однозначно приписать ту или иную комбинацию признаков. На первый взгляд может показаться, что это недостаток классификации, но на самом деле это ещё одно интересное свойство зачёркивания: литуратив амбивалентен по своей природе, и множественность возможных пониманий и создаёт особый стилистический и игровой эффект, которым пользуются авторы текстов. Так, при рассмотрении примера (5) уже отмечалось, что непонятно, что следует считать «абсолютной истиной», а что — «авторской». Однако более распространены случаи, когда неочевидно именно отнесение к одному из 6 пунктов нашей классификации, ср.:

(23) *Что-то на почве ~~зверского скандала~~ милой супружеской ссоры пробило меня на печальную мыслю: вот интересно, кто-нить из френдов согласился бы на мне жениться?!* (из блога, 2006)

Наиболее естественно трактовать литуратив в этом фрагменте как [+ замена][+ нарушение постулатов][− отрицательное впечатление]: читатель предполагает, что зачёркнутый текст нарушает постулат истинности («зверского скандала не было, а была просто небольшая супружеская ссора»), но подспудно рассматривает и возможность трактовки [+ замена] [− нарушение постулатов][+ отрицательное впечатление] («зверский скандал действительно был, но чтобы у вас не создалось отрицательное впечатление обо мне и моей семье, я зачёркиваю это и заменяю на более слабую характеристику»).

Таким образом, все литуративы можно распределить на 6 основных классов по трём бинарным признакам. Бывают случаи, когда принадлежность классу неясна, но такие неясности — неотъемлемое свойство литуративов: именно амбивалентность и делает их хорошим средством языковой игры.

6. Литуративы на фоне стандартных приёмов языковой игры

Литуратив является одним из приёмов языковой игры, к которому ближе всего примыкают ирония и сарказм. При этом важным отличием литуратива от остальных форм языковой игры является способ бытования: литуратив существует только в письменном языке, тогда как остальные приёмы языковой игры в первую очередь являются устными. Более того, иногда перевод в письменную форму делает такие виды языковой игры, как ирония и сарказм, непонятными, поскольку теряются важные дополнительные маркеры: интонация, мимика и т. п. В интернет-речи функцию этих маркеров часто берут на себя литуративы без замены. Фактически, они являются не особым новым приёмом, а графическим отображением явлений устной коммуникации.

Иначе обстоит дело с литуративами с заменой. В принципе, они близки к иронии и сарказму, однако в отличие от них имеют два плана выражения: «несказанный» и «сказанный». Так, согласно «Литературному энциклопедическому словарю», «ирония — иносказание, когда слово или высказывание обретают в контексте речи значение, противоположное буквальному смыслу или отрицающее его, ставящее под сомнение» [Кожевников, Николаев (ред.) 1987]. Можно попытаться формализовать это определение: «Я подразумеваю А, но по некоторой причине Х говорю противоположное ему В». Тогда определение литуратива с заменой в аналогичной терминологии может звучать так: «Я подразумеваю А, но по некоторой причине Х говорю не только А, но и В и при этом делаю вид, что не говорю А». Для иронии и сарказма свойственно, что говорится только В, а А так и остаётся подразумеваемым. Таким образом, в случае литуратива с заменой мы имеем два конкурирующих означающих с противопоставленными друг другу означаемыми, тогда как в случае иронии или сарказма представлено только одно означающее, которое имеет одно непосредственно выводимое означаемое и одно подразумеваемое. Причина, по которой подразумеваемое при иронии умалчивается, а при использовании литуратива — зачёркивается, может быть одна и та же, как, например, [+ отрицательное впечатление].

Экспликация обоих планов при иронии и сарказме в традиционных способах коммуникации встречается редко, ср.:

(24а) *Ты уснешь, окружен попечением / Дорогой и любимой семьи /* (Ждущей смерти твоей с нетерпением) (Н. А. Некрасов, «Размышления у парадного подъезда»).

В современной интернет-коммуникации данный пример легко мог бы содержать литуратив, который как раз и позволяет одновременно охарактеризовать семью персонажа двумя способами и сделать как бы невысказанным тот из них, который рискует создать отрицательное впечатление о говорящем ([+ отрицательное впечатление]), поскольку в обществе не принято прямым текстом говорить такие вещи:

(24б) *Дорогой и любимой семьи, / ~~Ждущей смерти твоей с нетерпением~~*

Практически любой литуратив с заменой, положительно охарактеризованный по параметру [± отрицательное впечатление], может быть заменён ироничным высказыванием, ср. (18) и сконструированный пример (25), областью бытования которого могла бы быть устная речь (для того, чтобы ирония надёжно распознавалась, должны быть применены специальные звуковые маркеры типа хмыканья или покашливания, а также мимика и интонация):

(25) *Последовал смс-диалог, потом звонки с надеждой, кхм, подружиться с ней.*

Причина иронической замены в данном случае, как и в случае с литуративом, заключается в невозможности сообщения прямым текстом о своих намерениях.

Отметим, что даже и в устной речи редко, но встречаются аналоги литуратива с заменой — намеренные оговорки с немедленным исправлением. При этом говорящий может либо просто исправиться без специальных маркеров, либо отметить намеренность оговорки звуковым маркером типа покашливания или выразить это вербально:

(26) *Вот способ мышления, который продемонстрировала Государственная дура — ой, простите, оговорился — Государственная Дума, ну, поражает…* (В. Познер, программа «Познер», 23.12.2012)

Таким образом, литуративы с признаком [– замена] являются способом письменного выражения традиционных приёмов языковой игры, в первую очередь иронии и сарказма, а литуративы с признаком [+ замена] — это новый приём, характерный для интернет-коммуникации и в устной речи встречающийся редко.

7. Литуративы и технические средства коммуникации

Основной сферой распространения литуративов в русскоязычном Интернете стали блоги, в первую очередь — платформа LiveJournal. Пользователям

блогов приходится владеть основами HTML-разметки для создания гиперссылок и шрифтовых выделений, в том числе и зачёркивания (тэги <s>…</s> или <strike>…</strike>), или же использовать соответствующие кнопки (в визуальном редакторе LiveJournal кнопка для создания зачёркивания находится рядом с кнопками для выделения полужирным шрифтом, курсивом или подчёркиванием). То, что для зачёркивания текста в блоге надо приложить не больше усилий, чем для использования полужирного или курсивного начертания, и привело к широкому распространению этого приёма.

В эпоху популярности форумов литуративы часто использовались и там: обычно для их создания применялся bbCode, создававшийся как упрощённый вариант HTML. Зачёркнутый текст в этом языке разметки оформляется тэгами [s]…[/s], которые также достаточно просты для запоминания и употребления; кроме того, на многих форумах есть и панель визуального редактора с кнопкой для создания зачёркивания.

Однако в последнее время литуративы становятся менее употребительными из-за того, что основные средства сегодняшней Интернет-коммуникации — социальные сети — не поддерживают даже простейшего форматирования текста. Ни ВКонтакте, ни Facebook, ни Twitter не позволяют пользоваться шрифтовыми выделениями, в том числе и зачёркиванием. Зачёркивание можно сымитировать, ставя с обеих сторон от каждой буквы объединённый диакритический знак ‘COMBINING LONG STROKE OVERLAY’ (U+0336). Для того, чтобы это делать автоматически, существуют даже специальные онлайн-сервисы (напр., http://vkontakte.doguran.ru/perecherknutyj-tekst-vkontakte.php), но очевидно, что это слишком трудоёмкий способ, чтобы он мог использоваться часто. Когда один из авторов этой статьи написал об этой технологии зачёркивания в записи ВКонтакте, он вскоре получил от одного из своих друзей комментарий: «Anyway, способ видится мне настолько костыльным, что мне, пожалуй, не очень нужны зачеркивания:) Вот что им стоило ввести поддержку HTML- или BB-кодов?»

Именно поэтому в социальных сетях, на которые приходится основная доля сегодняшней Интернет-коммуникации, зачёркивание практически не встречается. Можно наблюдать утрату зачёркивания даже в языке одного человека: пользователь, перешедший из ЖЖ на ВКонтакте, в первой своей записи еще может пытаться искать альтернативы зачеркиваню (27), но в дальнейших записях уже не пользуется этим средством:

(27) *В связи с пересмотром личностных приоритетов…* (зачёркнуто)
В связи с положением Сатурна относительно Венеры в плоскости эклиптики… (зачёркнуто)
Я стукнулась головой, поэтому (зачёркнуто)
Переношу часть он-лайн присутствия из жежешечки во вконтакт. Привет! :)

Технические средства для создания зачёркивания сохраняются в электронной почте, а также в некоторых службах мгновенных сообщений. Так, в визуальном редакторе почты на Mail.ru есть кнопка для создания зачёркивания. Такая же кнопка есть и в почте «Яндекса», хотя там по умолчанию письма

создаются в режиме plain text, и можно предполагать, что большинство пользователей не обращается к расширенному форматированию. Однако в web-интерфейсе Gmail средства для создания зачёркиваний отсутствуют даже несмотря на то, что чат Google Talk поддерживает зачёркивание, оформленное при помощи языка разметки Markdown: зачёркнутый текст должен быть с двух сторон окружён дефисами (т. е. «-зачёркнутый текст-» автоматически преобразуется в «~~зачёркнутый текст~~»).

Как бы то ни было, электронная почта и чаты находятся вне сферы публичной коммуникации, а в публичном Интернет-пространстве доминирующую роль в настоящий момент играют сервисы, не поддерживающие зачёркивание. Это позволяет говорить о том, что роль литуративов в современном языке Интернета снижается. Впрочем, нельзя исключать появления новых технических возможностей, которые снова вернут зачёркивание в число популярных приёмов.

Литература

1. *Грайс Г. П.* Логика и речевое общение // Новое в зарубежной лингвистике. Вып. 16. Лингвистическая прагматика. — М., 1985. С. 217–237.
2. *Гусейнов Г. Ч.* Неполная коммуникация в блогосфере: эрративы и литуративы, http://www.speakrus.ru/gg/liturative.htm
3. *Занегина Н. Н.* ~~Я этого не говорил~~: о литуративах, зачеркиваниях или мнимых текстах // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). — М., 2009, с. 112–115.
4. *Кожевников В. М., Николаев П. А.* (ред.). Литературный энциклопедический словарь. М., 1987.
5. *Attardo, Salvatore.* Linguistic Theories of Humor. Berlin / New York, 1994.
6. *Grice, H. P.* Logic and Conversation. In: Syntax and semantics, vol. 3. Ed. by P. Cole and J. L. Morgan. New York, 1975, pp. 41–58.
7. *Leech, Geoffrey N.* Principles of Pragmatics. London / New York, 1983.
8. *Mey, Jacob L.* Pragmatics: An Introduction. Malden, 2001.

References

1. *Attardo, Salvatore.* Linguistic Theories of Humor. Berlin / New York, 1994.
2. *Grice, H. P.* Logic and Conversation. In: Syntax and semantics, vol. 3. Ed. by P. Cole and J. L. Morgan. New York, 1975, pp. 41–58.
3. *Grice H. P.* Logic and Conversation [Logika i rechevoe obschenie], Novoe v zarubezhnoj lingvistike. Vyp. 16. Lingvisticheskaja pragmatika [New Trends in Foreign Linguistics. Vol. 16. Pragmatics in Linguistics]. Moscow, 1985, pp. 217–237.
4. *Guseinov, Gasan.* Incomplete communication in the blogosphere: erratives and lituratives [Nepolnaja kommunikatsija v blogosfere: èrrativy i lituraty], http://www.speakrus.ru/gg/liturative.htm
5. *Kozhevnikov, Vadim, and Nikolaev, Peter.* Encyclopaedic Dictionry of Literature [Literatyrnyj Ènciklopedicheskij Slovar']. Moscow, 1987.
6. *Leech, Geoffrey N.* Principles of Pragmatics. London / New York, 1983.
7. *Mey, Jacob L.* Pragmatics: An Introduction. Malden, 2001.
8. *Zanegina, Natalija.* ~~I didn't say that~~: on lituratives, strikethrough, or spurious texts [~~Ja ètogo ne govoril~~: o lituraty, zacherkivanijah ili mnimyh tekstah], Komp'juternaia lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii "Dialog 2009" [Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog 2009"]. Moscow, 2009, pp. 112–115.

НЕЧЕТКАЯ НОМИНАЦИЯ В РУССКОЙ РАЗГОВОРНОЙ РЕЧИ: ОПЫТ КОРПУСНОГО ИССЛЕДОВАНИЯ

Подлесская В. И. (podlesskaya@ocrus.ru)

Российский государственный гуманитарный университет, Москва, Россия

Ключевые слова: нечеткая референция, русский язык, корпус, устная речь

VAGUE REFERENCE IN RUSSIAN: EVIDENCE FROM SPOKEN CORPORA

Podlesskaya V. I. (podlesskaya@ocrus.ru)

Russian State University for the Humanities, Moscow, Russia

The paper focuses on phenomena that fall under a broad category of what is called “loose uses” of language or “vague reference”. These are lexical, grammatical and prosodic resources that allow the speaker to refer to objects and events for which the speaker fails to retrieve the exact name, or simply finds the exact name to be unnecessary or inappropriate. Based on first-hand corpus data of spoken Russian, the paper investigates expressions that are used in a language to temporarily substitute a delayed constituent, as well as those that do not imply any later substitution, but rather suggest an approximate nomination sufficient at the current moment of communication. These expressions can be used instead of their supposed exact correlate or together with it. The first option implies that an expression is used as a generic, or as a cover bleached nomination. The second option implies that the speaker doesn’t take the full responsibility for the given actual nomination the expression is added to, since it is in some sense incomplete or not fully appropriate. The study of lexical resources of vague reference in spoken Russian is complemented by investigating also the associated syntactic and prosodic patterns.

Key words: vague reference, Russian, spoken corpus, natural discourse

Памяти Александра Евгеньевича Кибрика,
любившего и научившего любить
живое в языке

1. Постановка задачи

Среди средств, которые задействует говорящий в проблемных точках речепорождения, имеется особая группа лексем, грамматических конструкций и просодических конфигураций, которые позволяют осуществить нечеткую, или приблизительную номинацию[1]. Эта группа средств реализуется в контекстах, где прямое точное называние объекта или положения дел оказывается невозможным или нежелательным. В частности, (а) в спонтанной речи говорящий может испытывать временные трудности при поиске нужного выражения; (б) подходящее выражение вообще может отсутствовать в арсенале говорящего; (в) нужное выражение в арсенале имеется, но говорящий считает его употребление по каким-либо прагматическим причинам неуместным — например, оно стилистически не вписывается в текущий регистр дискурса, табуировано и проч. В такого рода контекстах средства нечеткой номинации позволяют информировать слушающего о том, что говорящий снимает с себя ответственность за точность вербализации и предоставляет слушающему возможность посильного сотрудничества в реконструкции исходного смысла. Так, в следующих примерах этой задаче служат слова *что-то типа, что-то вроде…но…, или как там она называется*:

(1) ВИЖ[2]
там /\ну у нас были-и ээ ·· что-то типа вводных /\тренингов по поводу работы,

(2) НКРЯ
А что такое жюльен? / Это такая горячая закуска / что-то вроде салата / но горячее / из птицы или из грибов. [Микродиалоги // Из материалов

---

[1] Работа поддержана РФФИ (грант № 13-06-00179) и «Программой стратегического развития РГГУ».

[2] Используется материал трех экспериментальных корпусов устной монологической речи: «Рассказы о сновидениях», «Рассказы сибиряков о жизни» и «Весёлые истории из жизни», доступных в пилотном режиме на сайте spokencorpora.ru (разрабатывается коллективом исследователей при участии автора статьи). Использовались также данные Национального корпуса русского языка (ruscorpora.ru) Примеры даются в том формате, в котором они задокументированы в соответствующем корпусе. О деталях транскрипции, использующейся в экспериментальных корпусах см. подробнее [Кибрик, Подлесская 2009]. Для правильной интерпретации приводимых примеров достаточно знать, что: тональный тип акцента указывается перед словом иконически с помощью косых черт; ударный слог в слове — носителе фразового акцента подчеркивается; незавершенность открытого списка нотируется в транскрипте многоточием на границе иллокуции, и знаком «,,,» (три запятых) внутри иллокуции; речевой сбой маркируется знаком «==» на границе иллокуции, и знаком «||» внутри иллокуции. Ссылки на корпуса даются после номера примера в виде сокращений НКРЯ, РОС, РСЖ, ВИЖ, соответственно.

Саратовского университета, 1984–1985]

я все-таки не могу всерьез относиться к этой новой хронографии / хронологии… или как там она называется. [Интервью с Сергеем Лукьяненко на радиостанции «Маяк» (2006)]

Семантика и прагматика средств нечеткой номинации исследовалась достаточно подробно, правда, преимущественно на английском материале, ср. [Lakoff 1972], [Markkanen, Schröder 1997]; [Kaltenböck, Mihatsch, Schneider 2010], [Jucker et al. 2003]; [Sperber & Wilson 1991: 546], [Channell 1994], [Enfield 2003] и др. Что же касается их грамматики, то она изучена гораздо слабее. В данной работе я попытаюсь частично восполнить этот пробел, рассмотрев на русском корпусном материале основные способы интегрирования средств нечеткой номинации в структуру предложения. Кроме того, я покажу, что наряду с более распространенными — и, как следствие, более изученными — лексическими средствами нечеткой номинации в этой зоне активно эксплуатируются такие синтаксические стратегии, как открытые ряды сочиненных групп и аппозитивные конструкции; кроме того используется такая стилистическая фигура, как семантическая и фонетическая рифма, ср.:

(3) НКРЯ
Приданое еще — плошки, ложки, серебришко, золотишко, в двадцать пять тысяч не уложишь… [В. Я. Шишков. Угрюм-река. Ч. 1–4 (1913–1932)]

Я рассмотрю также особые просодические паттерны, которые задействуются для выражения нечеткой номинации, и покажу, что разноуровневые средства используются в зоне нечеткой номинации не изолированно, они имеют тенденцию объединяться в кластеры.

2. Интеграция лексических маркеров нечеткой номинации в структуру предложения: стратегия замещения и стратегия совмещения

При использовании маркера нечеткой номинации говорящий может следовать одной из двух стратегий: стратегии замещения и стратегии совмещения, см. подробнее [Podlesskaya 2010]. В первом случае маркер используется ВМЕСТО предполагавшегося или возможного точного наименования, отсылая к более размытой или более широкой категории. Во втором случае маркер используется СОВМЕСТНО с неким контекстно приемлемым способом именования — в качестве сигнала о неполном референциальном соответствии избранного способа.

Маркеры первого типа, заместители (английский термин placeholder, не имеющий устоявшегося русского аналога), обычно имеют местоименную природу. Чаще всего, это слова, относящиеся (или восходящие) к следующим

классам: указательные, вопросительные, универсальные или неопределенные местоимения; универсальные квантификаторы, существительные с максимально обобщенным значением *(вещь, дело, штука),* лексикализованные конструкции типа англ. *whatchamacallit* с местоименным компонентом, см. подробнее [Hayashi, Yoon 2006; Podlesskaya 2010].

(4) НКРЯ
я на рынок. Зелени надо купить / еще там / по мелочи… того-сего… [Разговор в автобусе (2006)]

(5) НКРЯ
над такого рода кубом появляется вот эта грандиозная штуковина / я так и не понял / это стела / или это здание / или там будет решетка какая-то смотровая [Беседа А. Гордона с Л. Кацисом об Апокалипсисе, НТВ, «Гордон» (2003–2004)]

Маркеры-заместители полноценно встраиваются в структуру предложения, ср. *того-сего*, как дополнение, оформленное родительным партитивным в (4), или *штуковина* как подлежащее в именительном падеже в (5).

Маркеры-заместители могут использоваться говорящим в двух режимах — долгосрочном и краткосрочном. При долгосрочном режиме говорящий в принципе отказывается от точной вербализации некоторого смысла, но вынужден по условиям локальной структуры текста употребить языковое выражение с заданными формальными свойствами, например, именную или глагольную группу. Таковы употребления. продемонстрированные выше в (4) и (5). Для долгосрочного режима характерно, в частности, употребление неопределенных местоимений, отсылающих к референту, известному говорящему, но необязательно известному слушающему (серия *кое-* местоимений в русском языке). Так, эти местоимения используются для намеренно нечеткой референции, когда отсылают к сущности, известной обоим локуторам, но намеренно не называемой:

(6) НКРЯ
Виктора Петровича кое-кто кое-куда пригласил и кое-что предложил. — Гость весело блеснул дымчатыми стеклами очков на Евгению: поняла ли она, о чем идет речь? Евгения глаза прикрыла и чуть заметно кивнула. Поняла: кое-кто кое-куда всуе не упоминается.. [Елена и Валерий Гордеевы. Не все мы умрем (2002)]

При краткосрочном режиме маркер-заместитель временно подставляется в структурную позицию составляющей, для которой — из-за трудностей речепорождения — говорящему сразу не удается подыскать адекватной реализации. После того, как говорящий справляется с проблемой, он восстанавливает отложенную составляющую. Краткосрочный режим обычно используется говорящим в тех случаях, когда проблема связана с «близким» поиском, т. е. когда предстоящая порция дискурса уже достаточно хорошо спланирована и затруднения касаются

выбора конкретного выражения из ограниченной зоны возможностей. Во многих языках, в том числе и в русском, в этой ситуации используются согласуемые маркеры-заместители, типа *этот (самый) / эта (самая), такой / такая*, как *его/её*, которые демонстрируют, что говорящий уже выбрал грамматическую форму планируемой группы и колеблется лишь в выборе конкретной номинации. Ср. следующий пример, где заместители *(эту, как её, с этим, как его)* полностью дублируют предложно-падежную форму отложенных составляющих:

(7) НКРЯ
Эту / как её / переписку Энгельса с этим… как его / дьявола / с Кауцким.
[Владимир Бортко и др. Собачье сердце, к/ф (1988)]

Нередко один и тот же маркер-заместитель используется в языке и в краткосрочном, и в долгосрочном режиме. Так, в следующем примере согласуемый местоименный маркер *этот* используется не в краткосрочном режиме, как в примере (7) выше, а в долгосрочном — он замещает словоформу *фантиками*, которая легко воссоздается слушающим по контексту. При этом в тексте нет явных симптомов речевого сбоя, так что говорящая не испытывала трудностей с поиском этой словоформы, не откладывала ее произнесения, а просто довольствовалась приблизительной номинацией, полагаясь на сотрудничество со слушающим:

(8) РСЖ
когда мы э= ··(0.4) =ти фантики /разгл<u>а</u>живали,
··(0.1) и ··(0.3) \мечт<u>а</u>ли,
как мы /\поменя<u>е</u>мся,,,
какими /э<u>т</u>и<u>м</u>и,
··(0.1) и мы в свою деревню ··(0.3) привезли очень много новых /\ф<u>а</u>нтиков.

Стратегия совмещения — в отличие от стратегии замещения — требует маркеров другого типа: они синтаксически обычно несамостоятельны, рекрутируются из неизменяемых слов или эволюционируют в неизменяемое слово. Эти маркеры — их чаще всего именуют аппроксиматорами — обычно формируют единую составляющую вместе с выражением, которое ими «семантически обслуживается». Типичными аппроксиматорами являются, например, маркеры, восходящие к словам и конструкциям со значением подобия. Так, классическим аппроксиматором в английском языке является *like*; в известной работе Andersen 1998, было показано, что *like* может входить в любые типы групп — именные, глагольные, квантитативные и др. В русском языке так же ведут себя такие аппроксиматоры, как *своего рода*, *типа*, *как бы* и др. Ср.

(9) НКРЯ
А ну там / как бы / уже вышел один мужик / он начал рассказать нам…
ну / как бы свой доклад читать / да… Вот / а нам / по идее / надо слушать / а потом… ну / такие / как бы преподавателям отчёт сдаём. [нрзб] .
[Рассказ о конференции (2006)]

(10) РСЖ
····(1.2) а ещё ···(0.8) там был ··(0.3) эээээ(1.1) ··(0.3) как бы ···(0.8) /отдел,
··(0.2) так ска= мм(0.2) || так \сказать,
···(0.6) где-е ··(0.4) ммм(0.5) были-и ···(0.5) ”(0.1) /–крокоди-илы-ы,,,
···(0.7) то есть эээ(0.6) ””(1.0) всякь= || ··(0.2) различные /–зме-еи,,,

Маркеры-аппроксиматоры, в отличие от маркеров-заместителей, выступают не «вместо» ненайденного языкового выражения, а «вместе» с одной или несколькими «пробными» попытками вербализовать некоторый смысл, т. е. как сигнал о том, что предпринятая попытка нуждается в обобщении или уточнении, но адекватного языкового выражения в арсенале говорящего в данный момент не нашлось. Ядром аппроксимативной зоны является количественная аппроксимация, т. е. обозначение приблизительного количества. Этот тип аппроксимации представлен в языках мира наиболее многообразно, однако вполне употребительны аппроксиматоры и при нечеткой номинации объектов, признаков и ситуаций (см. подробнее скрупулезное монографическое исследование [Адамович 2011], выполненное на материале русского, немецкого и белорусского языков). При выражении приблизительного количества сфера действия «оператора приблизительности» обычно не выходит за пределы количественной группы. Формально эта группа может иметь вершиной предлог *(около восьми метров)*, числительное *(приблизительно восемь метров, сорок с лишним метров)* или аппозитивное сочетание *(семь-восемь метров)*; может быть использована и чисто синтаксическая операция инверсии числительного и исчисляемого объекта *(метров восемь)*. Ср. употребление аппроксиматора с предложным статусом *порядка*:

(11) ВИЖ
самое /интересное оказалось –то,
что \действительно в-в /Англии порядка /–семидесяти ·· этих сервис-центров,,,
··· ээ во Франции порядка шестидесяти /трёх,,,
и так \далее,

Операторы приблизительности нередко используются совместно, образуя иногда изысканные конфигурации. Показательны два следующих примера, где используется инверсия плюс комбинация двух лексических аппроксиматоров — *наверно* и *точно*; последний, вопреки внутренней форме, обозначает не точное количество, а ‘не меньше, чем…’:

(12) РСЖ
мы очень долго /шли,
(Часов наверно ··(0.4) \пять.
···(0.5) \Точно шли.)

[Н]амотали ··(0.1) лишних наверно кило́метров \семьдесят,
т\очно,

нь= ==
ну не /семьдесят,
\пятьдесят.

Заметим, что в значении ‘не меньше чем…’ *точно* модифицирует глагольную группу и обычно является носителем коммуникативно значимого акцента, тогда как при обозначении точного численного значения *точно* (как и *ровно*, например) входит в количественную группу и, в общем случае, не акцентируется[3]:

(13) НКРЯ
Полученная величина и есть морская миля, которая для простоты принимается равной точно 1852 метрам. [А. Цыбин. Футы, метры и постоянная Планка // «Наука и жизнь», 2006]

Еще один парадоксальный семантический перенос наблюдается при употреблении неопределенного местоименного наречия *где-то* не в пространственном значении, а для обозначения приблизительного количества, а также — в качестве общеситуативного аппроксиматора:

(14) РСЖ
…(0.6) Что-о значит свет этот был в течение где-то двух /минут,

(15) РОС
>> [приснилось] что /я-а ==
….(1.5) (Как это лучше /сказать?)
…(0.7) ну’ ….(1.2) что я \поссорила маму с \папой.
”(0.3) Ну где-т= || ··(0.3) ну где-то \так.

Одной наиболее распространенных функций аппроксиматоров (и особенно дискурсивных слов, восходящих к конструкциям уподобления — *так, такой, типа*), является аппроксимативное введение чужой речи — в качестве сигнала о том, что говорящий не несет полной ответственности за точность и возможные интерпретации цитируемого или просто считает часть информации, содержавшейся в цитируемом фрагменте, контекстно избыточной. При этом цитирование возможно как в форме прямой, так и в форме косвенной речи (см. проницательный анализ употребления английского *like* в этой функции в [Fox Tree 2006; Fox Tree, Tomlinson 2008]).

---

[3] Анонимный рецензент справедливо указал на существование конструкций с постпозитивным акцентированным *ровно/точно* при выражении точного численного значения: *пришли в два часа \ровно*. Я думаю, что и постпозиция, и акцент в данном случае являются следствием эмфазы. Я глубоко признательна рецензенту и за проницательное замечание, касающееся возможного употребления *ровно* и *точно* в функции сравнительных союзов, ср. *говорит, точно (= как) пьяный*. Такое употребление лишний раз демонстрирует, что значения уподобления и аппроксимации неслучайным образом соседствуют на семантической карте.

В русском языке специализированным средством аппроксимативного введения чужой речи является, в частности, сочетание указательного местоимения или местоименного наречия с частицей *-то* в составе цитируемого фрагмента (часто с редупликацией), ср.

(16) НКРЯ
а другим пи… в письмах он пишет: «Я тогда-то буду там-то, а тогда-то собираюсь делать то-то», [Беседа А. Максимова с Т. Себенцовой в програме «Времечко», ТВЦ // Архив Хельсинкского университета, 2000–2005]

номер этого… участка записать / позвонить просто… друзьям / родителям рассказать / вот такой-то такой-то / там-то / там-то просит мои документы. [Разговоры на прогулке (2006)]

Между маркерами-заместителями и маркерами-аппроксиматорами нет непроходимой границы: один и тот же маркер может использоваться и в режиме замещения, и в режиме совмещения. Об этом красноречиво свидетельствует, например, многофункциональность неопределенного местоимения *что-то* (имеющего в разговорном языке редуцированный вариант *чего-то* > *чё-то*). В своем прототипическом употреблении местоимение полноправно встраивается в структуру предложения, в том числе, с модификаторами, ограничивающими сферу референции, но не снимающими неопределенности:

(17) РОС
…(0.5) и /нашли чего-то такое \желе̲зное.

Сохраняется оно в структуре предложения и тогда, когда используется в качестве маркера-заместителя для выражения нечеткой референции:

(18) НКРЯ
«Экологическая глобализация как аспект там чего-то чего-то… » / в общем / экономические аспекты глобализации / так скажем.. [нрзб] . [Рассказ о конференции (2006)]

а также в составе предложных количественных аппроксиматоров *(с чем-то, без чего-то)*:

(19) НКРЯ
Я купилась на это 36,2 евро / и когда мне посчитали в долларах / это получилось там 80 с чем-то / понимаете. [Беседа с социологом на общественно-политические темы (Москва) // Фонд «Общественное мнение», 2004]

В примерах (17)–(19) выше падеж местоимения лицензируется его синтаксическим хозяином, но этот статус местоимение утрачивает, когда употребляется в неизменяемой форме как собственно частица-аппроксиматор.

При этом может выражаться не только количественная аппроксимация, как в (20):

(20) НКРЯ
Там чего-то более полугодовой зарплаты в среднем / это очень бешенные какие-то деньги [Фонд «Общественное мнение», 2003]

но и значение нечеткого припоминания или неясной причины ситуации, неконтролируемой говорящим и, как правило, нежелательной для говорящего:

(21) НКРЯ
Мне тоже какие-то ужастики снились. Чё-то мы там на корабле плаваем / куда-то заплыли на остров незнакомый / и там еще какие-то бандиты ходят. [Телефонный разговор двух студентов (2005)]

пыталась с Интернета скачать / но у меня чё-то компьютер долго грузится… не идёт / в общем… [Разговоры на прогулке (2006)]
Да / вообще так / как… Чё-то тебя ваще плохо слышно. [Телефонные разговоры московских студентов (2008)]

Используется местоимение *что-то / чего-то* и для аппроксимативного введения чужой речи:

(22) РОС
…(0.6) ’А-а /дру̲г мне \говорит,
··(0.3) «/\Не̲т,
ты’= || ··(0.3) ты-ы ··(0.1) подойди чего-то **/сама̲-а** к нему¡…
*не /**о̲н** должен с-сюда перейти,*
а \ты̲ должна туда перейти¡»

Маркеры-аппроксиматоры и маркеры-заместители часто используются совместно. Показателен следующий пример, где говорящая, испытывая, очевидные трудности с вербализацией нужного понятия (‘служба помощи на дорогах’), сначала прибегает к помощи согласуемого заместителя *этого, как его*, из чего следует, что первоначально в этой позиции, скорее всего, предполагалось употребление одушевленного существительного мужского рода. Однако поиск не увенчался успехом, поэтому в соответствующую позицию помещается группа с аппроксиматором *типа*, а затем добавляется описание искомого референта через его функцию. До самого конца говорящая остается неудовлетворенной результатами поиска точной номинации, что подтверждается сигналами хезитации — длительной паузой (1.3 сек) и удлинением звуков в предпоследней строке примера:

(23) ВИЖ
Езжайте в автосервис с аварийкой быстре̲е@!
{СМЕХ}..(0.4) Или может быть вам вызвать /этого —

\как его,
— типа /“Ангела”?
Ну \вот,
то что-о ….(1.3) \служба-а,
\помощи.

Как будет показано в следующем разделе, кластеризация свойственна показателям нечеткой номинации, и лексические маркеры часто используются совместно с чисто синтаксическими, а также просодическими сигналами.

3. Синтаксические и просодические стратегии, опирающиеся на выражение незавершенности. Кластеры разноуровневых средств выражения нечеткой номинации

Выше уже приводился пример чисто синтаксического средства нечеткой номинации — инверсии числительного и единицы измерения для выражения приблизительного количества (*метров восемь*). Этот пример симптоматичный, но достаточно периферийный; гораздо более универсальным и распространенным синтаксическим приемом является использование сочинительной техники, а также функционально близких к сочинению аппозитивных конструкций. Структурный параллелизм сочиненных групп (как без союза, так и с повторяющимся союзом) используется как конструктивное средство предъявления открытого списка объектов или ситуаций, подразумевающего возможное, но не эксплицированное продолжение. В устной речи эта синтаксическая стратегия, как правило, поддерживается особыми просодическими средствами выражения незавершенности. В русском языке это, прежде всего, интонационная конструкция «открытого ряда», или «имитации ментальной деятельности (припоминания)» по Т. Е. Янко [Янко 2008: 109–117, 163–170] — с пологим подъемом тона и последующим ровным или слегка нисходящим тоном (часто — с растяжением ударного гласного)[4]:

(24) РСЖ
··(0.2) мы /пришли,
поставили /–пала-атки,,,
···(0.7) ээ(0.2) разожгли /–костё-ор…
····(1.2) /\Во-от,
··(0.4) потом ==
··(0.1) там рядом была /река,
··(0.3) и мы все пошли /–купа-аться…
···(0.8) {ЦОКАНЬЕ} /–Покупались,,,

[4] Напомним, что незавершенность этого типа нотируется в транскрипте многоточием на границе иллокуции, и знаком «,,,» (три запятых) внутри иллокуции (см. выше комментарий к примеру (1)).

····(1.1) ээээ(0.5) вечером мы жарили || ээ(0.1) жарили /–шашлыки-и,,,
····(1.7) ээ(0.2) пели /–песни-и…

Синтаксические и лексические средства чаще всего используются совместно. Прежде всего, присоединение маркеров-аппроксиматоров регулярно эксплуатирует сочинительную технику — маркер связывается сочинительным отношением с одним или несколькими членами сочинительного ряда по типу «*X, (Y, …) и другие*», «*X, (Y, …) и прочее*», «*X, (Y, …) и тому подобное*», «*X, (Y, …) и так далее*». В такого рода кластерах широко используются и разделительные союзы, образуя паттерны типа «*X или Y или что-то вроде того*»:

(25) НКРЯ
Смещение каких-то то ли дисков / то ли позвонков / то ли ещё чё-то /
я не знаю / э-э-э им тоже желательно спать на твёрдой поверхности.
[Разговор подруг // Из коллекции НКРЯ, 2007]

К сочинительному ряду могут подключаться и обобщающие слова — вместе с аппроксиматором или без такового. Наряду с собственно сочинением, используются и аппозитивные конструкции, в том числе, с семантически и фонетически рифмующимися элементами. Ср. следующий пример, где одновременно используется несколько способов приблизительной номинации мелких насекомых — рифмованная аппозитивная конструкция *жучки-паучки-червячки*, сочинительный ряд с аппроксиматором в составе обобщающего выражения *и прочие летающие насекомые* и сочинительный ряд с обощающим выражением, неверно употребленным без аппроксиматора *шмели ·· мухи ·· и насекомые*. Все это лексико-грамматическое разнообразие реализуется с использованием просодии отрытого списка:

(26) ВИЖ
··· Как /–известно,
·· мы ходим на \рыбалку.
·· Для рыбалки нам нужны –жучки-паучки-червячки,,,
·· /–мухи,,,
··· и прочие летающие \насекомые.
У нас \нет проблем ловить,
{ШМЫГАНЬЕ} эту \живность.
Мы просто открываем –окна,,,
·· и на мамины –гобелены,,,
·· каждую –весну,,,
··· шмели ·· мухи ·· и насекомые слетаются целыми \кучами.
И мы их просто снимаем с этого \гобелена.

Поскольку кластеры средств нечеткой номинации возникают в болезненных точках речепорождения, они часто сопровождаются метавысказываниями, в которых говорящий эксплицирует трудности с выбором номинации, ср. (27):

(27) РОС
и /мне вдруг встречается **абсолютно** || ··(0.2) /\челов<u>е</u>к,
который я не’ ==
которого я не /зн<u>а</u>ю,
и не /зн<u>а</u>л-л ээ(0.4) до= || ”(0.2) никогда,
····(1.2) /и’ ···(0.8) я’ ··(0.3) \говор<u>ю</u>:
····(1.4) «/Приветh ~»
··(0.4) эээ(0.8) ··(0.4) То ли /–Ф<u>е</u>дька…
или-и /–кт<u>о</u>-то…
я не \п<u>о</u>мню,
чего-нибудь /сказ<u>а</u>л там,
····(1.2) какое-то \<u>и</u>мя назвал,

4. Заключение

Данные живой речи позволяют убедиться, что в тех случаях, когда говорящий испытывает трудности при подборе адекватной точной номинации или просто избегает точной номинации в силу тех или иных прагматических причин, он обычно использует целый комплекс сигналов разного уровня, которые предупреждают слушающего об этой ситуации. Наряду с лексическими маркерами-заместителями и маркерами-аппроксиматорами эффективным средством нечеткого номинирования могут служить и чисто синтаксические стратегии, подкрепляемые в устной речи специализированными просодическими паттернами; это, прежде всего, относится к сочинительным конструкциям, выражающим незавершенность «открытого ряда», и к особым супрасегментным средствам их воплощения. Потребность в нечеткой номинации возникает у говорящего в проблемных точках порождения дискурса, и преодоление проблемы не всегда совершается в один шаг, поэтому типичным является использование сразу нескольких сигналов об имеющейся проблеме и, в частности, более одного показателя нечеткой номинации. Дальнейшее изучение задокументированных образцов живой речи позволит уточнить номенклатуру разноуровневых средств, обслуживающих эту зону значений.

Литература

1. *Адамович С. В.* (2011), Семантическая категория аппроксимации и система средств ее выражения. Гродно: ГрГУ им. Я. Купалы.
2. *Кибрик А. А., Подлесская В. И. (Ред.).* (2009). Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
3. *Янко Т. Е.* Интонационные стратегии русской речи в сопоставительном аспекте. ЯСК. Москва, 2008.

References

1. *Adamovich S. V.* (2011), Semanticheskaja kategorija approksimacii i sistema sredstv eë vyrazhenija [Semantics of approximation and the system of approximators], Grodno: J.Kupala State University of Grodno.
2. *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.
3. *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.
4. *Andersen, Gisle.* (1998). The pragmatic marker *like* from a relevance-theoretic perspective. A. H. Jucker & Y. Ziv (Eds.), Discourse markers: Descriptions and theory). Amsterdam: Benjamins, pp. 147–170
5. *Channell, Joanna.* (1994). Vague Language. Oxford: Oxford University Press,
6. *Enfield, Nicholas James.* (2003). The definition of what-d'you-call-it: Semantics and pragmatics of recognitional deixis. Journal of Pragmatics 35, pp. 101–117.
7. *Fox Tree, J. E.* (2006). Placing *like* in telling stories. Discourse Studies 8, pp. 723–743.
8. *Fox Tree, Jean E.; Tomlinson Jr., John M.* (2008). The Rise of *Like* in Spontaneous Quotations. Discourse Processes 45:1, pp. 85–102.
9. *Jucker, Andreas H.; Smith, Sara W.; Lüdge Tanja.* (2003). Interactive aspects of vagueness in conversation. Journal of Pragmatics 35, pp. 1737–1769.
10. *Hayashi, Makoto; Yoon, Kyung-Eun.* (2006). A cross-linguistic exploration of demonstratives in interaction: with particular reference to the context of word-formulation trouble, Studies in Language 30-3, pp. 485–540.
11. *Kaltenböck, Gunther; Mihatsch, Wiltrud; Schneider, Stefan (Eds.).* (2010). New Approaches to Hedging. [Studies in Pragmatics] Emerald.
12. *Lakoff, G.* (1972). Hedges: A study of meaning criteria and the logic of fuzzy concepts. P. Peranteau, J. Levi and G. Phares (eds.) Papers from the English Regional Meeting of Chicago Linguistic Society,. Chicago: Chicago University Press. pp. 183–228.
13. *Markkanen R., Schröder R.* (Eds.). (1997). Hedging and Discourse Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts, Berlin, New York: Walter de Gruyter, pp. 188–207.
14. *Podlesskaya Vera I.* (2010). Parameters for typological variation of placeholders // Nino Amiridze, Boid H.Davis and Margaret Maclagan (eds.) Fillers, Pauses and Placeholders. [Typological Studies in language (TSL), vol. 93]. Amsterdam/ Philadelphia: John Benjamins, pp. 11–32.
15. *Sperber, Dan & Wilson, Deirdre.* (1991). Loose talk. Steven Davis (ed.), Pragmatics. A Reader. Oxford: OUP, pp. 540–549.

ГРАММАТИЧЕСКИЙ СЛОВАРЬ ДЛЯ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ XVIII–XIX ВЕКА: ПЕРВЫЕ РЕЗУЛЬТАТЫ[1]

Поляков А. Е. (pollex@mail.ru)
НПБ им. К. Д. Ушинского РАО, Москва, Россия

Савчук С. О. (savsvetlana@mail.ru),
Сичинава Д. В. (mitrius@gmail.com)
Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

В статье излагаются основные принципы построения грамматического словаря и морфологического анализатора для текстов XVIII–XIX вв. веков с учётом орфографических, морфологических и лексических особенностей языка этого периода, выявленных на материале текстов Национального корпуса русского языка. Поскольку от данного анализатора требуется универсальность подхода и возможность работы с текстами разных типов и разными орфографическими режимами, то он должен состоять из нескольких модулей, применяемых к текстам различных типов в зависимости от степени проявления в них тех или иных орфографических и грамматических явлений. Его словарь построен на базе существующего грамматического словаря современного русского языка, а также словарей XIX в. и текстов Национального корпуса. Обсуждается несколько альтернативных возможностей реализации орфографических (предобработка, применение технологии параллельного корпуса, нормализация при разметке) и морфологических правил (нормализация при разметке, добавление нестандартных форм в парадигму). Проводится оценка первых результатов применения анализатора к текстам НКРЯ и предлагаются различные варианты улучшения результатов (введение новых правил, пополнение словаря и т. д.).

Ключевые слова: грамматический словарь, автоматический анализ текстов XVIII–XIX вв.

[1] Работа выполнена при поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика».

A GRAMMAR DICTIONARY FOR AUTOMATIC ANALYSIS OF THE XVIII–XIXTH CENTURY TEXTS: FIRST RESULTS

Polyakov A. E. (pollex@mail.ru)
Ushinsky's State Scientific Pedagogical Library RAE

Savchuk S. O. (savsvetlana@mail.ru),
Sitchinava D. V. (mitrius@gmail.com)
V. V. Vinogradov's Institute for the Russian Language RAS, Moscow, Russia

The paper presents the key principles of building a grammar dictionary and a morphological analyzer for XVIII–XIXth century Russian texts based on orthographical, morphological and lexical features exemplified by the Russian National Corpus (RNC). The analyzer should involve different modules applicable to different kinds of texts depending on their respective orthographical and grammatical phenomena. Several alternative ways of implementing orthographical and morphological rules are discussed (including pre-processing, online normalization etc.). Evaluation data of the first analysis results are presented.

Key words: grammar dictionary, automatic text analysis, texts of the XVIII–XIXth century

1. Постановка проблемы и способы ее решения

Последнее десятилетие характеризуется ростом интереса к сохранению письменного наследия — текстов предшествующих эпох, в частности, к оцифровке этих текстов и увеличению возможностей доступа и поиска (играет здесь роль и то, что старые тексты, созданные, как минимум, век назад, не охраняются авторским правом, находясь в общественном достоянии). Наблюдается рост числа исторических корпусов, электронных исторических библиотек; объёмы оцифрованных старых изданий, доступных в электронном виде, существенно выросли. Нужно упомянуть, в частности, систему Google Books, где возможен полнотекстовый поиск на разных языках, в том числе русском, по книгам и журналам, для раннего периода нередко представленным в полном виде и доступным для скачивания. Этот поиск уже становится неотъемлемым инструментом работы специалистов самого разного профиля, занимающихся историей тех или иных явлений. Уровень технологий вырос и позволяет изготавливать электронные издания в графических форматах на достаточно высоком уровне и в большем объёме.

Однако основная проблема создателей исторических корпусов и электронных библиотек с поиском по тексту по-прежнему остается открытой. Для морфологического анализа текстов предшествующих эпох (если такая задача вообще ставится) используется анализатор, рассчитанный на современный язык и частично (или даже полностью) на современную орфографию, потому что другого пока нет. Более того, от качества работы такого анализатора зависит уже качество распознавания отсканированного печатного текста в орфографии XVIII–XIX вв.; словоформы, не предсказываемые современным грамматическим словарём, могут быть распознаны неправильно (например, слово *пылъ* с конечным ером — как *пыль,* а *красныя* — как *красная* или *красный*).

Морфологическая разметка в Национальном корпусе русского языка производится (полу)автоматически с помощью анализаторов Dialing (подкорпус со снятой омонимией) и Mystem (тексты с неснятой омонимией в составе разных подкорпусов). В основе обоих анализаторов лежит грамматический словарь современного русского языка [Зализняк 1977/2003]; словарь анализатора Mystem, по сравнению со словарём Зализняка, пополнен на материале часто встречающихся в Интернете и в поисковых запросах неологизмов 1990–2000-х годов и имён собственных. Результат такой разметки дает значительное количество ошибочных и/или гипотетических разборов. Как было показано в [Савчук, Сичинава 2009], процент погрешностей в разборе текстов XVIII в. находятся на уровне современных текстов, не нормированных (тексты электронной коммуникации) или в принципе не ориентированных на литературную норму (записей диалектной речи).

Повышение качества анализа текстов с большим количеством отклонений от стандарта предлагалось проводить двумя путями: 1) использование стандартного анализатора, отражающего грамматические и орфографические нормы современного литературного языка, с подключением дополнительных правил для отдельных словоформ и категорий слов и 2) использование нового морфологического анализатора, настроенного на определенные тексты. Первый путь до сих пор использовался для улучшения качества разметки НКРЯ, однако в большом пополняющемся корпусе список словоформ, отклоняющихся от стандарта и требующих применения дополнительных правил, неуклонно растет, что делает этот способ неэффективным.

Между тем база текстов НКРЯ существенно растёт. Расширился объём текстов XVIII в. (к февралю 2012 г. 3,8 млн слов) и текстовпредшествующего периода — среднерусских текстов XIV–XVII веков (3 млн слов). Увеличение объема исторических текстов, а также их хронологического разнообразия делает более актуальным использование нового анализатора. Морфологической анализатор церковнославянского языка, разрабатываемый для церковнославянского подкорпуса НКРЯ [Поляков и др., 2012], для этой цели в текущем виде не подходит (его пробное применение к среднерусским текстам, без орфографической настройки, показало, что опознаётся лишь примерно половина словоформ). Он рассчитан на современную церковнославянскую орфографию (выработавшуюся с XVII в. и отличную как от предшествующей церковной, так и от последующей гражданской: с последовательным сложным распределением

пар омофоничных букв и т. п.), а также морфологический и лексический стандарт богослужебных текстов последних веков, когда церковнославянский язык ограничивается собственно функцией языка православной литургии. Между тем тексты XIV–XVII в., не говоря уже о XVIII–XIX вв., используют церковнославянские элементы лишь в комбинации с собственно русскими, и в ряде жанров собственно русские черты на всех уровнях существенно преобладают.

Возможно применить к текстам одновременно церковнославянский анализатор (ослабив в нём требования к орфографии) и современный анализатор, наоборот, добавив в него по крайней мере правила, упразднённые реформой 1918 года. Такая комбинация церковнославянского и современного анализаторов дает сравнительно неплохие результаты, но оставляет неразобранной некнижную и нецерковную лексику и грамматику, отсутствующую как в современном русском языке, так и в современном церковнославянском (например, компаратив типа *сильняе* или множественное число среднего рода типа *злодействы*), не говоря уже о словообразовательной и фонетической вариативности и проблеме совмещения разборов обоих анализаторов для совпадающих словоформ. Поэтому разработка нового анализатора должна вестись с учетом специфики конкретных текстов, при этом предусматривая появление в новых текстах новых аналогичных словоформ, для которых должна быть предусмотрена возможность правильного разбора.

2. Специфика корпуса исторических текстов

Корпус исторических текстов отражает эпоху, когда корреляция между языком и жанром была гораздо сильнее и прямолинейнее, чем в современной языковой ситуации. Общеизвестно принятое русским классицизмом разделение литературного языка на три «штиля» (высокий, средний и низкий), определенных, в том числе, по насыщенности текста славянизмами; каждый из них был закреплён за своим жанром. Для среднерусского периода было характерно сосуществование текстов с более сильными церковнославянскими либо народными тенденциями; представление о «диглоссии» — жёстком распределении жанров между двумя языками — несколько упрощено, хотя и отражает определённые установки писавших. Если верно, что церковно-богословские тексты, такие, как проповедь, ориентировались (в период до 1740-х гг.) на чистый церковнославянский язык, а бытовые тексты, такие, как «грамотки» XVII–XVIII в. или личные дневники более позднего времени — на разговорный русский, то столь же верно и то, что одновременно определённые народные языковые черты проникали и в первые, а церковнославянские — и во вторые. Более равноправное комбинирование славянских и народных элементов было характерно для языка летописей, посланий (так называемый «гибридный язык»), позже — официальных документов, дипломатических и учебно-научных текстов. Наконец, необходимо отметить и такое явление, характерное и для среднерусской эпохи, и для языка Нового времени, как церковнославянские цитаты (из Библии, тогда еще не переведённой на русский, или из богослужебных текстов, до настоящего времени в общепринятой практике

церковнославянских) в составе русских текстов. Это явление присутствует и в современном языке как в виде фразеологизмов-славянизмов, в том числе библеизмов (*ничтоже сумняшеся, возвращается ветер на круги своя*, *коемуждо по делом его* и т.д.), так и собственно цитат, причём не только в специально богословских текстах. Явлением более или менее тесного взаимопроникновения двух систем, их «гибридизации», собственно говоря, и объясняется необходимость учета элементов лексики и грамматики фактически другого языка при анализе русских текстов.

Тексты разных периодов могут значительно отличаться по характеру языка, поскольку конец XVII—середина XIX в — это самый интенсивный период формирования литературного языка нового типа. По количеству заимствованной лексики, по употребительности славянизмов, по орфографической нормированности достаточно существенно различаются между собой петровская эпоха, период классицизма, эпохи, связанные с именами Карамзина и Пушкина.

До недавнего времени Национальный корпус русского языка включал ранние тексты (XVIII — первая половина XX в.; о среднерусском периоде речь не идёт) только в современной орфографии, или, по крайней мере, с некоторыми отклонениями от современной нормы, но в целом ориентированные на орфографический режим 1918 г. или даже 1956 г. [Савчук, Сичинава, Гарипов 2006]. Это диктовалось, помимо традиции, и ориентацией на авторитетные научные издания советского и постсоветского времени. Вместе с тем давно обсуждается вопрос о включении в корпус значительного количества текстов в дореформенной орфографии [Соловьев, Ахтямов 2006; Савчук 2008], с сохранением упразднённых реформой 1918 года графем и орфографических правил, тем более что анализатор Mystem умеет учитывать большинство этих правил (использование *ѣ, ѳ, і*, конечного *ъ*, окончания *-ыя, -ія, -аго, -яго*). Ценность оригинальной орфографии для филологического изучения текста сейчас не нуждается в особом аргументировании (см., в частности, работы М. И. Шапира). В настоящее время в экспериментальном порядке в основной корпус НКРЯ входит один текст целиком в дореформенной орфографии по изданию 1857 года — роман «Черная рада» Пантелеймона Кулиша, впервые распознанный и вычитанный для Корпуса с прижизненного издания; его изданий в новой русской орфографии, кажется, не существует, поскольку наиболее актуальной для читателя уже порядка ста лет считается авторская версия этого текста на украинском языке. Вообще для текстов, никогда не переиздававшихся в новой орфографии, сохранение в Корпусе дореформенного правописания кажется особо важной задачей; безвозвратная утрата этой информации после трудоёмкого сканирования и вычитки текста была бы совершенно нерациональна. Вместе с тем встаёт проблема обработки и совместного поиска по текстам разных типов. Например, для основного корпуса НКРЯ желательна возможность индексировать корпус так, чтобы при поиске точных форм можно было найти одновременно и дореформенное, и современное написание (например, чтобы по точному запросу словоформы *пѣной* находилась бы и словоформа *пеной* в новоорфографических текстах).

3. Принцип работы анализатора. Составные элементы анализатора

Можно сформулировать требования, которым должен отвечать анализатор для обработки исторических текстов:

А) *Универсальность*: анализатор должен уметь работать с текстами, обладающими различными характеристиками:

1) обрабатывать тексты в новой и дореформенной орфографиях, учитывая информацию, которая несёт каждая из них (например, не просто приравнивать омофоничные буквы, но и отличать *всѣ* от *всё*);
2) обрабатывать тексты разных жанров от религиозных, написанных на церковнославянском или приближенном к нему, до бытовых писем в свободной орфографии (далекой от нормализации). В перспективе следует предусмотреть использование анализатора для обработки современных текстов, содержащих отступления от литературной нормы иного рода: текстов электронной коммуникации, записей речи на региональных вариантах русского языка (бытующих в русской диаспоре или иноэтнической среде) и пр.

Б) *Открытость* — способность пополняться и видоизменяться, настраиваться на разные типы текстов, возможность «обучаться» на основании пополненного словаря и т. п.[2]

Несмотря на требование универсальности, одновременно могут сосуществовать и модификации анализатора для текстов разных периодов и/или жанров, использующие ряд правил, специфических именно для этих текстов. Например, к текстам XX–XXI вв. едва ли нужно применять правило, согласно которому частицы *б(ы), ж(е)* и *ли/ль* могут писаться слитно с предыдущим словом (что особо часто встречается в XVIII в.); это приведёт исключительно к паразитическим разборам типа *мыль = мы + ль, стали = ста + ли, ниже = ни + же* и др. (в текстах XVIII в. разборы такого рода можно отсеивать при помощи специального правила). Аналогично, такие специфические для письменности раннего XVIII в., а также XVII и предшествующих веков правила, как пропуск мягкого знака *(толко = только)* и отсутствие в графической системе буквы *й* (*таино = тайно*) приведёт к излишним неправдоподобным разборам в современных текстах (типа *банка = банька, заика = зайка* и т. п.). Список периодов и жанров, требующих особых модификаций анализатора, нужно будет установить опытным путём после анализа текстов, входящих в Национальный корпус русского языка. В частности, можно предположить, что различные модификации потребуются для текстов бытовых грамоток XVII в., для текстов разных периодов,

---

[2] В отношении несовременных текстов (представляющих собой по определению закрытый, хотя и очень большой класс) о принципе открытости анализатора можно говорить с известной долей условности, однако это обстоятельство не играет практической роли до тех пор, пока все или почти все исторические тексты не будут включены в корпус. В настоящее время мы еще очень далеки от этого (достаточно сказать, что не только оцифровано, но и вообще издано лишь незначительное меньшинство сохранившихся от XVII–XVIII вв. текстов, в том числе и бытовых текстов в нестандартной орфографии).

ориентированных на церковнославянский язык, для собственно русских текстов XVIII в., первой половины XIX в. и нескольких дальнейших периодов.

Исходя из этих требований, перед разработчиками стоят два типа задач.

1. *Грамматические*: разработка грамматических парадигм для лексем, отсутствующих в современном словаре.

2. *Орфографические*: обеспечение лемматизации форм, имеющих отклонения от стандартных написаний. К решению второй (орфографической) задачи имеется ряд возможных подходов.

1) Предобработка (preprocessing) — нормализация теста, своего рода «перевод» его на стандартный язык, предваряющая морфологический анализ. Нормализация текста широко применяется в устных и диалектных корпусах разных языков и в подкорпусе электронной коммуникации НКРЯ [см. об этом Гришина, Савчук 2009] в ручном режиме. При подготовке диалектных текстов для диалектного подкорпуса НКРЯ предобработка текстов, записанных в фонетической транскрипции, осуществлялась в полуавтоматическом режиме по технологии, разработанной И. Б. Качинской и Т. А. Архангельским при участии одного из соавторов данной статьи. Сначала с исходным текстом в фонетической транскрипции (так называемый текст-1) работал автоматический модуль-детранскриптор, переводящий транскрипцию в условную орфографическую запись (текст-2), а потом эта запись редактировалась и вручную переводилась (в части грамматики и фонетики) на литературный язык; этот перевод (текст-3) в дальнейшем автоматически анализировался при помощи Mystem, после чего в нём вручную снималась омонимия и проставлялась специфическая для диалектного языка разметка [Качинская 2010]. Для исторических корпусов английского языка в автоматическом режиме применяется модуль VARD [Baron, Raison 2009], который, по оценке его авторов, дает хорошие результаты при обработке широкого спектра текстов с ненормативной орфографией. Для орфографической нормализации французских текстов эпохи Возрождения разработан инструмент VariaLog, который в принципе может быть использован и для других языков [Lay 2012]. Недостатки метода предобработки заключаются в значительной затрате ресурсов для подготовки нормализованной версии текста и зачастую неоднозначности такой нормализации при объёме корпуса в несколько миллионов слов[3].

---

[3] Вот как выглядел бы фрагмент текста после предварительной обработки (в данном случае нормализация потребовалась для половины словоформ):

```
<distinct form="Што">что</distinct>
<distinct form="касаетца">касается</distinct>
<distinct form="да">до</distinct> брата князь Александра
<distinct form="Михаиловича">Михайловича</distinct> я вам
<distinct form="ево">его</distinct>
<distinct form="ваяж">вояж</distinct> в
<distinct form="первам">первом</distinct>
<distinct form="писме">письме</distinct>
<distinct form="обстаятелна">обстоятельно</distinct>
<distinct form="аписывал">описывал</distinct> с
<distinct form="челавекам">человеком</distinct> Тургенева
```

2) Применение технологий параллельного корпуса — одновременное использование как оригинального текста, так и перевода орфографии на современные нормы, выровненных по предложениям или даже словоформам [Meyer 2009]. Национальный корпус русского языка уже поддерживает параллельную технологию — выравнивание текстов по предложениям, которая используется, прежде всего, в двуязычных и многоязычном параллельных подкорпусах НКРЯ (ср. [Добровольский, Кретов, Шаров, 2005], [Sitchinava 2012]). Фактически эта технология представляет собой расширение предыдущей: нормализованный текст становится доступен пользователю. Существенным недостатком такого подхода, с нашей точки зрения, является то, этот нормализованный текст по крайней мере в значительной части случаев не будет опираться на существующие критические или массовые издания (как это бывает в случае включения в корпус старых текстов в пореформенной орфографии), а будет представлять собой продукт деятельности разработчиков корпуса. Тем самым последние фактически берут на себя также функции текстологов, готовящих для широких кругов пользователей претендующее на научность издание старого текста в новой орфографии. Как представляется, подобный подход для нужд исторических корпусов НКРЯ нецелесообразен.
3) Учет различных уровней вариативности в грамматическом словаре анализатора. Именно данный подход лежит в основе предлагаемого в настоящей публикации. В грамматический словарь анализатора, выстроенный на базе лексикографических источников, дополнительно включаются словоформы, не распознаваемые или распознаваемые неправильно и неполно существующими анализаторами. Одновременно работают орфографические правила (они могут быть также включены в механизм индексации корпуса), позволяющие опознать в цепочке, отсутствующей в словаре, то или иное словарное слово (разбор слов с *і* как слов с *и*, разбор слов на *-тца* как слов на *-ться*); это алгоритмическая нормализация, не фиксируемая в виде какого бы то ни было промежуточного текста. При этом могут быть включены отдельным списком блокирующие правила для случаев частотных «паразитических» разборов (ср. обсуждаемые выше случаи типа *ниже = ни + же, стали = ста + ли*). Применение данного подхода к анализу конкретных текстов разных типов и периодов покажет, в какой степени он позволит сократить количество неправильных разборов, и, возможно, обнаружит участки, в принципе не поддающиеся автоматическому анализу и требующие ручной нормализации.

```
<distinct form="пакоинава">покойного</distinct> Ивана
<distinct form="Сергеича">Сергеевича</distinct>, и
<distinct form="пажалавал">пожаловал</distinct>
<distinct form="кушел">кушал</distinct> у меня на
<distinct form="другои">другой</distinct> день
<distinct form="возвароту">возвороту</distinct>
<distinct form="сваево">своего</distinct>
<distinct form="ис">из</distinct> Теплых Станов, и с тех пор не видал;
```
[Вас. Бор. Голицын Влад. Бор. Голицыну 8 апреля 1771 г.].

4. Грамматический словарь

4.1. Общие принципы

Общие принципы и алгоритм работы анализатора, области его применения были изложены в [Поляков 2012]. Грамматический словарь определяется как список лексем языка с приписанной информацией о словоизменении. Каждая лексема в словаре содержит, как минимум, следующую информацию: 1) основа с указанием чередований; 2) постоянные признаки лексемы (часть речи, род, одушевленность, переходность, и т. д.); 3) код словоизменительного типа (парадигмы). Вот пример записи для некоторых глаголов с чередованиями в основе:

но(с|ш)+ить	V,ipf,tr	V4
б(и|ь|е)+ть	V,ipf,tr	V11
пе(к|ч|)+ь	V,ipf,tr	V8
ж(г|ж|ег|е)+чь	V,ipf,tr	V8*g

Грамматический словарь анализатора складывается из нескольких модулей, соответствующих различным периодам истории русского языка. При анализе конкретного текста выбирается модуль, соответствующий типу и периоду создания текста.

1) Современный модуль строится на основе Грамматического словаря Зализняка [Зализняк 1977/2003], который фиксирует лексический и грамматический стандарт конца XX века. Тем не менее, этот модуль позволяет анализировать значительную часть словоформ, встречающихся в текстах XVIII–XIX века, если они не имеют существенных орфографических и грамматических отличий от современной нормы.
2) Модуль XVIII–XIX века строится на основе анализа корпуса реальных текстов, а также исторических словарей русского языка, включая:
   - Словарь Академии Российской (1789–1794);
   - Словарь церковнославянского и русского языка (ЦСРЯ) (1847);
   - Полный русский орфографический словарь (1898);
   - Словарь русского языка XVIII века.
3) Модуль XVII века и более ранних периодов строится в основном на основе анализа корпуса реальных текстов и лишь частично на основе исторического словаря (Словарь русского языка XIV–XVII века).
4) Модуль для церковнославянского языка сейчас существует в рамках отдельного церковнославянского корпуса, а возможность его применения для исторического корпуса пока неочевидна.

Для адекватного анализа текстов XVIII–XIX века, часть которых представлена в дореформенной орфографии, необходимо добавить в грамматическую модель анализатора следующие формы:

1) формы, характерные для всех периодов:

- деепричастия совершенного вида от основы презенса (*прийдя, увидя, взгромоздясь*), которые вполне употребительны в современном языке, а тем более в языке XVIII–XIX века;
- вариант частицы *-ся* после гласных (*валюся, валилася*), который употребляется в современном языке (в некоторых идиолектах), а также в языке XVIII–XIX века;
- сравнительная степень на *-ей* (*сильней*) и с префиксом *по-* (*посильнее, посильней*);

2) формы, характерные для XIX века и более ранних периодов:
- адъективные флексии (*-аго/-яго, -ыя/-ія*);
- особые формы местоимений (*ея, онъ, однъ, однъхъ*);
- творительный падеж 3-го склонения на *-ію* (*милостію, помощію*);

3) формы, характерные для XVIII века и более ранних периодов:
- усеченные формы прилагательных (*красна/о/ы/у*), которые формально совпадают с краткими формами, но имеют другое грамматическое значение;
- сравнительная степень на *-яе* (*сильняе, скоряе*);
- глагольные флексии *-ти* и *-ши* (*ходиши, ходити*), которые, скорее всего, должны трактоваться как церковнославянизмы (см. ниже).

4) церковнославянские формы:
- формы имперфекта (*творяше, творяхомъ, творяху*) и аориста (*творихъ, творихомъ, твориша*), которые нередко бывают омонимичны (*делахъ, делахомъ*);
- частотные формы косвенных падежей существительных (*градомъ, градъхъ, отроцы, отроцъхъ*);
- частотные лексемы (*иже*, *яко, понеже, вельми*);

и т. д.

Анализатор, помимо словаря, сопровождается отдельными правилами для анализа текстов в разных орфографических режимах — алгоритмической нормализацией, одновременной с приписываемой морфологической разметкой. Они работают по следующему принципу: если словоформа не предсказывается на основании грамматического словаря (а может получить только гипотетический разбор), то предпринимаются попытки, исходя из её буквенного состава, регулярной замены в ней тех или иных букв и анализа получившейся нормализованной словоформы. Если эта словоформа получает негипотетический анализ (и, возможно, этот анализ удовлетворяет тем или иным грамматическим ограничениям), то он подставляется в качестве её разбора. Таковы графические (типа *ѳ => ф,* см. ниже, 4.3.А) и орфографические (типа *цы => ци*, см. ниже, 4.3.Б) правила, применимые ко всем словоформам, удовлетворяющим данным критериям.

Вместе с тем должен быть задействован также ряд индивидуальных правил, вводящих конкретные орфографические варианты для конкретных лемм; например, такова индивидуальная вариативность типа *естьли~если* (неизменяемое), *потчевать~подчивать, ветчина~вядчина* (проводится по всей парадигме).

4.2. Источники для формирования словника грамматического словаря

Источниками для формирования словника грамматического словаря являются, с одной стороны, существующие исторические словари русского языка (см. выше, п. 4.1), с другой — результаты анализа корпуса реальных текстов. Лексемы из этих двух источников будут добавляться к основному модулю, составленному на основе Грамматического словаря Зализняка. Этот путь представляется нам самым коротким для достижения практических целей — адекватного анализа текстов исторического корпуса. Почему нецелесообразно ограничиваться только словарями? Прежде всего потому, что значительную часть словника исторических словарей составляют книжные и устаревшие слова (включая церковнославянизмы), которые не очень часто встречаются в реальных текстах, особенно в художественных и бытовых. В то же время широко представленные в текстах собственные имена в словарях не описаны. Таким образом, создание электронного грамматического словаря исключительно на базе лексикографических источников, значительная часть которого не будет «работать» при анализе текстов, кажется нам неэффективным. Напротив, словник, полученный на основе текущего состояния корпуса и пополняемый при добавлении новых текстов, как раз и будет отражать реальное употребление.

Корпусной словник создается на базе частотного списка словоформ, извлеченных из текстов, которым при автоматическом анализе приписываются леммы и грамматические характеристики. Большая часть словоформ успешно разбирается современным анализатором; меньшую часть (около четверти) составляют словоформы, которые отсутствуют в современном грамматическом словаре. При автоматическом анализе они получают гипотетические разборы, в дальнейшем им вручную приписываются правильные леммы и грамматическая информация. По составу это а) церковнославянская лексика (актуальная для корпуса), частично перекрывается с лексикой из ЦСРЯ б) варианты, в) собственные имена и отыменные прилагательные [Савчук 2012].

Пополнение словника грамматического словаря из обоих источников будет осуществляться поэтапно, по мере подготовки электронных версий словарей, одной стороны, и развития корпуса исторических текстов, с другой.

4.3. Методы анализа вариативности.

Для правильного анализа и сведения к единой лемме языковых вариантов на различных уровнях анализатор может использовать правила следующего типа.

А) *Графические правила*

Графические правила основаны на приравнивании графем, встретившихся в тексте, графемам, входящим в порождаемые словарём формы, например, *ѳ => ф; ѯ => кс*. В текстах, где распределение омофоничных графем

было неустойчивым и зависело от конкретной орфографической школы (прежде всего это среднерусские тексты и тексты раннего XVIII в., особенно бытовые), может быть задействован модуль, рассматривающий такие пары букв как равнозначные варианты и при нормализации заменяющий один на другой глобально. Вместе с тем для текстов, где достаточно устойчиво установилась смыслоразличительная функция ряда омофонов (например, в русской «гротовской» орфографии *миръ* vs. *міръ*), может использоваться модуль, учитывающий эти различия при простановке леммы и морфологическом разборе.

Б) *Орфографические правила*

Орфографические правила связаны с нормализацией определённых орфограмм: иными словами, определённые буквы заменяются на другие лишь в контексте некоторых третьих, при этом обе буквы присутствуют в алфавите грамматического словаря. Таково, например, правило *цы => ци*, при соблюдении которого стандартизированный разбор получают написания типа *цыновка, цыдулка, цыгарка, цыгейка*, нормативные до 1956 г., причём некоторые из них встречаются в текстах и после этой даты, или правило *тца => ться/тся*, при котором восстанавливаются оба (финитный и инфинитивный) разбора для глагольных словоформ, где конечное сочетание фонетически совпало и в таком виде до XVIII в. отражалось на письме в текстах, не считавшихся грубо неграмотными. Особую роль такие правила играют в бытовой письменности XVII–XVIII вв., где полностью разрешёнными приёмами (как и, например, в поздних берестяных грамотах) является упрощение двойных согласных, передача на письме оглушения и озвончения, пропуск знака для ь и т. п.

Пример работы орфографического правила:

в тексте представлена словоформа *надеютца;*
из-за наличия конечного *тца* проверяется правило *тца=ться/тся*;
словоформа *надеються* словарем не порождается;
словоформа *надеются* словарем порождается и разбирается как lex=НАДЕЯТЬСЯ gr=praes,3p,pl;
словоформа *надеютца* получает разбор lex=НАДЕЯТЬСЯ gr=praes,3p,pl=distort.

Аналогично словоформа *носитца*, для которой возможны две нормализации, получает два разбора — инфинитивный *(носиться)* и финитный *(носится),* а словоформа *отца* только разборы от слова ОТЕЦ (ввиду отсутствия словоформ **от(ь)ся*).

В) *Морфологические правила*

Морфологические правила, в отличие от орфографических, устроены с учётом информации о грамматическом разборе словарной словоформы. Например, вводится правило, которое можно записать как *-яе*, comp,anom <=

-ее, comp. Это значит, что если в тексте встретилась не получающая словарного анализа словоформа, кончающаяся на *-яе*, например, *сильняе*, а замена *-яе* на *-ее* в этой словоформе даёт словарную форму (*сильнее*), и притом эта форма есть словоформа сравнительной степени, то словоформа типа *сильняе* получает такой же грамматический разбор, что и словоформа типа *сильнее* плюс помету anom («аномальная морфологическая форма»). Правило это может быть сформулировано и более широко, так, чтобы учитывать односложные окончания *(сильняй)* и префиксальную сравнительную степень (*посильняе*). Аналогичные правила могут вводить (с добавлением пометы anom) ненормативные варианты постфикса *-ся* вместо *-сь (оставалося)* или наоборот *(запершегось)*.

Альтернативный способ получения адекватного морфологического анализа несловарных форм состоит в том, чтобы включить нужные формы в состав парадигм. В процессе совершенствования анализатора предполагается опробовать оба этих способа.

Пример применения морфологического правила:
в тексте представлена словоформа *имееши*;
из-за наличия конечного *ши* проверяется правило *ши=шь*;
словоформа *имеешь* словарем порождается и разбирается как
lex=ИМЕТЬ gr=praes,2p,sg;
из-за наличия разбора 2p,sg словоформа имееши получает разбор
lex=ИМЕТЬ gr=praes,2p,sg=anom.

При этом, например, несловарная словоформа *воши* не получает разбора, идентичного разбору словоформы *вошь* (поскольку она не глагольная).

Г) *Списочный способ* — задание вариативности списками, на ограниченных классах единиц. Таково, например, орфографическое правило, согласно которому ряд конкретных корней может (или даже практически обязан) в среднерусских и церковнославянских текстах сокращаться (записываться под титлом), ср. такие словоформы, как *Г(о)с(по)дь, м(е)с(я)ц, гл(аголе)ть* и т. д. К списочным правилам относятся и блокирующие правила для частотных паразитических разборов типа *стали = ста ли*.

Указанные правила имет определенный порядок применения; так, в первую очередь применяются графические правила, в том числе списочные; затем — орфографические, морфологические и списочные правила, блокирующие паразитические разборы.

В процессе формирования словаря и оптимизации работы анализатора предусматривается несколько циклов обработки текстов. Каждая новая версия анализатора будет проходить проверку на корпусе: после анализа результатов в словарь и правила анализатора будут вноситься пополнения и коррективы, после чего этот цикл повторяется уже с новой версией. Далее мы изложим оценку результатов первой версии анализатора на корпусе текстов XVIII–XIX веков.

5. Оценка результатов

5.1. Состав несловарных словоформ

Первый вариант анализатора на основе первой версии словаря был опробован на экспериментальном корпусе объемом около 4 млн словоупотреблений[4], который включает 256 тыс. различных словоформ. Результаты представлены в таблице.

разобрано	185 221	72,4 %
гипотезы	63 904	25,0 %
не разобрано	6 780	2,6 %

Наибольший интерес для дальнейшей разработки словаря представляет анализ словоформ, которые не были опознаны как слова русского языка или получили гипотетические разборы, и оценка предложенных гипотез. Список неопознанных форм в настоящее время полностью проанализирован, всем формам вручную приписаны леммы, и они пополнили список разобранных форм. Перечислим наиболее массовые случаи.

Не распознаны или неправильно опознаны сочетания знаменательных частей речи с частицами *-то(-та, -ат), же(ж), ли(ль), бы(б), -де, -ка*, в написании которых в разные периоды наблюдались колебания. Согласно современным правилам, частицы *-то, -де, -ка* пишутся через дефис, *бы, ли, же* — раздельно. Раздельное написание этих частиц рекомендовалось уже в XIX в [Грот, 1873], однако и в XX в. вплоть до реформы 1956 г. активно использовались дефисные написания. В изданиях XVIII в. в написаниях частиц не было последовательности, можно встретить дефисные, слитные и раздельные написания: *них-же, месте-же, нихже, мыже, таковаже, пили-б, ожидали-б, где-б, пилиб, ожидалиб, еслиб.* В НКРЯ сочетания с раздельным и дефисным написанием частиц анализируются как две леммы, при этом автоматический анализ в большинстве случаев правильный. Эти же решения следует использовать и в исторической части словаря. Основную трудность как в НКРЯ, так и в конструируемом анализаторе представляет правильная идентификация частиц в составе сочетаний при слитном написании (*ежелиж*, *былаб* и под.).

Другую большую группу словоформ, не получивших разборов, составляют архаические формы склонения и спряжения. Среди них заметное

---

[4] Экспериментальный корпус составлен на основе текстов XVIII — первой трети XIX вв., входящих в состав НКРЯ. По жанровому составу корпус XVIII в. разнообразен: доля художественных текстов и публицистики — по 24 %, церковно-богословские тексты составляют 19 %, научные тексты ф — 17 %, официальные документы — 11 %, бытовые тексты (письма, дневники) — 5 %. Приблизительно в тех же пропорциях представлены и тексты XIX в. Хронологически тексты экспериментального корпуса распределяются следующим образом: 1700–1730 — 6 %, 1731-1780 — 43 %, 1781–1799 — 30 %, 1800–1830 — 21 %. Подробнее о составе корпуса XVIII в. см. [Савчук, Сичинава 2009]).

место занимают: существительные, прилагательные в форме тв. п. на *-ою* (263 формы): *Гришкою*, *Кабардою*, *прежестокою*; причастия в форме им. п. мн. ч. на *-ии* (101) */ -ыи* (18): *входящии*, *нарицающии*, *приежжаемыи*, *украшенныи*; причастия на *-яй* (128): *возвышаяй*, *вступаяй*; краткие причастия на *-ущ / ащ* (27): *блистающ*, *властвующ*; формы имперфекта: *живяше*, *знаяше* (10); *бяху, стояху, мняху* (22). Часть глагольных форм была лемматизирована, но все предложенные гипотезы оказались ошибочными: форма 2 л. наст. в. на *-ши* (278): *деeши, жаждеши*; имперфекта (122): *поучаше, презираше, моляшеся; бежаху, зваху, побиваху, нарицахуся, удивляхуся, являхуся*; аориста: *искусиша, победиша* (39)*; несохом, победихом* (50) и др. К архаичным глагольным формам примыкают также диалектно-просторечные формы 3 л. ед. ч. на *-ут/-ют* для глаголов 2 спряжения (40): *купют, просют, проводют, посмотрют, готовются, находются*. Все они займут свое место в парадигмах исторического модуля словаря.

Третью многочисленную группу составляют орфографические варианты: *безщчетну, возмеш, баталиах, полицыи, прокломации, поосчрении* (написания с *сч* вместо *щ* совершенно регулярны, например, у Татищева, который в своем орфографическом трактате утверждал об избыточности буквы *щ* и свою фамилию писал как *Татисчев*) и др. Здесь анализ отдельных групп вариантов приведет к формулированию частных формальных правил анализа соответствующих орфограмм (см. п. 4.3).

Четвертую группу составляют многочисленные собственные имена — топонимы, имена, фамилии и отчества лиц, литературных героев и мифологических персонажей, причем многие из них присутствуют в текстах в нескольких вариантах: *Шлиссельбург* (утвердившийся впоследствии вариант)*, Шлюссембурх, Шлютенбурх, Шлютелбург, Шлютельбург, Слюсинбург, Слютелбург, Слютельбург, Валпарейсо, Валпарейзо, Вальпарейзо* (в современной передаче — *Вальпараисо*, город в Чили)*, Елизавета, Елисавета, Елисавет, Елисавефь, Елисафет, Ньютон, Нейтон, Невтон, Дон-Кихот, Дон-Кишот, Донкишот, Микель-Анджело, Мишель Анжело* и др. Примыкают к этой группе производные от собственных имен — прилагательные и существительные: *европскии, коперниканскии, ефесския, Антошка, Бомонтша* и т. д.

Часть неопознанных форм (около 1 %) объясняется ошибками набора и сканирования, в результате проверки таких «псевдоформ» по текстам вносятся исправления.

5.2. Оценка гипотетических разборов.

Среди словоформ, получивших гипотетические разборы, можно выделить зоны с высоким уровнем предсказуемости (25–30 %, то есть одна гипотетическая лемма из трех или четырех предложенных оказывается правильной), зоны средней предсказуемости (предлагается от пяти гипотез, из них одна правильная) и зоны с нулевой предсказуемостью (ни одна из предложенных гипотез не является правильной).

Примеры высокой предсказуемости обнаружили, в частности, существительные с основой на *-к-* (около 4% словоформ, получивших варианты разборов), спрягаемые формы глаголов, за исключением архаичных (более 6% словоформ), формы прилагательных (около 35% всех словоформ с гипотетическими разборами):

доимка доимка?=N,f,inan=sg,nom=N33* | доимок?=N,m,inan=sg,gen=N13* | доимка?=N,f,anim=sg,nom=N33*

напояют напоять?=V,ipf=ind,pres,pl,3,act=V1 | напоять?=V,ipf,intr=ind,pres,pl,3,act=V1 | напоять?=V,pf=ind,fut,pl,3,act=V1

комфузные комфузный?=A=pl,nom/acc=A1* | комфузной?=A=pl,nom/acc=A1b | комфузная?=N,f,inan=pl,nom/acc/acc=A1 | комфузный?=A=pl,nom/acc=A1 | комфузные?=N,pl,inan=pl,nom/acc/acc=A1

Из архаических форм в зоне высокой предсказуемости находятся формы прилагательных мн. ч им. п. на *-ия /-ыя, -аго*:

одинакия одинакий?=A=pl,nom/acc=A3 | одинакий?=A=pl,nom/acc=A3* | одинакая?=N,f,inan=pl,nom/acc/acc=A3 | одинакой?=A=pl,nom/acc=A3b | одинакое?=N,n,inan=pl,nom/acc/acc=A3

Эти формы правильно опознаются как прилагательные и отграничиваются от форм существительных с омонимичными финалями *-ия:*

министерия министерия?=N,f,inan=sg,nom=N37 | министерия?=N,topn,f,inan=sg,nom=N37 | министерий?=N,persn,m,anim=sg,gen/acc=N17 | министерий?=N,m,inan=sg,gen=N17 | министерия?=N,persn,f,anim=sg,nom=N37

Пример средней степени предсказуемости:

неведь неведь?=N,f,inan=sg,nom/acc=N41 | неведь?=N,m,anim=sg,nom=N12 | неведь?=N,topn,m,anim/inan=sg,nom=N12 | **неведь**?=CONJ/PART | неведь?=PREP

Наибольшие сложности представляет идентификация архаических глагольных форм. Все они попадают в зону нулевой предсказуемости. Причина объяснена выше — отсутствие в словнике в достаточном объеме церковнославянских глаголов и парадигм спряжения.

Форма **вознесоста** (2 л. двойст. числа аориста глагола *вознести*) получает гипотетические разборы, ни один из которых не является верным:

```
вознесост?=N,m,anim=sg,gen/acc=N11 | вознесост?=N,m,inan=sg,gen=N11
    | вознесоста?=N,f,inan=sg,nom=N31 | вознесост?=N,persn,m,anim=
    sg,gen/acc=N11 | вознесост?=N,topn,m,inan=sg,gen=N11
```

Форме **видиши** (2 л. ед. ч. наст. вр. глагола *видети*) приписаны ошибочные разборы[5]:

```
видиша?=N,f,inan=pl,nom/acc¦sg,gen=N34 |
    видисать?=V,pf=imp,sg,2,act=V6t | видисать?=V,ipf=imp,sg,2,act=V
    6t | видиша?=N,persn,m,anim=pl,nom¦sg,gen=N34
```

Нулевую предсказуемость имеют формы им. п. мн. ч. прилагательных на *-ии* (современное *-ие*).

```
великии — великия?=N,topn,f,inan=pl,nom/
    acc¦sg,dat¦sg,gen¦sg,loc=N37

глухии — глухия?=N,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37 |
    глухия?=N,topn,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37
```

В отличие от прилагательных существительные на *-ия* опознаются с высокой степенью правильности:

```
девизии — девизия?=N,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37 |
    девизия?=N,topn,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37 |
    девизие?=N,n,inan=sg,loc=N27

енергии — енергия?=N,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37 |
    енергия?=N,topn,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37
```

Если учесть, что прилагательных в этой зоне в 2 раза больше, чем существительных, в разбор словоформ следует включить граммему прилагательного, применив специальные фильтры. В частности, для форм, оканчивающихся на *-нии* (их около 700), справедливо следующее наблюдение:

---

[5] Архаические глагольные формы — проблема и для анализатора Mystem, который в отдельных случаях «взрывается» ложными гипотезами. В частности, для формы доставляеши в НКРЯ предлагается 8 лемм и 37 гипотетических вариантов анализа.

Лемма доставляеши — сущ, фам, одуш, м, 0 / сущ, неод, с, ед, 0 /сущ, имя, одуш, м, 0
Лемма доставляеший — прил, мн, кратк
Лемма доставляешить — глаг, нп, нсв, повел, действ, 2л, ед / глаг, нп, св, повел, действ, 2л, ед / глаг, перех, нсв, повел, действ, 2л, ед / глаг, перех, св, повел, действ, 2л, ед
Лемма доставляеш — сущ, фам, одуш, м, мн, им / сущ, неод, м, мн, вин, геогр / сущ, неод, м, мн, им, геогр / сущ, имя, одуш, м, мн, им /
Лемма доставляеша — сущ, фам, одуш, м, мн, им / сущ, фам, одуш, м, ед, род / сущ, фам, одуш, м–ж, мн, им / сущ, фам, одуш, м–ж, ед, род
и т. д.

- часть основы на гласную + *-нии* => существительное (*Гавании, Казании, Исмении, докончании, догорении);*
- часть основы на согласную + *-нии* => прилагательное (*главнии, вчерашнии, доволнии*);
- часть основы на н + *-нии* => прилагательное или причастие (*украшеннии, смиренниии*).

В зоне нулевой предсказуемости находится также подавляющее большинство (более 300) форм тв. п. существительных и прилагательных на *-ою* (*ескадрою, Козмою, кавалерственною, манетною*), притяжательные прилагательные на *-ин* (*богинин, венерин, минервин* и др.).

Список словоформ с гипотетическими разборами находится в стадии обработки: словоформам с ошибочными разборами вручную приписываются леммы и грамматические признаки, для форм со средней предсказуемостью рассматриваются способы сокращения предлагаемых гипотез.

5.3. Сокращение количества гипотез

Сокращения предлагаемых гипотез можно добиться путем использования правил, которые будут применяться к определенным классам словоформ, образующим закрытые или пополняемые списки. Часть правил будет основана на учете связи морфологических признаков с морфемной структурой слова. Так, например, для форм на *-ения* анализатор предлагает 5 гипотетических лемм:

```
напоения напоение?=N,n,inan=pl,nom/acc¦sg,gen=N27 |
    напоения?=N,f,inan=sg,nom=N37 | напоения?=N,topn,f,in
    an=sg,nom=N37 | напоения?=N,persn,f,anim=sg,nom=N37 |
    напоений?=N,m,inan=sg,gen=N17

посмотрения посмотрение?=N,n,inan=pl,nom/acc¦sg,gen=N27 | посмот
    рения?=N,f,inan=sg,nom=N37 | посмотрения?=N,topn,f,inan=sg,no
    m=N37 | посмотрения?=N,persn,f,anim=sg,nom=N37 | посмотрений?
    =N,m,inan=sg,gen=N17
```

При этом статистически формы распределяются следующим образом. Из 444 форм 439 (98%) являются формами род. п. существительных среднего рода, 2 формы — род. п. мужского рода и 3 — им. п. женского рода (имена собственные). Следовательно, формам на *-ения* целесообразно приписывать леммы существительных среднего рода как наиболее статистически значимые, а существительные женского и мужского рода задать списком и включить в словарь.

Аналогичная картина наблюдается у существительных с суффиксом *-ствиј-* (*-ствие, -ствия, -ствии, -ствию* и т. д.): 206 из 207 словоформ относятся к существительным среднего рода, 1 форма — предлог (*вследствии*). Следовательно, набор лемм, который предлагается при анализе словоформ этого достаточно продуктивного класса, можно сократить до одной.

```
неблагодарствии неблагодарствие?=N,n,inan=sg,loc=N27 |
    неблагодарствия?=N,topn,f,inan=pl,nom/
    acc¦sg,dat¦sg,gen¦sg,loc=N37 |
    неблагодарствия?=N,persn,f,anim=pl,nom¦sg,dat¦sg,gen¦sg,loc=N37 |
    неблагодарствий?=N,persn,m,anim=pl,nom¦sg,loc=N17 |
    неблагодарствия?=N,f,inan=pl,nom/acc¦sg,dat¦sg,gen¦sg,loc=N37
```

Среди небольших групп, в которых целесообразно уменьшить количество вероятных гипотез, можно назвать характерные для языка XVIII века германизмы со специфическими исходами: формы на *-берг* (топонимы и фамилии), *-бург* (топонимы), на *-ау* (топонимы славянского происхождения, например, *Бункау, Лаубау*); на *-мейстер, -мистр* (сущ. муж. рода, одушевленные *геролдмейстер, ширмейстер, вафмистр, виц-вахмистр, квартермистр, секунд-ротмистр*) и др.

Другой источник сокращения количества гипотез, как уже говорилось в п. 4.3, видится в анализе орфографической вариативности, реально представленной в текстах, и применении орфографических преобразований. Приведем несколько примеров.

Формы с приставками на *-з*: формы с начальным *без-/в(о)з-/из-/низ-/раз-/роз-/ч(е)рез-/з-* перед глухими анализировать как формы с начальным *бес-/в(о)с-/ис-/нис-/рас-/рос-/ч(е)рес-/с-* (для приставки *без-* такое обобщение уже сделано).

Слитное написание *не* с глаголами: *небудет, неисповедует, ненайдет, нетребует*;

сч => щ: *поосчрение=поощрение, немосчный=немощный* и др., более 500 форм;

жы, шы => жи, ши: *показавшые, вашы, стражы, лжы* (55);

конечное *-ыя => ия*, *-ыи => ии*: *Францыя, полицыя, позицыя* (10), *амуницыи, полицыи* (18), конечное *-иа => -ия*: *библиа, благополучиа* (145), *-иах => иях*: *баталиах, материах* (23);

-лск(ий, ого, аго, ому и т.д) *=> льск*(ий, ого, аго, ому и т.д): *посолский, тоболской, неприятелское, евангелский* (87), *-лст- => -льст-*: *лстивый, жителство, посолство, ловителствуя, началству, обстоятелство* (более 130). Список орфографических особенностей текстов в старой орфографии см. также в [Поляков 2012].

6. Заключение

Принципы формирования грамматического словаря для автоматического анализа текстов XVIII–XIX вв., описанные в настоящей статье, были опробованы на экспериментальном корпусе, составленном из текстов XVIII — 1-ой трети XIX в. Результаты автоматической морфологической разметки показали, что более 73 % словоформ корпуса получили однозначные разборы, что свидетельствует о степени совпадения текстов данного периода и современных текстов по словарному составу. Однако следует учесть, что относительно высокий процент совпадения в значительной мере объясняется составом корпуса: в нем мало текстов начала XVIII в., кроме того, большая часть текстов, в соответствии

с ориентацией на авторитетные научные издания советского и постсоветского времени, представлена в современной орфографии или с некоторыми отклонениями от современной нормы. Очевидно, что увеличение доли оригинальных текстов XVII и первой трети XVIII в. изменило бы результат анализа в сторону уменьшения доли совпадающих словоформ и привело бы к росту количества ошибочных разборов.

Изучение и классификация ошибок анализатора позволило наметить пути дальнейшего расширения и совершенствования словаря. Это во-первых, ручной анализ неопознанных форм и включение соответствующих лемм в словник; во-вторых, пополнение списка парадигм за счет церковнославянских и старорусских форм, а также включение вариативных грамматических форм в состав отдельных современных парадигм; в третьих, составление орфографических правил для алгоритмической нормализации. В ближайшие планы входит внедрение всех изменений в словарь, тестирование его новой версии на экспериментальном корпусе, откорректированном с учетом выявленных ошибок, а также на отдельных текстах более раннего периода, относящихся к разным жанрам.

Литература

1. *Большаков И. А., Большакова Е. И.* Автоматический морфоклассификатор русских именных групп // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2012). Вып. 11. М., 2012. С. 81–92.
2. *Гришина Е. А., Савчук С. О.* Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 129–149.
3. *Грот Я. К.* Спорные вопросы русского правописания от Петра Великого доныне. СПб, Типография императорской АН, 1873.
4. *Добровольский Д. О., Кретов А. А., Шаров С. А.* Корпус параллельных текстов: архитектура и возможности использования. // Национальный корпус русского языка: 2003–2005. — М.: Индрик, 2005. С. 263–296.
5. *Зализняк* 1977/2003 — Зализняк А. А. Грамматический словарь русского языка. Изд. 1-е. М., 1977 (4-е изд., испр. и доп., М. 2003)
6. *Качинская И. Б.* Диалектный подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место // О. Ю. Крючкова и др. (ред.) Русская устная речь. Материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения» и межвузовского совещания «Проблемы создания и использования диалектных корпусов». Саратов, Издательский центр «Наука», 2011. С. 245–255
7. *Поляков А. Е.* Проблемы и методы анализа русских текстов в дореформенной орфографии // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2012). Вып. 11. М., 2012. С. 536–547.

8. *Поляков А. Е., Добрушина Е. Р., Иванова-Алленова Т. Ю.* Корпус церковнославянских текстов в составе НКРЯ, первая версия: проблемы и решения. Информационные технологии и письменное наследие. Материалы международной научной конференции. — Петрозаводск, 2012. С. 211–215.
9. *Савчук С. О., Сичинава Д. В., Гарипов И.* Подкорпус текстов XVIII века в составе Национального корпуса русского языка: из опыта работы // Web Journal of Formal, Computational & Cognitive Linguistics. Специальный выпуск (Труды Российского научно-образовательного центра по лингвистике им И. А. Бодуэна де Куртенэ), 2006.
10. *Савчук С. О.* Корпус русских текстов XVIII века в составе Национального корпуса русского языка: проблемы и перспективы // Информационные технологии и письменное наследие. Материалы международной научной конференции. Казань, 2008. С. 241–244.
11. *Савчук С. О., Сичинава Д. В.* Корпус русских текстов XVIII века в составе НКРЯ: проблемы и перспективы // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 52–70.
12. *Савчук С. О.* Электронный словарь вариантов на основе текстов XVIII в. Информационные технологии и письменное наследие. Материалы международной научной конференции. — Петрозаводск, 2012. С. 241–244.
13. *Соловьев В. Д., Ахтямов Р. Б.* Корпус русского языка XVIII века: текущее состояние/ Материалы международной научной конференции Ижевск, 13–17 июля 2006 г. Ижевск, 2006. С. 156–160.
14. *Baron, A., Raison, P.* (2009) Automatic standardization of texts containing spelling variations. How much training data do you need? // In M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23, http://ucrel.lancs.ac.uk/publications/CL2009/314_FullPaper.pdf
15. *Lay, M. H.* (2012) VariaLog: how to locate words in a French Renaissance Virtual Library // Digital Humanities Conference, University of Hamburg, Germany, 2012, available at: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/varialog-how-to-locate-words-in-a-french-renaissance-virtual-library/
16. *Meyer, R.* (2009) Semi-automatic morphosyntactic tagging of a diachronic corpus of Russian // In M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23. http://ucrel.lancs.ac.uk/publications/cl2009/abstracts.htm#347
17. *Sitchinava D.* (2012) Parallel corpora within the Russian National Copus // *Prace Filologiczne*, LXIII, 2012. С. 271–278.

References

1. *Baron, A., Raison, P.* (2009), Automatic standardization of texts containing spelling variations. How much training data do you need? M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23, available at: http://ucrel.lancs.ac.uk/publications/CL2009/ 314_FullPaper.pdf
2. *Bolshakov, I. A., Bolshakova, E. I.* (2012), An Automatic morphological classifier of noun phrases in Russian [Avtomaticheskij morfoklassifikator russkih imennyh grupp]. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" 2012 [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj mezhdunarodnoj konferencii "Dialog" 2012]. Bekasovo, pp. 81–92.
3. *Dobrovol'skij D. O., Kretov A. A., Sharov S. A.* (2005), Parallel Corpus: architecture and usability [Korpus parallel'nyh tekstov: arhitektura i vozmozhnosti ispol'zovanija], in Russian National Corpus: 2003–2005 [Nacional'nyj korpus russkogo jazyka: 2003–2005], Indrik, Moscow, pp. 263–296.
4. *Grishina E. A., Savchuk S. O.* (2009), Spoken texts in the RNC: composition and structure [Korpus ustnyh tekstov v NKRJa: sostav i struktura], in Russian National Corpus: 2006–2008. New Results and Perspectives [Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy]. Nestor-Istorija, SPb, pp. 129–149.
5. *Grot Ja. K.* (1873) Controversial issues of Russian spelling since Peter the Great until now [Spornye voprosy russkogo pravopisanija ot Petra Velikogo donyne]. Tipografija imperatorskoj AN, SPb.
6. *Kachinskaja I. B.* (2011), Dialectal subcorpus of the RNC. The new standards. New workplace [Dialektnyj podkorpus NKRJa. Novyj standart podachi. Novoe rabochee mesto], in O. Ju. Krjuchkova i dr. (red.) Russian speech. Proceedings of the International Conference "Barannikovskie reading. Spoken speech: Russian dialect and colloquial vernacular culture of communication" and intercollegiate conference "Development and Use of dialect corpora" [Russkaja ustnaja rech'. Materialy mezhdunarodnoj nauchnoj konferencii "Barannikovskie chtenija. Ustnaja rech': russkaja dialektnaja i razgovorno-prostorechnaja kul'tura obshhenija" i mezhvuzovskogo soveshhanija "Problemy sozdanija i ispol'zovanija dialektnyh korpusov"], Izdatel'skij centr "Nauka", Saratov, pp. 245–255.
7. *Lay, M. H.* (2012), VariaLog: how to locate words in a French Renaissance Virtual Library, Digital Humanities Conference, University of Hamburg, Germany, 2012, available at: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/varialog-how-to-locate-words-in-a-french-renaissance-virtual-library/
8. *Meyer, R.* (2009), Semi-automatic morphosyntactic tagging of a diachronic corpus of Russian, M. Mahlberg, V. González-Díaz, and C. Smith (eds.), Proceedings of the Corpus Linguistics Conference, CL2009. University of Liverpool, UK, pp. 20–23, available at: http://ucrel.lancs.ac.uk/publications/cl2009/abstracts.htm#347

9. *Poljakov A. E.* (2012), Problems and methods in analysis of Russian texts in the pre-reform spelling [Problemy i metody analiza russkih tekstov v doreformennoj orfografii], Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog" 2012 [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Po materialam ezhegodnoj mezhdunarodnoj konferencii "Dialog" 2012]. Bekasovo, pp. 536–547.
10. *Poljakov A. E., Dobrushina E. R., Ivanova-Allenova T. Ju.* (2012), Corpus of Church Slavonic texts in the RNC, the first version: problems and solutions. [Korpus cerkovnoslavjanskih tekstov v sostave nkrja, pervaja versija: problemy i reshenija], Information technology and the written heritage. Proceedings of the International Conference [Informacionnye tehnologii i pis'mennoe nasledie. Materialy mezhdunarodnoj nauchnoj konferencii], Petrozavodsk, pp. 211–215.
11. *Savchuk S. O., Sichinava D. V., Garipov I.* (2006), Subcorpus of the XVIIIth century texts in the Russian National Corpus [Podkorpus tekstov XVIII veka v sostave Nacional'nogo korpusa russkogo jazyka: iz opyta raboty], Web Journal of Formal, Computational & Cognitive Linguistics. Special Issue (Proceedings of the Baudouin de Courtenay Russian Research and Educational Center in linguistics) [Special'nyj vypusk (Trudy Rossijskogo nauchno-obrazovatel'nogo centra po lingvistike im. I. A. Boduena de Kurtene)], available at: http://fccl.ksu.ru/fcclpap.htm.
12. *Savchuk S. O.* (2008), Corpus of the Russian XVIIIth century texts in the Russian National Corpus: problems and prospects [Korpus russkih tekstov XVIII veka v sostave Nacional'nogo korpusa russkogo jazyka: problemy i perspektivy]. Information technologies and written heritage. Proceedings of the International Conference [Informacionnye tehnologii i pis'mennoe nasledie. Materialy mezhdunarodnoj nauchnoj konferencii], Kazan, p. 241–244.
13. *Savchuk S. O., Sichinava D. V.* (2009), Corpus of the Russian XVIIIth century texts in the RNC: Problems and Perspectives [Korpus russkih tekstov XVIII veka v sostave NKRJa: problemy i perspektivy], in Russian National Corpus: 2006–2008. New Results and Perspectives [Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy], Nestor-Istorija, SPb, pp. 52–70.
14. *Savchuk S. O.* (2012), Electronic dictionary of variants based on the 18th century texts [Elektronnyj slovar' variantov na osnove tekstov XVIII v.], Information technology and the written heritage. Proceedings of the International Conference [Informacionnye tehnologii i pis'mennoe nasledie. Materialy mezhdunarodnoj nauchnoj konferencii], Petrozavodsk, pp. 241–244.
15. *Sitchinava D.* (2012), Parallel corpora within the Russian National Corpus. *Prace Filologiczne*, LXIII, 2012, pp. 271–278
16. *Solovyev V. D., Akhtyamov R. B.* (2006), Corpus of the XVIIIth century Russian: the present state of affairs. [Korpus russkogo jazyka XVIII veka: tekushhee sostojanie]. Proceedings of the International Conference [Materialy mezhdunarodnoj nauchnoj konferencii]. Izhevsk, pp. 156–160.
17. *Zaliznjak, A. A.* (1977/2003), Grammatical dictionary of the Russian language [Grammaticheskij slovar' russkogo jazyka], Moscow (4 ed. 2003).

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ПРАВИЛ ДЛЯ СНЯТИЯ МОРФОЛОГИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ

Протопопова Е. В. (protoev@gmail.com),
Бочаров В. В. (victor.bocharov@gmail.com)

Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия

Ключевые слова: омонимия, морфологическая разметка, русский язык, неконтролируемое обучение

UNSUPERVISED LEARNING OF PART-OF-SPEECH DISAMBIGUATION RULES

Protopopova E. V. (protoev@gmail.com),
Bocharov V. V. (victor.bocharov@gmail.com)

Saint Petersburg State University, Saint Petersburg, Russia

Morphological disambiguation is one of the key aims of part-of-speech tagging. The task is considered to be solved, though all the tools for disambiguation use a lot of manually created data. This paper describes an attempt to disambiguate Russian corpus without manually annotated data. The method used was proposed about twenty years ago but has not been applied to synthetic languages yet. The main idea of our approach is to derive disambiguation rules automatically from a corpus with ambiguous annotations using only a few statistical data. It can be done in a simple way by means of unsupervised learning. The results are quite high and can be compared to results of existing systems. We also tried to measure the size of the corpus necessary to produce a reasonable set of disambiguation rules and showed that it can be comparable in size with the corpora used to train statistical disambiguation models.

Keywords: ambiguity, Russian language, morphological annotation, unsupervised learning

1. Introduction

Although we have observed a great improvement in POS-tagging for English and other European languages and works on this topic became quite rare during past ten years, the number of works devoted to POS-tagging of Russian language is rather low. The existing systems [Zelenkov et al. 2005, Sokirko, Toldova 2005, Sharoff, Nivre 2011] use machine learning methods which require a lot of manually annotated data. In this work we have tried to apply an unsupervised algorithm described in [Brill 1995]. This method has several advantages:

- It is based on automatically annotated corpus and the manually created annotation is necessary only for evaluation.
- The output of the system is a list of rules which can be understood and explained.

In this paper we are going to examine, what accuracy we can achieve using this unsupervised model, what corpus do we need to train the model effectively and how the size of corpus affects the resulting rules.

Our approach is based on that described in [Brill 1995]. We should mention that we did not take into account morphological categories other than part-of-speech, though case ambiguity, for example, is very common in Russian. The core idea of Brill's approach is to gather statistical information about POS tags and their distribution. The rule to transform tags A_B into tag B can be obtained if the system has seen that tag B is more frequent than tag A in the observed context.

We conducted a number of experiments to assess the size of training corpus. The algorithm was applied to corpora of various size — from 1K to 170K sentences and we obtained different lists of rules thereby. We present and compare the results in Section 4.

2. Related work

All known approaches to disambiguating Russian corpora are trained on manually tagged part of Russian National Corpus. The first such algorithm presented in [Zelenkov et al. 2004] is based on statistical knowledge about tags and their contexts and its accuracy is more than 97%. The probability of each possible tag is computed as sum of probabilities of this tag in the context multiplied by their scores ("influence" on its environment).

A tagger based on Hidden Markov Model is described in [Sokirko, Toldova 2005] and achieves up to 98% accuracy. The algorithm takes into account trigram probability for tags (tag0 was seen after tag1 and tag2) and bigram probability for words (word1 was tagged as tag1 after tag0). Manually annotated part of RNC (5M words) was used as training data.

Later research in [Sharoff, Nivre 2009] showed that HMM with some improvements (guessing inknown words by their ending) can be successfully applied for POS tagging and achieves a competitive result of about 97% on a reduced tagset. The authors mention common problems of statistical POS tagging stating however that "a completely automatic machine learning procedure can quickly produce a fast and reliable NLP component".

3. Methods and data used

3.1. The original algorithm

The idea was in short described above and here we want to mention some details. We did not change the original algorithm but there are a few details in implementation.

First, a text is annotated by an initial-state annotator, which can assign a word either a random structure or an output of a manually-created dictionary. Then a learner is given transformation templates such as following:

Change tag from x to Y in context C.

where x is set of tags (i.e. morphological hypothesis) assigned to a word, $Y \in x$. C is one context feature: one word or tags to the left or to the right are considered as context features.

Unlike supervised learner, the unsupervised one cannot measure the accuracy of the transformation, hence a special scoring function is used to find more reliable disambiguation contexts. In each learning iteration the score is based on the current tagging of a corpus.

Computing the score for the transformation above includes three steps:

1) For each tag $Z \in x$, $Z \neq Y$ compute

$$\frac{freq\,(Y)}{freq\,(Z)} \cdot incontext(Z, C)$$

2) Let

$$R = argmax_z \frac{freq\,(Y)}{freq\,(Z)} \cdot incontext(Z, C)$$

3) Then score for this transformation is

$$score = incontext(Y, C) - \frac{freq\,(Y)}{freq\,(R)} \cdot incontext(R, C)$$

where *freq(X)* is number of occurences of words unambiguously tagged with tag X and *incontext(X,C)* is number of occurences of words unambiguously tagged with tag X in context C.

On each interation the algorithm searches the transformation which maximizes the scoring function and the learning stops when no positive scoring transformations can be found.

In [Brill 1995] the algorithm was applied to tagging English corpus and resulted in 95.1% accuracy on 200K words part of Penn Treebank and 96.0% accuracy on 350K words part of Brown corpus.

In our implementation we took into account only nearest right and left context including punctuation and sentence borders. If several rules have the same score and it is the best score on this iteration, they are applied in descending order of the chosen tag frequency.

The rules were obtained from 10 random corpora of each size and are written in the following way:

ADJF NOUN → NOUN | 1:tag=PNCT

that is

Change tag from ADJF NOUN to NOUN if next tag is PNCT.

We also store the following information for each rule: its score, number of rule applications on this iteration, number of occurences of an ambiguous tag.

3.2. Differences between English and Russian tagset

The tagset for inflective languages such as Russian is bigger and different in its structure from one for English because it includes not only part of speech tags but also grammatical categories such as case, number, gender, tense etc. For one word form a morphological hypothesis is a set of tags that can be considered as a set of key-value pairs where key is a grammatical category: part-of-speech — noun, number — single, case — nominative etc. Each morphological interpretation is a set of morphological hypothesis (i.e. set of tag sets). The following example includes three hypothesis (one hypothesis per line) for form “чай” that can be both imperative mood of verb “чаять” (to hope, to expect) and nominative or accusative case of noun “чай” (tea):

чай VERB impf sing excl tran impr

чай NOUN masc sing inan nomn

чай NOUN masc sing inan accs

According to the morphological dictionary we used there are 4369 different tag sets (lines in the example above) and 1678 of them are assigned to more than 10 forms. It makes a little sense to use such a big tagset in machine learning because most of items are very rare. The obvious solution is to split disambiguation task into several steps: one step per grammatical category (as in [Acedanski, Gołuchowski, 2009], the approach described below). In this paper we describe the first step where only part of speech tags are considered. The example with form “чай” now looks much simpler and the step-wide tagset is reduced to only part-of-speech tags:

чай VERB
чай NOUN

Our annotator assigns unknown words tags UNKN (unknown sequence of cyrillic characters), LATN (unknown sequence of latin characters), NUMR (numeric characters) and PNCT (punctuation). The algorithm also uses tags SBEG and SEND as sentence borders context.

A similar but supervised approach was applied to Polish POS-tagging [Acedanski Gołuchowski, 2009]. Tagging was performed in two phases, first of all POS, case and person were disambiguated and then other categories.

3.3. Training and test corpora

We used articles from http://www.chaskor.ru/ as training data. 15M tokens were annotated using OpenCorpora morphological dictionary and ten disjoint random corpora of each size (from 1K to 170K sentences) were derived from it. The annotation is represented in a simple way:

48,501 Согласно	328,254 согласно	328,255 согласно	328,258 согласен
	ADVB	PREP	ADJS Qual neut sing

Each hypothesis include lemma's id, lemma and morphological annotation itself.

For the test set we have taken a random sample of manually disambiguated sentences from OpenCorpora project.

4. Results

Our first aim was to find out what corpus size is required to produce a reasonable set of rules. Since there is no straightforward way to measure it, we made a number of experiments. 325 lists of rules obtained from different training sets were compared and the results are described below.

4.1. Disambiguation rules

First of all, the lists of rules were compared according to their size and content. The results (fig. 1) show that the number of rules increases if we increase the size of training corpus because the number of contexts increases respectively. The number of unique rules increases less but it did not become absolutely stable on big corpora (more than 3M words).

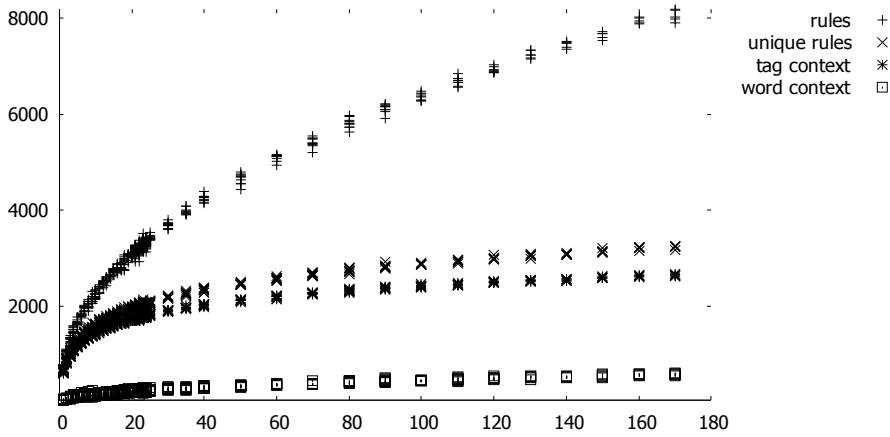


Fig. 1. Number of disambiguation rules obtained from different corpora

The similarity between lists of rules was also measured as Spearman rank correlation coefficient, which is higher for bigger training corpora (fig.2). The increasing correlation coefficient shows that same rules are ranked in almost the same order in lists obtained on big corpora.

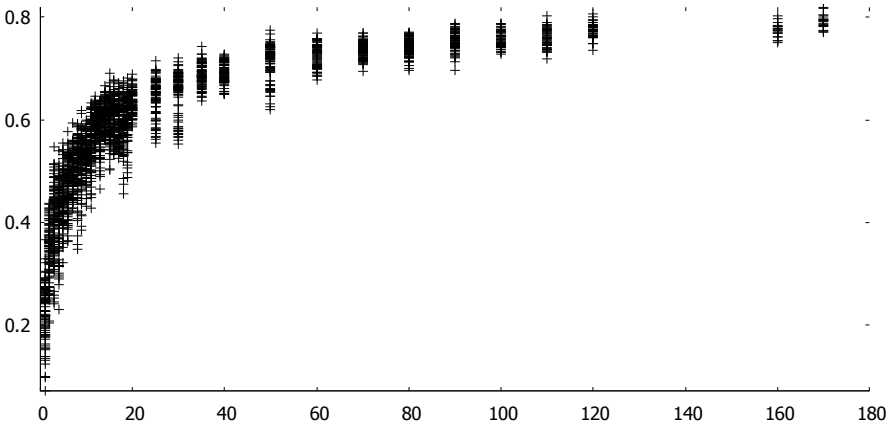


Fig. 2. Spearman rank correlation between each two sets of rules

4.2. Unsupervised annotation results

Another way to check if we got enough rules for tagging is to examine the difference in annotation after applying different rule lists. A test set containing 1K sentences was disambiguated using various sets of rules. First of all, we compared the resulting annotations with each other. We see that the difference decreases if we increase size of the training corpus (fig. 3).

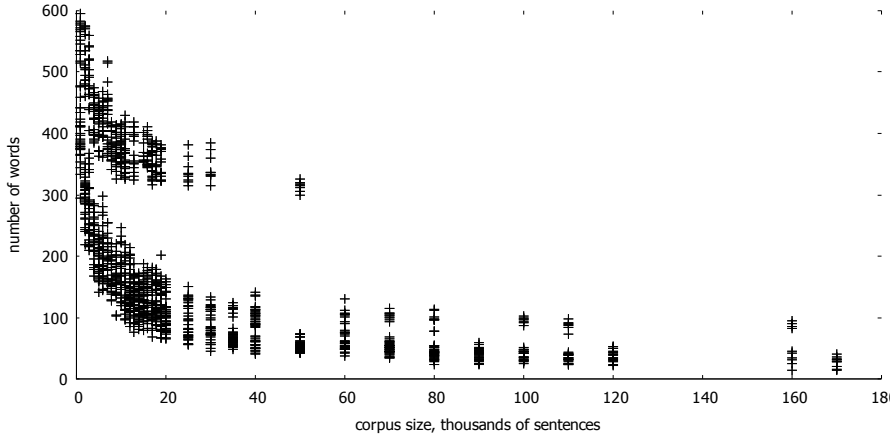


Fig. 3. Number of words tagged differently depending on the size of the training corpus

The increase in training corpus size causes the increasing number of context features hence the number of words with changed annotation grows (from 13 to 15%) and the number of ambiguous annotations left decreases (fig. 4). We should mention that 40% of words in the corpus were ambiguous.s

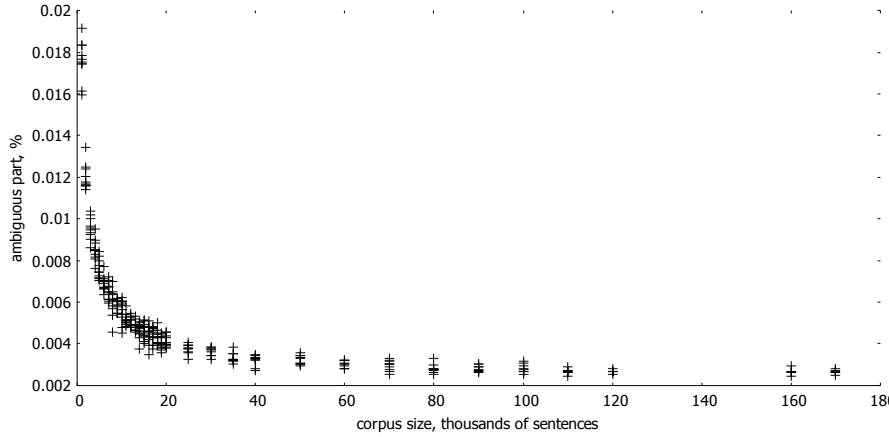


Fig. 4. Number of ambiguous annotations left

However, the growing number of context features does not always correspond to the increasing accuracy in disambiguation as it is shown below. We also compute the recall of our algorithm as fraction of unambiguous annotations in the test corpus. This metric shows the same results as those described above.

Most of ambiguous tags left are tags for related parts-of-speech (CONJ, INTJ, PRCL) which can appear in any context and tags which include these function words classes and are used for only several words such as *uzhe 'already'* — ADVB/COMP/NOUN/PRCL.

4.3. Accuracy

Tagging accuracy was measured on manually disambiguated 100 sentences selected from OpenCorpora project corpus. The test corpus includes sentences of different genres and syntactic structure. They were disambiguated with sets of rules described above and this annotation was compared with manually created one. Number of tagging mistakes decreases according to size of training corpora.

For each mistake an initial ambiguous tag is regarded as a mistake types. We took three most frequent mistakes for each list of rules and rank them according to the average number of their occurences. These major types of mistakes are shown below. Some of them are just mistakes in tagging one word (as *как* — ADVB/CONJ/NPRO), others are mistakes in tagging words that can appear almost in any context (CONJ/PRCL: *и*). Such cases as ADJF/NOUN are due to the limited set of context features: the rule *"ADJF NOUN -> ADJF | 1:tag=NOUN"* does not cover the situation when two adjectives are followed by a noun.

Table 1. Most frequent tagging mistakes

ADJF/NPRO	7.49417852523
CONJ/PRCL	6.26098191214
ADVB/CONJ	6.24516129032
PRCL/CONJ	6.17464424321

One of the major mistakes — tagging conjunction as adverb — is caused by tagging parenthesis as a conjunction, so that this type of ambiguity is heterogenous: out tagger mixes the adverbs used as parentheses with the words which really can be tagged either as conjunction or as an adverb (*kogda 'when', kak 'how', tak 'so'*).

Tagging accuracy is computed as a fraction of correct tags in test corpus (fig. 5). We suppose that the dispersion of results is due to the genre peculiarities of the training corpora and to the size of test corpus itself. However, we can see that stable reasonable results were obtained on the corpora bigger than 50K sentences. The highest average precision is observed on training corpora of 19–20 thousands of sentences. It should be noticed that tagger which chooses a random tag for each ambiguous word achieves accuracy about 93% in our case.

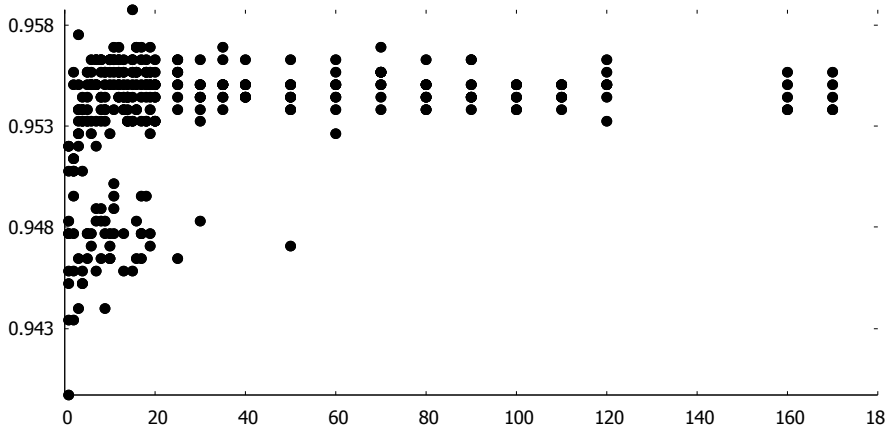


Fig. 5. Tagging accuracy

5. Conclusion

Taking into account the simplicity of the algorithm, we can say it has achieved quite reasonable accuracy and the performance of the system can be improved. We have shown that the morphological disambiguation task for Russian language can be solved almost without any special linguistic work using only corpora and morphological dictionary and our results are practically the same as those obtained in Brill's work. We have not come to a definite conclusion about the sufficient size of the training corpus, though several evaluations show that systems trained on corpora of 60K sentences and bigger achieve quite high results and produce practically the same number of rules.

There are several ways to improve our system. First of all, in this work we have used few context features. The context can be extended to four words (two left and two right neighbours) and the learner can take into account some more features including some lexical and grammatical categories. The algorithm should also be tuned to solve the ambiguity of word-forms (such as case ambiguity) as its supervised version was adapted for Polish morphological disambiguation. Another way for improvements is to study the influence of training corpus genre on the resulting set of rules. These improvements may require more linguistic knowledge which nevertheless cannot be compared to the task of creating fully manually annotated corpus.

References

1. *Acedański S. and Gołuchowski K.* A Morphosyntactic Rule-Based Brill Tagger for Polish. Recent Advances in Intelligent Information Systems, Kraków, Poland, 2009, pp. 67–76.
2. *Brill E.* Unsupervised Learning Of Disambiguation Rules For Part Of Speech Tagging. Proceedings of the Third Workshop on Very Large Corpora. Cambridge, Massachusetts, USA, 1995.
3. *Sharoff S., Nivre J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011". Bekasovo, 2011.
4. *Sokirko A., Toldova S.* Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian,
available at http://aot.ru/docs/RusCorporaHMM.htm
5. *Zelenkov J., Segalovich I., Titov V.* Probabilistic model for morphological disambiguation based on normalising substitutions and adjacent words positions. [Verojatnostnaja model' cn'atija morfologicheskoj neodnoznachnosti na osnove normalizujushchih podstanovok i pozicij sosednih slov]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005". Zvenigorod, 2005.

КОНДУКТОР, НАЖМИ НА ТОРМОЗА…[1]

Рахилина Е. В. (rakhilina@gmail.com)

НИУ Высшая школа экономики, Москва, Россия;
Институт русского языка им. В. В. Виноградова
РАН, Москва, Россия

В статье описываются семантико-синтаксические особенности русской формы *постой* — аттенуативного императива от глагола *постоять*. Ее свойства рассматриваются в ряду других квазиграмматических показателей континуативного прохибитива: *прекрати, перестань, хватит, будет, оставь, хорош* и др. Все они используются как побуждение прервать некоторую ситуацию. Показано, что специфической частью семантики *постой* является аттенуативность, так что соответствующее действие (конкретное, происходящее в данный момент, но при этом не выраженное поверхностно) прерывается на время; это время говорящий предлагает использовать для того, чтобы каким-то образом оптимизировать ход событий, а бессоюзное придаточное, которое входит в конструкцию с *постой*, эксплицирует предложение говорящего.

Ключевые слова: лексическая семантика, грамматическая семантика, квазиграмматические маркеры, грамматикализация, прохибитив

CONDUCTOR, PRESS THE BRAKES…

Rakhilina E. V. (rakhilina@gmail.com)

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

The paper examines the lexical semantics and syntax of the form *postoj* (attenuative imperative of Russian verb *postojat'* 'stand (for a while)') describing it as one of the quasi-grammatical markers of continuous prohibitive, such as *prekrati, perestan', xvatit, budet, ostav', xoroš,* etc. All of them mark the illocution for interrupting the ongoing situation. *Postoj* differs from the other markers by its attenuative semantics, so that the situation (definite and taking place at the moment of speech, but not explicated in the sentence) has to be interrupted only for a while. The speaker offers to use this short span of time to improve it with some additional means; asyndetic clause, which follows *postoj*, explicates the speaker's suggestion.

Key words: lexical semantics, semantics of grammar, quasigrammatical markers, grammaticalization, prohibitive

---

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2013 году.

0. Введение

С лингвистической точки зрения начало известной песни, строчка из которой вынесена в заглавие статьи (*Постой, паровоз! Не стучите, колеса!*), нестандартно: оба императива имеют в качестве адресата неодушевленные объекты. Понятно, что используя императив, который обычно обращен к одушевленному адресату (*Дети, почему у вас такой грохот? Не стучите так громко, дедушка работает!*), в контексте предметного имени, автор имел в виду метафорический сдвиг, или одушевление: как если бы колеса стали контролирующим действие участником и сами могли перестать стучать. Но если с глаголом *стучать* такой сдвиг вполне возможен, то к императивной форме *постоять* он плохо применим и дает эффект грамматической неправильности, потому что исходное значение формы *постой* ('не двигайся некоторое время') в русском языке уже претерпело некоторые изменения и даже самым обычным семантическим сдвигам так просто не подчиняется.

Справедливости ради надо сказать, что свою каноническую интерпретацию аттенуативного императива *постой* все же сохранило — но только в некоторых специальных контекстах.

Во-первых, это контексты, задающие какие-то физические характеристики пространственной ситуации, которую описывает этот глагол: локативные (*постой у порога*), временны́е (*постой пока, Постой! … Постой одну минуту! Не убегу, успеешь захватить!* [С. Я. Надсон. Царевна Софья (1880)]), образа действия (*постой спокойно*) и под. Во-вторых, это предложения, в которых делается акцент на контрасте покоя адресата и движения какого-то другого участника, например, субъекта: *ты постой, а я пройдусь* или: *Поднимись на откос и постой, оглядись* [Н. В. Крандиевская. «Как песок между пальцев, уходит жизнь…» (1938–1940)].

Для сравнения — аттенуативная интерпретация абсолютно недопустима для *постой* при удвоении, которое можно было бы назвать морфологическим и которое достаточно характерно для этой формы. При морфологическом удвоении — в отличие от простого повтора лексемы — две формы произносятся неразрывно и акцент переносится на вторую: *Постой-постой, — сказал я. — Повтори-ка ещё раз!* [Вера Белоусова. Второй выстрел (2000)], но не: **постой-постой, а я пройдусь*! (ср. здесь приведенный выше пример из Надсона с простым повтором, где обе формы равно значимы, и значение аттенуативно).

Не очень распространена аттенуативная интерпретация и для изолированных употреблений императива, или изолированных с постпозицией адресата — то есть как раз построенных по образцу *постой, паровоз!* Скорее, им свойственна семантика прохибитивного континуатива[2], т. е. императива, прерывающего (в нашем случае — на время, потому что аттенуативность в этом значении *постой* сохраняется) уже идущее действие, ср. у Пушкина:

---

[2] Такое значение не отмечено в Майсак 2005 как дериват от 'стоять', но полностью соответствует переходу "STOP > PROHIBITIVE", который засвидетельствован в Heine, Kuteva 2002.

постой (= ‘прерви свою речь на время’) — *а карантин!* Призыв прервать ситуацию, выраженный таким образом, может быть использован далеко не всегда. Например, нельзя (по крайней мере в современном русском языке) сказать: *Постой!,* обращаясь к человеку спящему, размышляющему или мечтающему. В то же время, круг ситуаций, к которым этот императив, в принципе может быть применен, достаточно широк, ср.: *постой* (= ‘не садись пока / не наступай, не ложись’), *я тут вытру; постой* (= ‘не ешь пока’), *я сметаны добавлю; постой* (= ‘не чисти пока’), *я тебе хорошую щетку дам; постой* (= ‘не плати пока’), *я мелочь достану; постой* (= ‘не включай пока телевизор’), *я очки надену* и др. под. Ср. также: *Постой, постой!* (= ‘не пей’) … *Ты выпил… без меня?* [А. С. Пушкин. Моцарт и Сальери (1830)] или: *Рыдает бедная хозяйка./ Хозяйка милая, постой* (= ‘не рыдай’)*,/ На картах лучше погадай-ка.* [Д. Хармс. «Григорий студнем подавившись…» (1937.02.20)] и др.

Судя по данным НКРЯ, частотность *постой* — в том числе и, видимо, как прохибитивного континуатива, заметно падает[3]: параллельно в русском языке в том же значении требования временно прервать идущее действие используются более продуктивные *погоди* и *подожди.* Интересно, что все три формы работают, в общем, параллельно и чаще всего встречаются в речевых контекстах. Причины, по которым возникает требование прервать речь адресата, у них одни и те же:

(1) Говорящий заметил противоречие в речи адресата;

(2) Говорящий видит незаметную для адресата логическую связь, вывод и проч.;

(3) Говорящий осенен новой собственной идеей, что-то вспомнил, увидел и проч.;

(4) Говорящий хочет совершить другое действие, ср.: *постой* (= ‘прерви свой рассказ, не говори пока’), *я дверь открою*.

1. О системе прохибитивно-континуативных значений

Русский язык обладает довольно сложной системой лексических маркеров прохибитивно-континуативных значений, в которую *постой*, вместе с *подожди* и *погоди*, довольно хорошо встраивается. Все это слабо грамматикализованные показатели, специального грамматического средства выразить этот смысл в русском языке нет. Как мы показали на материале количественных значений (Рахилина, Ли Су Хен 2005, 2010), особенность квазиграмматических маркеров в том, что они, частично сохраняя «следы» своей исходной семантики, привносят

[3] Косвенное свидетельство этому — высокая частотность *постой* в поэтическом подкорпусе НКРЯ XIX — первой половины XX в.

в систему противопоставлений, бедных и простых в «чистой» грамматике, качественные характеристики. А поскольку сами конструкции достаточно частотны и продуктивны, соответствующие им качественные значения оказываются выделены языком как когнитивно релевантные. С этой точки зрения их исследование особенно интересно — нам важно знать (в том числе и для последующих типологических исследований), какие именно различия в прохибитивной зоне настолько значимы, что хотя бы в каком-то языке могут лексикализоваться.

Итак, русский язык развил лексические показатели прохибитивного континуатива для:

(1) Неприятных, вредных, отрицательно оцениваемых действий: *прекрати, перестань* (а также более разговорное *кончай*) — *прекрати безобразничать / таскать кота за усы*;

(2) Избыточной (но не вредной, и, может быть, даже положительно оцениваемой) деятельности: (устар.) *будет, по́лно*: *Да будет тебе сердиться*; *— Да уж вышло, Егор! Будет тебе скромничать! — Не совсем, Валя, не совсем!* [А. П. Платонов. Эфирный тракт (1926–1927)]; *Не прилично, что ли? Да полно форсить-то!* [Ф. М. Достоевский. Идиот (1869)].

(3) Избыточного количества объектов или вещества, передаваемого адресатом субъекту: *хватит, довольно, достаточно,* также *будет, хорош,*

(4) Движения: *стой, остановись, стоп*

(5) Сна: *проснись, вставай*,

(6) Предположений или рассуждений — (устар.) *по́лноте,* (разгов.) *брось*: *Физик, о чем ты? Брось./ Законы природы, закон Иеговы —/ То вкривь, то вкось.* [И. В. Чиннов. «Мир, созданный Богом, и мир, возникший…» (1984)]

а также некоторых особых речевых ситуаций, а именно:

(7) Таких ситуаций, в которых говорящего раздражают вопросы или советы адресата — *отстань*: *Кстати, Тёткин, что такое эполета? — Отстань. Дело не в этом*. [И. Грекова. На испытаниях (1967)]

(8) (устар.) Утешения — *оставь / оставьте*: *Оставьте… помощь ваша бесполезна и для соперника вашего и для моей соперницы…* [А. Ф. Вельтман. Эротида (1835)]

Некоторые лексемы из этого списка узко специализированы, ср. (4)–(8), так что эти случаи лучше квалифицировать как лексикализацию прохибитива.

Между тем другие, как *прекрати* и *перестань,* имеют яркую устойчивую семантику и могут претендовать на статус квазиграмматических маркеров (см. также Кустова 2011). Они настолько связаны с отрицательной оценкой, что в их контексте любое доброе дело переинтерпретируется как вредное или неуместное (с презумпцией, что эта точки зрения говорящего известна и адресату), ср. *прекрати улыбаться* (*— стоишь, как дурак*!), *прекрати мыть полы* (*— опять спина болеть будет!*). Так же ведет себя и *хватит*. Однако в отличие от *хватит, прекратить* и *перестать*, не используются в количественных значениях (ограничивающих ситуацию через количество передаваемого объекта) — и тем самым выделяют количественные употребления (*хватит уже воды*, *больше не лей*!) в особый класс. Ср. здесь *достаточно*, для которого, напротив, естественны количественные контексты: **достаточно дергать кошку за усы!*

Другое свойство *прекратить* и *перестать* состоит в том, что они представляют ситуацию как дробную, состоящую из множества мелких, т. е. служат инструментом ее мультипликации. Поэтому предложения типа *прекрати поднимать руку / вытягивать шею / улыбаться* и проч. интерпретируются не как единичное непрерывное действие, которое нужно остановить на середине (ср.: **прекрати открывать глаза*), а как ***последовательность*** таких действий, которую нужно прервать.

Особняком стоят речевые ситуации — их свойства с точки зрения естественного языка настолько разнообразны, что они допускают любые неспециализированные маркеры континуативного прохибитива. Действительно, с одной стороны, речь — это процесс, который легко представить как прерывистый и который может оцениваться и положительно, и отрицательно (*прекрати кричать / будет тебе его уговаривать*), а с другой стороны — речь это передача информации от адресата к говорящему и количество этой информации можно ограничить (*хватит / довольно / достаточно — переходите к следующему вопросу билета*).

Еще одно важное противопоставление в этой семантической зоне, которое не хотелось бы упустить, касается степени участия в ситуации говорящего и слушающего: как видно из примеров, в стандартном случае слушающий ситуацию контролирует, а говорящий прекращает. Однако предикативы, например, *довольно* или *хватит,* легко допускают и включение в ситуацию говорящего, демонстрируя своего рода инклюзивность, ср. известное: *Довольно кукситься! Бумаги в стол засунем!*, при *будет* возможно отсутствие контроля ситуации со стороны адресата: *Будет тебе мыкаться! Женись и живи*. [И. А. Гончаров. Обрыв (1869)], а *хватит* может быть обращено к третьему лицу: *Хватит ему мыкаться!*

Бегло очертив контуры системы русских прохибитивных континуативов[4], определим место *постой* (*погоди / подожди*) в этой системе. На наш взгляд,

[4] Кажется, что в ряду прохибитивных континуативов следует рассматривать и конструкцию больше не: *больше не (при)ходи ко мне.* Тем не менее, на наш взгляд, семантика ее скорее инцептивная, чем континуативная, потому что к моменту произнесения такой прохибитивной формулы соответствующее действие или последовательность действий уже должно прекратиться, так что требование состоит не в том, чтобы оно не продолжалось, а в том, чтобы оно не возобновлялось.

постой — это именно *континуативный* прохибитив, он не используется в количественных контекстах и, в отличие от других маркеров этой зоны, может прерывать и гомогенный процесс (а не только последовательность действий), всегда контролируемый адресатом. Важной частью семантики *постой* остается аттенуативность, поэтому соответствующее действие прерывается на время; это время используется говорящим для того, чтобы каким-то образом оптимизировать ход событий. Как видим, *постой* и группа его близких синонимов соотносится с довольно своеобразным классом ситуаций, не похожих на другие в том же ряду.

2. Денотативный статус

Но зачем вообще нужны особые маркеры прохибитива, почему русскому языку недостаточно простого отрицания императивной формы, применимого (так по крайней мере кажется на первый взгляд), к любому глаголу: ***не*** *ешь,* ***не*** *пей,* ***не*** *спи,* ***не*** *ходи*?

Дело в том, что денотативный статус обычной прохибитивной конструкции, как правило, неопределен и совсем необязательно соотносится с актуальной, конкретно-референтной ситуацией. Указания типа: *Не ешь соленую капусту, тебе вредно*, универсальны, они касаются любого случая, в котором адресату представится данная возможность — требование состоит в том, чтобы в этот произвольный момент такой возможностью сознательно пренебречь. Но и в целом семантика императива как формы манипуляции адресатом больше нацелена на будущее, чем на прошлое, в котором уже ничего нельзя изменить: не важно, идет ли это действие сейчас — важно, чтобы в будущем оно уже не осуществлялось.

Такая семантическая стратегия гораздо ближе к инцептиву, чем к континуативу, и для некоторых русских глаголов именно инцептивная интерпретация прохибитива (вопреки Храковский, Володин 1986: 99) оказывается единственно возможной. Ср.: *не ходи туда* или характерное *не спи* (= ‘не засыпай’); эта формула используется тогда, когда человека будят, он слегка просыпается, но еще не пришел в себя и может заснуть снова. Любопытно, что есть случаи, когда простой прохибитив кажется вовсе грамматически неправильным, как **не иди*[5] или по крайней мере значительно менее приемлемым в континуативном контексте, чем особая аналитическая квазиграмматическая форма из тех, что мы обсуждали. Ср. здесь пары типа: **не ходи, голова болит* VS. *перестань ходить, голова болит* или: *эй, ты, если не трус, перестань прятаться*! (А. Слаповский) — *?не прячься!* (в актуальном значении). В любом случае, денотативная разница между, например, *не разговаривай* и *прекрати разговаривать* очевидна: первое действительно ориентировано на отсутствие действия

[5] В НКРЯ встретилось всего 78 примеров, большинство из которых либо идиомы (*не иди замуж / на поводу / иди во власть* и др. под.), либо с противительным отрицанием как в: *Ты наверняка сказала ему не “иди” домой, а “езжай” домой*. Для сравнения — на *не стой* корпус выдает 149 «полноценных» употреблений.

в будущем безразлично к тому, насколько оно актуально, а второе ясно указывает на конкретно-референтный процесс, который нужно прервать[6].

3. Дейктичность

Раз континуативные прохибитивы, в отличие от обычных, непосредственно связаны с актуальной ситуацией, значит, соответствующие показатели должны тяготеть к дейктичности. Действительно, большинство из них способно и даже склонно употребляться изолированно и не называть прерываемую ситуацию, которая легко восстанавливается из контекста, ср.: *Остановись!* (**остановись идти*), *Отстань!* (**отстань задавать вопросы*), *Постой!* (**постой говорить*[7]), *Вставай*! (**вставай спать*) и др. — и в том же ряду изолированные употребления *Хватит! Перестань! Прекрати!* и под.

Любопытно, что недейктические контексты, содержащие в предложении глагол, не просто отсылают с его помощью к прерываемой ситуации (которая, повторим, и так известна из внеязыкового контекста), а вводят каждый свою дополнительную семантику. Такова, в частности, роль инфинитива при *перестань — прекрати*: в основном он служит не для того, чтобы ввести актуальное действие как таковое, а чтобы ввести его оценочную интерпретацию (другими словами, обидеть адресата), ср.: *Прекрати безобразничать / дурачиться / вредничать / притворяться!* Количественные контексты тоже, как правило, дейктичны, но могут и содержать именную группу, описывающую ту субстанцию, которая в данном случае ограничивается — особенно если есть выбор объектов: *чаю достаточно, а сахара добавь*.

Что касается формы *постой*, то и она не всегда употребляется изолированно: свойственная ей конструкция вводит предикатную лексему как сказуемое (обычно в форме будущего времени) сочиненного с ним предложения: *постой, я тебе вот что скажу*. Однако эта лексема обозначает действие, внешнее

---

[6] Как видим, замеченное противопоставление, заставило нас использовать в отношении глагольной лексики нетрадиционную для нее референциальную терминологию — впрочем, нам и раньше казалось, что референциальные отношения всегда были несколько искусственно сужены до предметной зоны. В свое время мы показали их применимость к отпредикатным именам (Крейдлин, Рахилина 1981), подробная лексическая классификация прохибитивов могла бы, как кажется, продвинуть теорию денотативных статусов для предикатной лексики в целом.

[7] Заметим, что в XIX и даже в XX употребление инфинитива при постой еще было не запрещено, ср. *Нет, постой коня седлать!* .. [А. С. Норов. Чельд-Гарольд (1824.08.20)] или: «Мой желанный! Мой любимый! » — *Нет, постой меня ласкать*. [К. Д. Бальмонт. Нереида 1903] и: — *Ты постой обижаться,* — остановил гнев Прокофия старик. [А. П. Платонов. Чевенгур (1929)]. Гораздо более частотны и современны контексты с инфинитивами для *подожди* (86 примеров в основном корпусе и 20 в устном) и *погоди* (150 в основном и 13 в устном), ср.: Ты подожди расстраиваться / может и не надо ничего. [Разговор матери с сыном // Из материалов Ульяновского университета, 2006]. Правда, это ничтожная часть по сравнению с числом их дейктических употреблений (без инфинитива) — соответственно, 4064 и 4239.

по отношению к уже идущему — а именно то, которое, в соответствии с семантикой *постой*, поможет оптимизировать ситуацию в целом и ради которого она должна быть приостановлена. Таким образом, в отношении временно прерываемой ситуации *постой* всегда дейктично — и это дополнительное свидетельство его «полноценности» как показателя именно континуативного прохибитива.

4. Семантические возможности развития *постой*

Значение отложенного продолжения действия порождает по крайней мере еще один смысл, который можно было бы назвать «отложенным угрозативом»: повеление ждать момента, когда от говорящего последует наказание. Это значение встраивается в общую парадигму сдвигов, свойственных (*подо*)*ждать* и (*по*)*годить*, следующих в русском сходной словообразовательной модели (ср. также англ. *Just you wait!,* фр. *Attends tu vas voir!*). В то же время, в современном русском языке оно почти совершенно вытеснено известным *Ну погоди* и синонимичным ему *подожди!* и для *постой* может считаться практически утраченным. Ср. явно устаревшие примеры типа: «*Постой же! — /Думает, — я тебя так не оставлю, проказник; поймаю / И за побег накажу — накормлю хорошенько*! [Н. А. Некрасов. Карп Пантелеич и Степанида Кондратьевна (1844–1845)] Мыслит царевич: «*Добро же! постой! » / За косу ловко схватил он рукой*. [М. Ю. Лермонтов. Морская царевна (1841)] и вполне современное: *Подлец! Ну, подожди у меня! — Что ты ему сделаешь?* [Василь Быков. Знак беды (1982)] или: *А подожди у меня еще. Какой у нас козырь?* [Разговоры за игрой в карты (2009)]

С теоретической точки зрения, это очень интересно: наблюдая квазиграмматикализацию серии сложных конструкций (Летучий, Рахилина 2012, Рахилина, Ли Су Хен 2010), мы видим, что внутри своей семантической зоны (множественности, итеративности и др.), они выстраиваются в своего рода шкалу, упорядоченную по степени грамматикализованности и по умолчанию предполагаем их постепенное развитие как продвижение по этой шкале. Между тем, никакого продвижения может не произойти: бывает, что единица застывает в каком-то фиксированном контексте и лексикализуется в нем в новом качестве надолго. *Постой* служит примером еще одной альтернативы: достаточно нетривиальное значение развилось и встроилось в систему прохибитивных континуативов, поддерживаемое морфосинтаксически и семантически близкими «соседями», а потом, ими же вытесненое, исчезло. Система откатилась назад, демонстрируя память о своих предшествующих, исходных этапах — в частности, не вполне грамматически правильными, но прекрасно поддающимися анализу употреблениями типа *Постой, паровоз!*

Литература

1. *Крейдлин Г. Е., Рахилина Е. В.* 1981. Денотативный статус отглагольных имен. НТИ, сер. 2, , N 12, с. 17–22.

2. *Кустова Г. И.* 2012. Об иллокутивной фразеологии Ю. Д. Апресян, И. М. Богуславский и др. (ред.) Смыслы, тексты и другие захватывающие сюжеты М., Языки славянской культуры 349–367.
3. *Майсак Т. А.* 2005. Типология грамматикализации конструкции с глаголами движения и глаголами позиции М., Языки славянской культуры.
4. *Рахилина Е. В., Летучий А. Б.* Русские конструкции с временным значением: о границах настоящего времени В кн.: Präsens: Сборник научных трудов. Москва: Олма Медиа Групп, 2012. С. 224–242.
5. *Рахилина Е. В., Су Хен Ли* 2009. Семантика лексической множественности в русском языке Вопросы языкознания, N 4, 13–40.
6. *Рахилина Е. В., Су Хен Ли* 2010. О категории лексической множественности // Е. В. Рахилина (отв. ред.). Лингвистика конструкций. М.: Азбуковник, — С. 352–397.
7. *Храковский В. С., Володин А. П.* 1986. Семантика и типология императива: Русский императив. Л. Heine B., Kuteva T. 2002. World Lexicon of Grammaticalization Cambridge University Press, Cambridge.

References

1. *Heine B., Kuteva T.* (2002), World Lexicon of Grammaticalization. Cambridge, Cambridge University Press.
2. *Hrakovskij V. S., Volodin A. P.* Semantics and typology of imperative [Semantika I tipologija imperativa]. Russkij imperativ [Imperative in Russian]. Leningrad, 1986.
3. *Krejdlin G. E., Rahilina E. V.* (1981), Denotation of deverbal nouns [Denotativnyj status otglagol'nyh imën]. Nauchno-tehnicheskaja informatsija [Scientific-technical information], series 2, no. 12, pp. 17–22.
4. *Kustova G. I.* On the illocutive phraseology [Ob illokutivnoj frazeologii]. Smysly, teksty i drugie zahvatyvajushchie sjuzhety [Meanings, texts, and other exciting things]. Apresjan Ju.D, Boguslavskij I. M. et al. (eds). Moscow, Jazyki slavjanskoj kul'tury, 2012, pp. 349–367.
5. *Majsak T. A.* (2005), Tipologija grammatikalizatsii konstruktsii s glagolami dvizhenija I glagolami pozitsii [The typology of grammaticalization in the construction with verbs of movement and position]. Moscow, Jazyki slavjanskoj kul'tury.
6. *Rahilina E. V., Letuchij A. B.* Constructions with a temporal meaning in Russian: On the borders of Present tense [Russkie konstruktsii s vremennym znacheniem: o granitsah nastojashchego vremeni]. Präsens: Sbornik nauchnyh trudov [Präsens: A research essays volume]. Moscow, Olma Media Group, 2012, pp. 224–242.
7. *Rahilina E. V., Li S.-H.* (2009), The semantics of lexical plurality in Russian [Semantika leksicheskoj mnozhestvennosti v russkom jazyke]. Voprosy jazykoznanija [Issues of linguistics], no. 4, pp. 13–40.
8. *Rahilina E. V., Li S.-H.* On the category of lexical plurality [O kategorii leksicheskoj mnozhestvennosti]. Lingvistika konstruktsij [Construction linguistics]. Rahilina E. V. (ed.), Moscow, Azbukovnik, 2010, pp. 352–397.

ИСПОЛЬЗОВАНИЕ КОНТРАСТА И ЭМФАЗЫ ДЛЯ ПЕРЕДАЧИ ИМПЛИЦИТНЫХ СМЫСЛОВ

Савинич Л. В. (savinitch@iitp.ru)

Институт проблем передачи информации
им. А. А. Харкевича РАН, Москва, Россия

Ключевые слова: коммуникативная стратегия, контраст, эмфаза, импликация

USE OF CONTRAST AND EMPHASIS FOR CONVEYING IMPLICIT MEANINGS

Savinitch L. V. (savinitch@iitp.ru)

A. A. Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia

The paper analyzes contrast and emphasis, modifiers of communicative meanings, their semantics and accent structure in the sentences examined. We argue that contrastive and emphatic highlighting of one of the utterances components in the given examples are made by the speakers strategically, in order to convey occasional implicit meanings. All examples are illustrated with graphs displaying tone fluctuations, sound intensity, modulation of sound, and other prosodic features.

Key words: communicative strategy, contrast, emphasis, implication

Исследуя особенности судебного дискурса, мы обратили внимание на различное просодическое оформление говорящим идентичных компонентов высказывания: сначала — без контрастного выделения, затем — с контрастом. Возник вопрос: делалось ли акцентное выделение случайно или стратегически осознанно, то есть для выражения определённого коммуникативного намерения говорящего? Если стратегически осознанно, то с какой целью? Таким образом, возникла задача данного анализа: в рассматриваемых высказываниях определить коммуникативные стратегии говорящего, исследовав семантику акцентов, просодические характеристики словоформ-акцентоносителей, акцентную структуру предложения.

Прежде, чем приступить к анализу наших примеров, мы приведём определение коммуникативной стратегии в теории речевых актов. «Коммуникативная стратегия говорящего состоит в выборе коммуникативных намерений, распределении квантов информации по коммуникативным составляющим и выборе порядка следования коммуникативных составляющих в предложении» [Янко 2001: 38]. Этой же проблеме посвящена работа [Бергельсон, Кибрик 1987: 52–63], в которой рассматриваются приоритетные стратегии, являющиеся механизмом текстообразования. Данные механизмы выборов осуществляют формирование высказывания, выделяя коммуникативно более значимые компоненты смысла и ослабляя менее важные. В другой работе на данную тему [Всеволодова, Ященко 2008: 11–14.] анализируется «комплекс коммуникативных задач» в процессе вербализации ситуации.

Коммуникативные стратегии говорящих реализуются в структурах носителей коммуникативных значений и могут выражать намерения говорящих сделать сообщение, задать вопрос, высказать просьбу, отдать приказ и другие. В данной статье будет рассматриваться роль модифицирующих коммуникативных значений, которые не относятся к категории основных иллокутивных значений (как сообщение, вопрос, просьба, мольба), а только модифицируют основные типы иллокуций и их коммуникативных компонентов. К таким модификаторам относятся контраст, верификативное, или да-нет, значение и эмфаза. Их семантика и акцентная структура подробно описаны в [Янко 2001; 2008; эмфаза — в Фаустова 2009]. Мы рассматриваем только два из них — контраст и эмфазу.

1. Характеристика словоформ-акцентоносителей

Наш первый пример относится к области юриспруденции и записан на аудионоситель во время выступления в суде государственного обвинителя. Текст зачитываемого обвинения подготовлен прокурором по установленному канону: пример начинается с лаконичной формулировки происшедшего события; затем следует описательная часть — детальное изложение происшествия в хронологическом порядке; в завершении событие квалифицируется в соответствии с существующим законодательством. Нас будет интересовать только предложение в заголовочной части, которое формулирует основной состав преступления:

(1) ***Торлосу́нов↘ Алимбек***[1]*↗ обвиняется в том, что он применил насилие, не опасное для жизни и* ***здоровья,↗*** *в отношении представителя* ***власти↘↗*** *и с исполнением им своих должностных* ***обязанностей****.↘*

В этом примере полужирным шрифтом выделены словоформы — акцентоносители коммуникативных значений; изменение частоты основного тона отмечено стрелками, которые в примерах расположены после словоформы-акцентоносителя; физические параметры записи представлены на двух панелях ниже, которые мы условно называем тонограммой (см. Тонограмму 1). Верхняя панель представляет собой осциллограмму, или преобразованный в машинную форму след, который оставляет на материале, (например, на бумаге) игла, возбуждённая звуковыми волнами. Осциллограмма отражает структуру слога и паузы в пределах анализируемого предложения; собственно тонограмма, представленная на нижней панели фиксирует изменения частоты основного тона в герцах, воспринимаемые как понижения и повышения тона и их комбинации.

Мы не будем останавливаться на принципах выбора акцентоносителей (об этом см. [Янко 2001: 69–84; Янко 2008: 38–60; Кодзасов 1999: 196–216; Кодзасов 2009: 73–93]. Мы только охарактеризуем все носители акцентов в анализируемом примере и укажем их некоторые просодические характеристики.

В начале предложения носителями акцентов служат два компонента имени собственного ***Торлосу́нов↘ Алимбек↗*** (Причём, в судебном дискурсе именно данная последовательность «фамилия–имя» или «фамилия–имя–отчество» является канонической.) Первый акцентоноситель этой группы выражен словоформой ***Торлосу́нов↘*** с нисходящей интонацией по типу ИК-2[2], характеризующейся более интенсивным падением тона на ударном слоге, чем у ИК-1, и подъёмом тона на предударных слогах, если они есть, что и обеспечивает крутое падение на ударном слоге, как зафиксировано на начальном фрагменте Тонограммы 1, ограниченного курсорами. (Описание интонационной конструкции ИК-2 см. в работах [Брызгунова 1980: 98–111; Янко 2008: 32, 189]. Об экспансии акцента ИК-2 и характерном его использовании в речи лекторов и дикторов см. в [Янко 2004].)

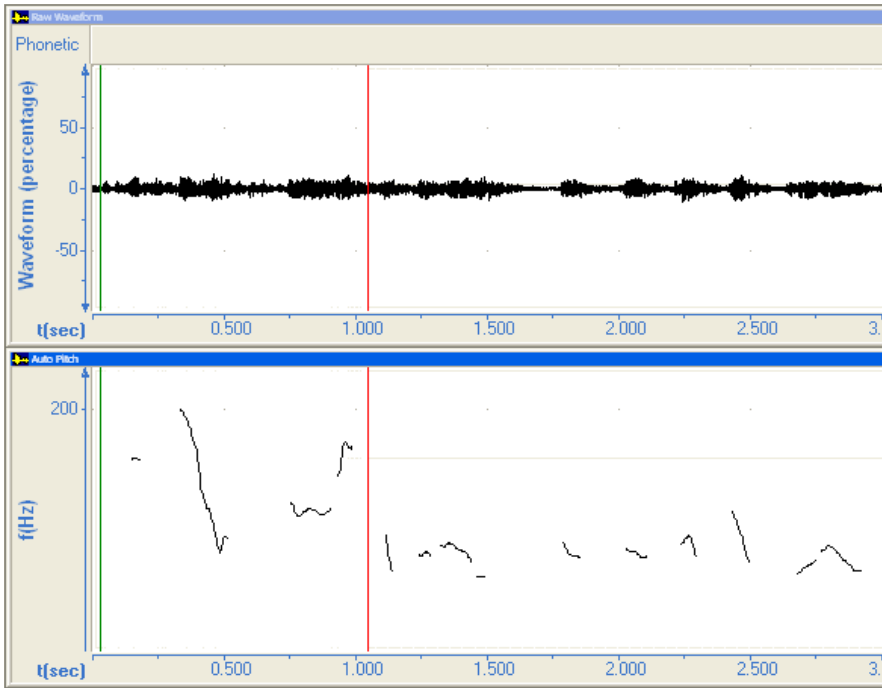
Второй акцентоноситель выражен словоформой ***Алимбек↗*** и произносится с повышением тона на последнем ударном слоге по типу ИК-3, маркируя тему высказывания. Это повышение отчётливо видно на нижней панели Тонограммы 1 перед вторым курсором.

---

[1] Для соблюдения этики анализа звучащих записей мы заменили имена собственные на вымышленные с сохранением фонетическо-акцентной структуры.

[2] Обозначение интонационных конструкций представлено в формулировке Е. А. Брызгуновой [Брызгунова 1980].

Тонограмма 1



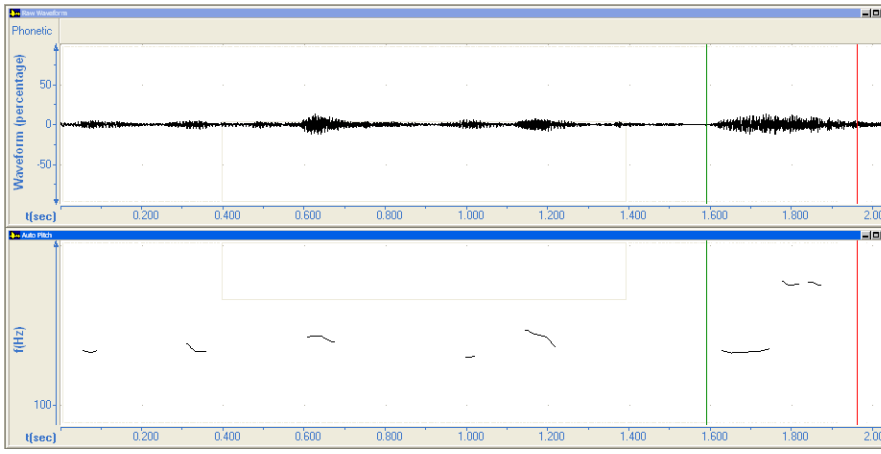
Торлосу́нов↘ Алимбек↗…

Следующая, третья, словоформа-акцентоноситель ***здоровья↗*** является членом атрибутивной группы *не опасное для жизни и* ***здоровья↗***. Она произносится с подъёмом тона на ударном слоге и падением на заударной части по типу ИК-3. Восходящая интонация ИК-3 в данном случае является восходящим акцентом незавершенности, то есть выполняет не локальную, или относящуюся к формированию отдельного речевого акта, а дискурсивную функцию [Янко 2008: 128–170]. Иными словами повышение интонации на данном акцентоносителе не маркирует одну из коммуникативных составляющих, например тему, как на словоформе ***Алимбек,↗*** а обеспечивает связность дискурса, то есть указывает, что данный фрагмент текста не последний, и за ним следует продолжение.

На нижнем графе Тонограммы 2 зафиксирован лёгкий подъём тона на ударном слоге ***-ро-*** (к сожалению, из-за быстрого темпа речи последний слог ***-вья*** произносится неотчётливо). Обратим внимание, что вся предшествующая акцентоносителю часть, как это наглядно зафиксировано на тонограмме, произносится практически на одном ровном тоне, без резких частотных колебаний.

Четвёртый акцентоноситель, словоформа ***власти,↘↗*** произносится с понижением тона на ударном слоге и повышением на заударном слоге, по типу ИК-4, являясь акцентом незавершённости текста. И наконец, последний, пятый, акцентоноситель в данном предложении — словоформа ***обязанностей↘*** — произносится с понижением тона на ударном слоге и последующих заударных слогах по типу ИК-1, маркируя конец предложения.

Тонограмма 2



не опасное для жизни и ***здо - ровья***

Таким образом, в процитированном отрывке из выступления государственного обвинителя мы представили некоторые просодические характеристики словоформ-акцентоносителей, обозначили коммуникативные составляющие данного предложения и можем заключить, что основной коммуникативной стратегией прокурора было выступить с сообщением о факте происшествия.

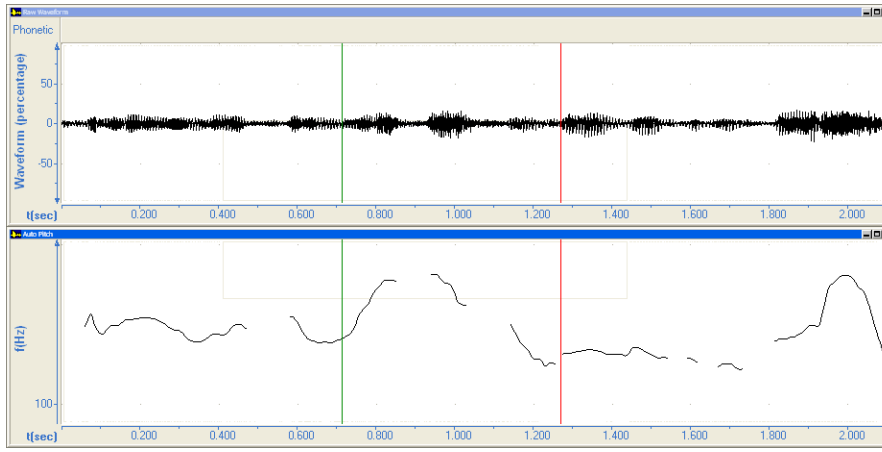
2. Характеристика контрастного выделения акцентоносителя

При зачитывании описательной части совершённого преступления, государственный обвинитель ещё раз повторил произнесённую ранее атрибутивную группу *насилие, не опасное для жизни и здоровья*, но уже иначе расставляя коммуникативно релевантные акценты:

(2) …*насилие,* ***не опасное***↘ *для жизни и* ***здоровья***↗ —

с отчётливым акцентным выделением постпозитивного прилагательного ***не опасное,***↘ повышением тона на его предударных слогах и падением на ударном и заударных слогах по типу ИК-2, как зафиксировано на Тонограмме 3 между курсорами. (Безударная словоформа ***не***, являясь проклитикой, составляет одно фонетическое слово с последующим прилагательным и имеет с ним одно ударение.)

Тонограмма 3



не о-пас-ное↘ *для жизни и* ***здоровья***↗

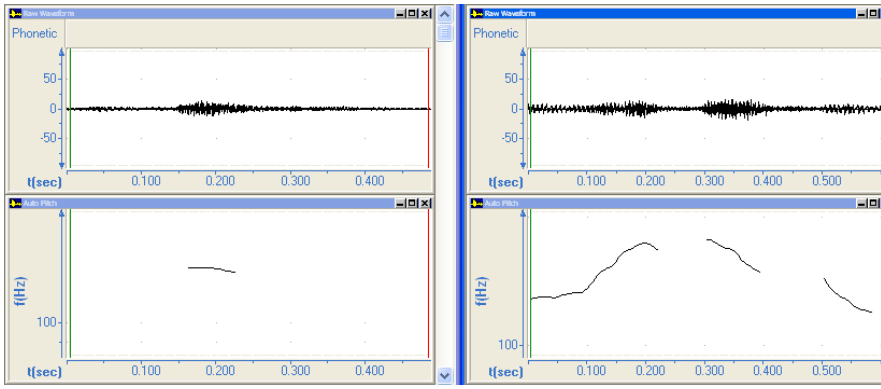
В последнем варианте новый акцент модифицирует значение коммуникативной рематической составляющей и приобретает новое значение — контраста.

Семантика контраста в данном примере связана с мысленной процедурой выбора из множества вариантов, ассоциирующихся с интонационно выделенным компонентом и известных собеседникам [Янко 2001: 47][3]. Эта же мысль высказана в другой работе, в которой контраст характеризуется «способностью соотносить содержание соответствующих высказываний с некоторым концептуальным множеством, по определённым правилам выводимым с учётом как содержания высказывания, так и различного рода знаний» [Паршин 1988: 10]. В нашем случае это множество может ограничиваться, например, вариантами: *не опасное* vs. *опасное*.

Просодическое выражение контрастной ремы с контрастом на прилагательном, как зафиксировано на графике правой нижней панели Тонограммы 4, существенно отличается от неконтрастного варианта на графике левой нижней панели.

[3] Мы не останавливаемся на других примерах контрастов, имеющих инклюзивное значение и значение опровержения. [см. Янко, 2001: 47]

Тонограмма 4



не опасное ***не о - пас - ное↘***

На следующих ниже графиках интенсивности — акустического коррелята громкости — иллюстрируется интенсивность звука обоих вариантов. На верхней панели, в варианте с неконтрастным компонентом, пик интенсивности на словоформе *не опасное*, отмеченный курсором, равен –19,4 децибела. На графике нижней панели, представляющем контрастный рематический компонент, пик интенсивности на акцентоносителе ***не опасное↘*** составляет –17 децибел. (Поясним эти цифры: абсолютная величина числа |19,4| больше абсолютной величины |17|, но эти числа со знаком минус, поэтому отрицательное число –19,4 меньше отрицательного числа –17, то есть контрастный вариант звучит громче.)[4]

Графики интенсивности

[4] Следует также пояснить, почему данные величины приводятся со знаком минус. Децибел (дБ) — логарифмическая единица отношения сигнала приёмника (воспринимаемого сигнала) к сигналу источника (посылаемого сигнала). Если сигнал приёмника меньше сигнала источника, то в данном случае мы получаем отрицательное значение децибела.

При этом отчётливо видно, что и предударный, и заударный сегменты акцентоносителя контраста (нижняя панель) также произносятся с повышенной интенсивностью тона по сравнению с аналогичными фазами неконтрастного варианта (верхняя панель). Помимо этого, наблюдается заметное увеличение длительности предударной части акцентоносителя контраста, поэтому курсоры, показывающие максимальное значение интенсивности, не совпадают.

Таким образом, во втором примере наблюдается стратегически осмысленное выделение коммуникативно значимого компонента предложения с целью, как указывалось выше, констатировать сделанный выбор из множества вариантов, ассоциирующихся с интонационно выделенным компонентом и известных собеседникам. (В нашем случае это два противопоставленных варианта: *не опасное* vs. *опасное*.) Между тем, мы полагаем, что в использовании коммуникативной стратегии имеется и другой, возможно, более существенный аспект: прокурор имплицитно оценивает, имплицитно квалифицирует совершённое преступление в соответствии с действующим законодательством. Подтверждает последний тезис прежде всего соотнесение с законодательным документом, из которого процитирована упомянутая атрибутивная конструкция и в соответствии с которым законом предусмотрены надлежащие меры воздействия.[5]

3. Эмфаза как модификатор коммуникативного значения

Наш следующий пример относится к эмфатическому выделению коммуникативного компонента.

Эмфаза (от греч. *ἔμφασίς* — разъяснение, указание, выразительность) — это выделение важной в смысловом отношении части высказывания (группы слов, слова или части слова), обеспечивающее *экспрессивность* речи[6]. Экспрессивность тесно связана с категорией эмоциональной оценки и в целом с выражением эмоций у человека. В то же время некоторые лингвисты утверждают, что не следует отождествлять эти категории. Экспрессивность — более широкое понятие, чем эмотивность, способная в ряде случаев занимать доминирующее положение в категории экспрессивности. Однако в языке существуют средства, которые имеют своей целью логически выделить определённые ча-

---

[5] См.: Уголовный Кодекс Российской Федерации.
Статья 318. Применение насилия в отношении представителя власти.
1. Применение насилия, не опасного для жизни или здоровья, либо угроза применения насилия в отношении представителя власти или его близких в связи с исполнением им своих должностных обязанностей — наказывается штрафом в размере до двухсот тысяч рублей или в размере заработной платы или иного дохода осужденного за период до восемнадцати месяцев, либо принудительными работами на срок до пяти лет, либо арестом на срок до шести месяцев, либо лишением свободы на срок до пяти лет.
2. Применение насилия, опасного для жизни или здоровья, в отношении лиц, указанных в части первой настоящей статьи, — наказывается лишением свободы на срок до десяти лет.

[6] Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990.

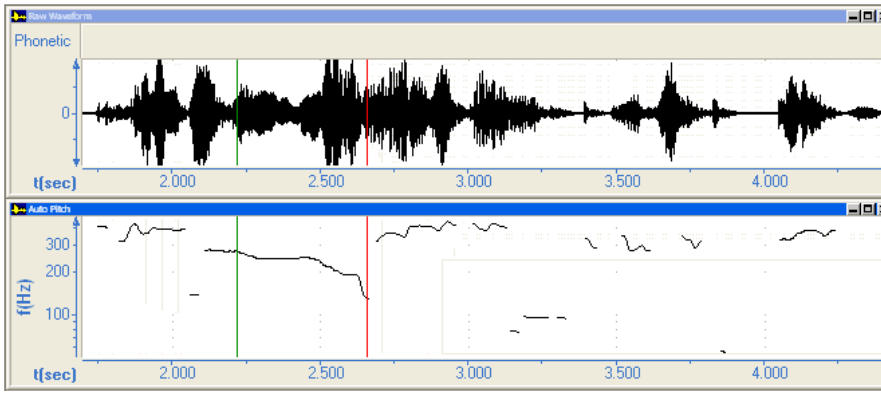
сти высказывания и которые не вызывают никаких чувств, но только служат цели вербальной актуализации высказывания [Galperin, 1977: 26–27] .

Исследование коммуникативного значения эмфазы, коммуникативные задачи, которые она призвана решать, а также средства её выражения содержатся в работах [Янко 2001: 64–67; 2008: 83–97; Фаустова 2006: 255–259, 2009: 96–101].

Наш следующий пример взят из записанного на аудионоситель чтения отрывка из художественного произведения в исполнении профессионального артиста, т.е. воспроизведение осуществляется, как и в первом примере, в «режиме озвучивания письменного текста». Возможно, для более полного понимания интересующего нас предложения, следует привести предшествующий ему текст.

(3) *<«Диктатор, не привычный к капризам примадонн, начал проявлять нетерпение. Он послал своего адъютанта передать антрепренёру, что, если занавес не будет сию же минуту поднят, всю труппу незамедлительно отправят в тюрьму. Хотя мысль о необходимости прибегнуть к таким мерам наполняет скорбью сердце президента.>*
В Макуто ↱ ***умели***↘ *заставить* ***пт****и****ч****ек* ***п****е****т****ь…»*

Тонограмма 5

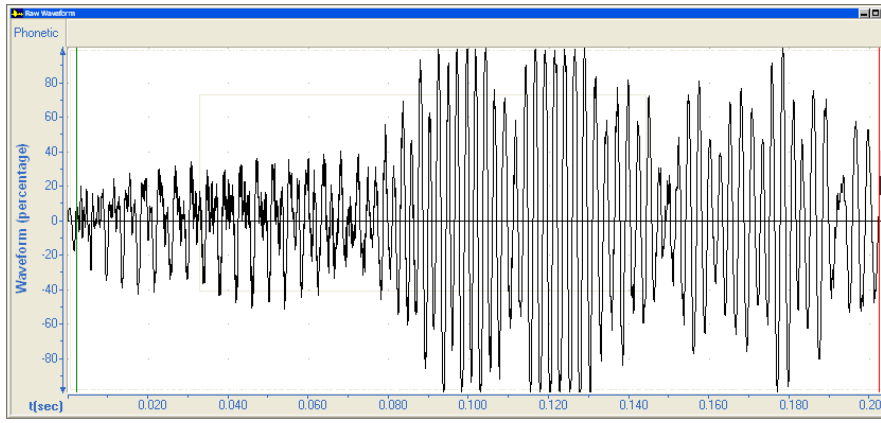


В Макуто ↱ ***уме-е-е-ли***↘ *заставить* ***пти-чек*** ***пе - ть.***

Рассмотрим словоформы-акцентоносители последнего предложения. Первая из них, ***В Макуто*** ↱ характеризуется восходящей интонацией на ударном слоге по типу ИК-6, маркируя тему предложения, так как название места действия, упоминавшееся ранее, уже известно слушателю. Акцент ИК-6 является «фонетическим вариантом» акцента ИК-3, но только при выражении темы высказывания. Отличие двух коммуникативно релевантных акцентов ИК-3 и ИК-6, которые оба характеризуются повышением частоты основного тона на ударном слоге акцентоносителя, состоит в том, что при акценте ИК-3 наблюдается падение тона на заударных слогах, а при акценте ИК-6 продолжается ровное движение тона после ударного слога, что отчётливо видно на нижнем графе Тонограммы 5 перед первым курсором.

Следующий акцентоноситель — словоформа ***умели***↘ является компонентом рематической составляющей и произносится с понижением тона на ударном слоге по типу ИК-1-Э[7]. При этом наблюдается существенное удлинение ударного гласного словоформы, зафиксированное на графике тонограммы нижней панели. Долгота звучания слога составляет 0,14 сек. Ударный слог ***-ме-*** акцентоносителя представлен на Осциллограмме 1.

Осциллограмма 1



Слева, в начальной фазе слога, осциллограмма с меньшим диапазоном колебаний фиксирует процесс образования губной смычки на звуке [м']. В дальнейшем диапазон частот увеличивается, однако пики колебаний то возрастают, то уменьшаются, обозначая своим контуром воображаемую волнистую линию. Эти перепады на слух воспринимаются как тоновые модуляции в пределах звучащего гласного [е]. На нижней панели Тонограммы 5 виден едва заметный изгиб тонограммы с лёгким подъёмом, плавным снижением и падением на заударном слоге ***-ли***. Такая модуляция тона в пределах одного продлённого ударного слога создаёт экспрессивность звучания.

В эмфатическом коммуникативном компоненте эмфатический акцент наблюдается не только на акцентоносителе, но и на других словоформах эмфатической коммуникативной составляющей, причём на акцентоносителе акцент наиболее интенсивный. Так, отмечается явление аспирации смычных согласных фонем [п], [т], [ч] в двух последних словоформах ***(пти-чек пе-ть)*** и увеличение интенсивности их звучания. Выразительность эмфазы поддерживается также лексическим способом — метафорическим переносом (*птичек* — 'певиц'). И наконец, для большей экспрессии артист выдерживает длительную паузу перед последней словоформой *петь*, длящуюся 0,14 секунд, что зафиксировано на осциллограмме верхней панели Тонограммы 5.

---

[7] Данное обозначение заимствовано из [Янко 2001: 64]. Акцентным выражением эмфазы служит эмфатический коррелят акцента ИК-1.

Как подтверждает проведённый анализ, интонационное выделение компонента предложения маркирует эмфазу на акцентоносителе рематической составляющей. При этом модифицированный компонент коммуникативной составляющей служит не для выражения чувств говорящего, как часто наблюдается при эмфазе, а именно для логического выделения коммуникативно значимого компонента предложения и в конечном итоге с намерением передачи имплицитного смысла — намёка на используемые противоправные меры.

Заключение

Подытоживая проведённый анализ, можно сделать вывод, что контраст и эмфаза, произнесённые с понижением тона на ударных и последующим падением на заударных слогах, были употреблены не только для модификации основных иллокутивных значений — сообщения (в первом примере) и связного нарратива (во втором примере), — но использованы осмысленно, намеренно как импликаторы неявных смыслов [об импликации см.: Grice H. P. 1975: 41–58; Грайс Г. П. 1985: 217–237].

Так, в первом примере из судебного заседания государственный обвинитель первоначально сообщает о случившемся происшествии, употребляя атрибутивную конструкцию *насилие, не опасное для жизни и* ***здоровья↗***. Конструкция прочитана ровным тоном, с восходящим акцентом на последнем компоненте. В данном случае восходящий акцент является акцентом незавершенности, обеспечивая связность дискурса. В дальнейшем, при повторном прочтении этой же конструкции, прокурор, как ожидается, мог бы произнести её аналогично первому варианту, без изменения интонационного контура. Однако говорящий, стратегически осознанно, акцентно выделяет ещё один компонент атрибутивной группы: *…насилие,* ***не опасное↘*** *для жизни и* ***здоровья,↗*** тем самым модифицируя его первоначальное значение в новое значение контраста. Семантика контраста, как описано выше, связана с мысленной процедурой выбора из множества вариантов, ассоциирующихся с интонационно выделенным. Атрибутивная конструкция процитирована прокурором из действующего Уголовного Кодекса. Таким образом, говорящий имплицирует, что в статье Уголовного Кодекса предусмотрены варианты взысканий для лиц, совершивших подобного рода правонарушения. Эти варианты различаются в зависимости от причинённого потерпевшему вреда: «не опасного для жизни и здоровья» и «опасного для жизни и здоровья». Выделяя акцентно один из вариантов, прокурор называет выбранную им альтернативу и указывает на предусмотренные Кодексом меры наказания.

Члены коллегиального суда, с большой вероятностью, знакомы и с действующим законодательством, и с имеющейся в статье Уголовного Кодекса альтернативой. Поэтому импликатура им будет понятна. С другой стороны, остальные присутствующие на заседании люди могут быть не знакомы ни с действующим законодательством, ни с мерами наказания за совершённое правонарушение, поэтому, возможно, не поймут смысла акцентного выделения компонента высказывания.

Во втором проанализированном примере эмфатическое выделение коммуникативно значимого компонента высказывания способствует его логическому выделению и в конечном итоге — к осмыслению его имплицитного содержания: применение властями противоправных мер.

Таким образом, по результатам анализа следует заключить, что контраст и эмфаза могут выступать не только как модификаторы коммуникативных значений, но также использоваться при передаче окказиональных имплицитных смыслов высказывания, тем самым расширяя число известных коммуникативных стратегий. Вопрос о том, вносит ли интонационное выделение компонента высказывания независимый вклад в интерпретацию высказывания при передаче имплицитных смыслов, пока остаётся гипотезой и требует дальнейших исследований.

Хотя, например, в английском языке имплицитный смысл высказывания может передаваться интонационно, при помощи падающей и далее восходящей интонации (fall-rise intonation) для выражения неуверенности, неопределённости [Ward G., Hirschberg J. 1985: 747–777].

Литература

1. *Бергельсон М. Б., Кибрик А. Е.* (1987) Прагматический принцип приоритета и его отражение в грамматике языка // Моделирование языковой деятельности в интеллектуальных системах. М., с. 52–63.
2. *Брызгунова Е. А.* Интонация // Русская грамматика. Т. 1. М.: Наука, 1980, с. 96–122.
3. *Всеволодова М. В., Ященко Т. А.* Коммуникативные языковые механизмы. Коммуникативная структура предложения // Причинно-следственные отношения в современном русском языке. М. Издательство ЛКИ, 2008, с. 6–26.
4. *Грайс Г. П.* Логика и речевое общение // Новое в зарубежной лингвистике. Вып. XVI. М.: Прогресс, 1985, с. 217–237.
5. *Кодзасов С. В.* Уровни, процессы и единицы в интонации // Проблемы фонетики. М.: Наука, 1999, с. 196–216.
6. *Кодзасов С. В.* Исследования в области русской просодии. М.: Языки славянских культур, 2009.
7. Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990.
8. *Паршин П. Б.* Сопоставительное выделение как коммуникативная категория. Автореферат диссертации на соискание учёной степени кандидата филологических наук. М., 1988.
9. *Фаустова Н. А.* Эмфаза и её интонационное выражение // Вопросы филологии. М., 2006, № 6, с. 255–259.
10. *Фаустова Н. А.* Интонационное выражение иллокутивных значений (на материале русского, французского и английского языков. Диссертация на соискание учёной степени кандидата филологических наук. М., 2009.
11. *Янко Т. Е.* Коммуникативные стратегии русской речи. М.: Языки славянской культуры, 2001.
12. *Янко Т. Е.* Материалы к речевому портрету лектора и диктора. Элементы русской просодии // Семантика и прагматика языковых единиц. Е. А. Земская, А. Н. Еремин (ред.). Калуга, 2004.
13. *Янко Т. Е.* Интонационные стратегии русской речи в сопоставительном аспекте. М.: Языки славянских культур, 2008.
14. *Galperin I. R.* Stylistics. Moscow: Higher School, 1977.
15. *Grice H. P.* Logic and conversation. Speech acts (Syntax and semantics. Vol. 3) ed. by Peter Cole & Jerry Morgan. New–York: Academic Press, 1975, с. 41–58.
16. *Ward G., Hirschberg J.* Implicating uncertainty: the Pragmatics of Fall-Rise Intonation // Language. Vol. 61, No 4, 1985, с. 747–777.

References

1. *Bergelson M. B., Kibrik A. E.* (1987) Pragmatic principle of priority and its reflection in language grammar [Pragmaticheskij printsip prioriteta i jego otrazhenije v grammatike jazyka]. Modelirovanije jazykovoj dejatelnosti v intellektualnyh systemah. Moscow, pp. 52–63.
2. *Bryzgunova E. A.* Intonation [Intonatsyja]. Russian grammar. Vol.1. Nauka, Moscow, 1980, pp. 96–122.
3. *Faustova N. A.* (2006) Emphasis and its intonation expressing [Emfaza i ejo intonatsionnoje vyrazhenije], Journal of Phylology [Voprosy filologii] No 6. Moscow, pp. 255–259.
4. *Faustova N. A.* Intonation expressing of illocutive meanings (on the examples from Russian, French, and English languages) [Intonatsyonnoje vyrazhenije illokutivnyh znachenij (na materiale russkogo, frantsuzskogo i anglijskogo jazykov)]. Ph.D. dissertation. Moscow, 2009.
5. *Galperin I. R.* (1977) Stylistics. Higher School, Moscow.
6. *Grice H. P.* (1975) Logic and conversation. Speech acts (Syntax and semantics. Vol. 3) ed. by Peter Cole & Jerry Morgan. New–York: Academic Press, pp. 41–58.
7. *Grice H. P.* (1985) Logic and conversation [Logika i rechevoje obzhchenije]. Novoje v zarubezhnoj lingvistike, XVI. Progress, Moscow, pp. 217–237.
8. *Janko T. E.* (2001) Russian speech communicative strategies [Kommunikativnyje strategii russkoj rechi]. Slavic cultures languages, Moscow.
9. *Janko T. E.* Materials for speech portrait of lecturer and announcer. Russian prosody elements [Materialy k rechevomu portretu lektora i dictora. Elementy russkoj prosodii. Ed. E. A. Zemskaja, A. N. Eremin]. Kaluga, 2004.
10. *Janko T. E.* (2008) Russian speech intonation strategies in comparative aspect [Intonatsionnyje strategii russkoj rechi v sopostavitelnom aspecte]. Slavic culture languages, Moscow.
11. *Kodzasov S. V.* Levels, processes, and units in intonation [Urovni, protsessy i jedinitsy v intonatsii]. Problemy fonetiki], III. Nauka, Moscow, 1999, pp. 196–216.
12. *Kodzasov S. V.* (2009) Researches in the field of Russian prosody [Issledovanija v oblasti russkoj prosodii]. Slavic cultures languages, Moscow.
13. *Linguistic* Encyclopedia. (1990) Soviet Encyclopedia, Moscow.
14. *Parshin P. B.* Comparative singling out as a communicative category. Synopsis of theses. Moscow, 1988.
15. *Vsevolodova M. V. Jashchenko T. A.* Communicative language mechanisms. Communicative structure of sentence [Kommunikativnyje jazykovyje mehanizmy. Kommunikativnaja structura predlozhenija]. Prichinno-sledstvennyje otnoshenija v sovremennom russkom jazyke. LKI Publ., Moscow, 2008, pp. 6–26.
16. *Ward G., Hirschberg J.* (1985) Implicating uncertainty: the Pragmatics of Fall-Rise Intonation. Language. Vol. 61, No 4, pp. 747–777.

О ПОЛИСЕМИИ «ПАРАМЕТР — БОЛЬШОЕ ЗНАЧЕНИЕ ПАРАМЕТРА»

Семенова С. Ю. (sonya_sem@mail.ru)

Институт научной информации по общественным наукам РАН, Российский государственный гуманитарный университет, Москва, Россия

Ключевые слова: регулярная полисемия, параметрическое существительное, семантическая деривация, метонимия, имена пространственных параметров, семантика количества

ON THE REGULAR AMBIGUITY: ‘PARAMETER VS. HIGH VALUE OF PARAMETER’

Semenova S. Yu. (sonya_sem@mail.ru)

Institute of Scientific Information on Social Sciences RAS, Russian State University for the Humanities, Moscow, Russia

The paper is concerned with regular ambiguity of the type ‘parameter — high value’ for Russian quantitative parametric nouns like *glubina* (depth), *davlenie* (pressure), etc. This type of ambiguity is shown to be heterogenous. For some dimensional nouns the ambiguity is caused by the metonymic shift from the meaning of a magnitude to the meaning of a spatial area where the value of this magnitude is high. For most parametric nouns this ambiguity is revealed in combinations with the verbs of surprise like *udivljat’sja* (‘be surprised’). The ambiguity has some analogs in non-quantitative parametric nouns, e.g. ‘parameter — Bon [the lexical function]’ for the non-quantitative parameter *kachestvo* (‘quality’).

Keywords: regular ambiguity of sense, parametric noun, dimensional parametric noun, quantitative meaning, semantic derivation, metonymy

1. Постановка задачи

Номинативная параметрическая количественная лексика (т. е. слова типа *глубина*, *скорость*, *давление*) постоянно, в течение, по крайней мере, нескольких десятилетий, привлекает к себе внимание исследователей. Интерес обусловлен целым рядом ее особенностей, в числе которых:

- ставшая уже традиционной значимость параметрической лексики, в первую очередь пространственной, для когнитивных исследований, в том числе для изучения закономерностей языкового кодирования визуальной информации;
- специфическая логико-семантическая природа параметрических имен, в частности, их близость к математическим функциям (способность принимать разные числовые значения при варьировании характеризуемого объекта, выступающего в роли аргумента функции);
- необходимость релевантного лексикографического представления параметрической лексики в прикладных системах, вытекающая из актуальности параметрической информации для задач извлечения знаний;
- развитая полисемия параметрической лексики (имеющая, в основном, регулярный характер), а также склонность такой лексики к дальнейшей семантической деривации.

Каждый из указанных пунктов является многомерным и подразумевает отдельное направление исследований; библиографические описания основных отечественных работ о параметрической лексике приведены в [8]. Ранее (напр., [4–7]), мы предлагали свое видение некоторых вопросов в рамках указанной проблематики. Так, в [7] систематизирован ряд типов полисемии — метонимических сдвигов и метафорических переносов, характерных для русских параметрических имен. Внимание уделено, с одной стороны, «дальней» метафоризации, связанной с переходом значения имени в более абстрактную сферу; условно говоря, от *глубины колодца* к *глубине просмотра гипертекста* и к *глубине интерпретации произведений Баха*; а с другой стороны, «ближним» метонимическим переходам, например, без выхода из пространственной сферы: от *высоты дерева* к *высоте треугольника* (т. е. от величины к фигуре).

В данной работе нас также будет интересовать полисемия русских параметрических существительных, ее «ближняя» сфера, а именно, регулярная полисемия «параметр — большое числовое значение параметра».

Этот вид полисемии был описан Ю. Д. Апресяном [1]. В названной работе это свойство отмечается прежде всего у имен линейных размеров *высота*, *глубина*, *толщина*: «(*прыгнул с высоты*), т. е. ‘с большой высоты’», *ушел в глубину* т. е. ‘на большую глубину’; *Толщины* [*дам города*N] *никак нельзя было приметить* (Гоголь), т. е. ‘большой толщины’» ([1]). Далее в цитируемой работе справедливо указывается, что данная «комбинация значений составляет отличительную черту более крупного лексикографического типа, в который входят и другие параметрические существительные; ср. *иду на скорости*, т. е. ‘на большой скорости’; *обрабатывать под давлением*, т. е. ‘под большим давлением’»

(там же). В цитируемой статье отражается и тот факт, что названным типом полисемии обладают не все параметрические имена. Например, утверждается, что он отсутствует у линейных размеров *длина* и *ширина*.

Возникает потребность изучить ситуацию подробнее, и в данной работе ставится задача выявить:

- типы контекстов, в которых проявляется второе из указанных значений;
- семантические процессы, приводящие к такой полисемии;
- круг параметрических имен, для которых эта полисемия возможна.

Материалом является, с одной стороны, перечень русских имен количественных параметров, выявленных на словнике «Толкового словаря русского языка» С. И. Ожегова (М., 1987), с другой стороны, ресурсы Интернет-пространства, которые просматриваются с помощью поисковой системы «Яндекс» с целью наблюдения за семантическим поведением параметрических имен. Используются также модельные примеры предложений и словосочетаний с параметрическими именами.

Перечень имен количественных параметров включает немногим более 200 слов (что согласуется с данными Ю. Д. Апресяна [2]). Внутренняя структура класса количественных параметрических имен, выделяемые в ней тематические и синтаксические группы, в том числе периферийные, отражены в [6 и др.].

2. Некоторые характерные контексты и семантические процессы

Итак, рассмотрим некоторые случаи полисемии указанного типа.

Для краткости обозначим названный тип полисемии через «P — G», где под символом P будем понимать лексему со значением «параметр», а под символом G (great) — лексему того же параметрического слова, со значением (или с семантическим компонентом) «большая величина».

Отметим, что в рассматриваемой паре лексических значений второе из них (значение G) является производным первого, параметрического. При этом целый ряд параметрических имен сами являются дериватами лексем, выражающих большое значение параметра. Таковы отадъективные имена, в основном исконные, образованные от прилагательных верхнего полюса измерительной шкалы: *высота* от *высокий*, *плотность* от *плотный* и т. п.; т. о. для отадъективных параметрических имен можно констатировать переходы от значения верхнего полюса к параметру и обратно. Актуальны и последующие метафорические и метонимические изменения [7]. Так что в целом полисемия параметрических имен представляет собой сложный комплекс, с неоднократными преобразованиями значений.

Перейдем к рассмотрению контекстов.

1. Обратимся снова к статье [1]. В двух первых упомянутых примерах (*прыгнул с высоты*; *ушел в глубину*) имена *высота* и *глубина* обозначают не собственно большие величины, а области пространства, расположенные на большом расстоянии от «нулевого уровня» вверх и вниз по вертикальной оси. Тем

самым, наблюдается метонимический сдвиг от обозначения величин к обозначению пространственных сущностей с координатами, представляющими собой большие величины. Он проявляется, например, в конструкциях с локативными предлогами: *монтажные работы на высоте / на глубине*; *«…когда усталая подлодка из глубины идет домой»*

(С. Гребенников, Н. Добронравов) и т. п. Данная метонимия отчетливо видна и в случае множественного числа у данных параметрических имен: *морские глубины*, *небесные высоты*. Как отмечалось в литературе (напр., [3, с. 210–212], см. также [5, 7]), метонимический сдвиг этого типа имеет место и для «дальних», т. е. метафорических, значений: *в глубине сада / веков / души*.

Данный вид метонимии справедлив только для параметрических имен *высота* (вместе с более старой формой *вышина*, не имеющей, впрочем, множественного числа) и *глубина*, которые выделяются этим свойством даже в весьма однородной группе имен линейных размеров.

(Правда, у других имен линейных размеров (*длина, ширина*, *толщина*) есть однокоренные аналоги со значением G, обслуживающие предметную сферу — *даль, ширь, толща* [у *высоты* и *глубины* тоже есть корреляты — *высь* и *глубь*]: *степные дали*; *Нам немало дано: ширь земли и равнина морская*. (К. Ваншенкин); *Настоящие хозяева водной толщи океанов и морей — мелкие ракообразные, головоногие моллюски и медузы*. При этом имя *толща* ближе к указанному типу, чем *даль* и *ширь*; оно, в отличие от двух последних, выражает идею концентрации в определенной среде и, тем самым, в определенном участке пространства. Метафорическим аналогом имени *толща* выступает имя *гуща* в переносном значении: *оказаться в гуще событий*).

2. Более широкий круг параметрических имен, в том числе основная масса пространственных параметров, склонны к типу G в сочетании с местоимением *весь*: *по всему периметру, по всей ширине, по всей высоте* и т. п.; ср. *Движение ограничено по всей ширине проезжей части*; *перетяжка скрепляющих (стяжных) колец по всей высоте трубы*. Встречаются примеры и с другими предлогами: *вдоль всей ширины*; *на всю длину волос я наношу льняное масло* и т. п., хотя доминирует предлог *по*. Представляется, что данный местоименный контекст обеспечивает, как импликатуру, тот факт, что значение параметра является немалым. Для пространственных параметров в этом контексте так же, как и в случае 1, имеет место метонимический переход от величины к пространственному ареалу (но переход, в отличие от 1, не связан с местонахождением некоторого объекта в некой «высоко»/ «глубоко» расположенной области; здесь, в силу семантики предлога, отмечается не идея концентрации, а идея распределения внутри достаточно вместительного пространства).

На то, что слово *весь*, изначально обозначающее логический квантор (вне количественной оценки), способно, все же, выражать такую оценку, обращал внимание Э. Сепир: «Следует отметить, что такие слова, как *all* и *whole*, нередко употребляются скорее во вторичном, оценочном смысле, нежели в их основном, абстрактном смысле.» [9, с. 398]. В самом деле, раз в тексте, с помощью местоимения *весь*, отражено, что определяемый объект квантуем, значит, в прототипе, он является достаточно большим для квантования.

Кванторное слово *весь* может употребляться и с некоторыми непространственными параметрическими именами, придавая им коннотацию G: *весь тираж / срок / период* и др.: *Весь тираж распродан*; *на весь срок президентских полномочий* и т. п. Эти примеры также свидетельствуют о нахождении имени параметра в предметной сфере: для имени *тираж* такая сфера изначальна («допараметрична»), а для имени *срок* она есть результат метонимического сдвига от параметрического значения.

3. Одним из контекстов, способных показывать, что употреблена лексема G, является контекст предикатов восприятия; см., например, в [1] пример из «Мертвых душ» Н.В. Гоголя с глаголом *приметить*. В самом деле, *примечают* не параметр как отвлеченную величину, а некоторое его значение, как правило, нестандартное, «заметное» у наблюдаемого объекта.

Особенно «сильным» контекстом, указывающим на нестандартность воспринимаемого, выступает контекст эмоциональной реакции — удивления, восхищения, испуга, страха и т. п.: *Меня ужасает не мой возраст, а возраст моих ровесников* [«Сводная энциклопедия афоризмов», dic.academic.ru/dic. nsf/aphorism/69/ВОЗРАСТ (от 20.04.2013)]. Контекст такого рода способен указывать на лексему G в том числе для имен *длина* и *ширина*, у которых (в отличие от *высоты* и *глубины*) не отмечается метонимического значения «область пространства с большой координатой по вертикальной оси» (см. выше случай 1); ср. примеры из Интернета: *Удивите длиной своих волос! Скидка… на наращивание волос в салоне красоты…*; *Он дышал со мной морским воздухом во Владивостоке, вкушал запахи тайги в Хабаровске, удивлялся ширине и мощи Енисея в Красноярске.*

Контекст удивления для имени параметра практически всегда указывает на значение величины (а не на нее саму), ведь для удивления нужна причина, и причина состоит в нестандартности значения.

В принципе, причиной удивления может быть и аномально малое значение параметра, ср.: *Платье? Блузка? Но длина удивила* (пример из Интернет-форума, в котором девушки обсуждают наряды; если говорящая интерпретирует предмет одежды как платье, то предметом ее удивления служит малая длина). Правда, подобные ситуации нетипичны; как правило, предметом удивления выступает количественное значение большего полюса.

Для отадъективных параметрических имен преимущественная интерпретация конструкции «удивился параметру» по типу G (при этом именно большое, а не малое значение параметра) вытекает не только из положения вещей в действительности, но и из «памяти значения» — из обозначения большего полюса мотивирующим прилагательным.

Отдельное место занимают боязнь *высоты* и *скорости*, поскольку это не только констатация эмоционального восприятия больших значений параметров, но и обозначение специфических ситуаций и психологических состояний; по сути дела, это термины (так, *боязнь высоты* означает возникновение неприятных, даже опасных, ощущений у человека, который находится на некоторой высоко расположенной площадке, скорее всего не огражденной, и смотрит вниз).

Примечательно, что встретился контекст предиката восприятия, в котором возможна двоякая интерпретация значения параметрического имени; это

иллюстративный пример из статьи параметрического имени *частота* в МАС: *Люди сидят близко друг к другу и могут слышать частоту дыхания соседа*. (Бахметьев, Преступление Мартына). Можно понимать по-разному, какую из лексем имени *частота* (как характеристики механического процесса), P или G, имел в виду автор текста; мнения информантов (в студенческой аудитории) разошлись. Действительно, ряд предикатов восприятия, в том числе *слышать*, способны употребляться с лексемой P, без обращения к собственно количественному значению. В первую очередь, это относится к предикатам целенаправленного восприятия (принадлежащих к категории агентивных действий); можно, например, *заметить время* (по часам), *пощупать пульс*, *уделить внимание длине изделия* (при его изготовлении или выборе) и др.

Но контексты эмоциональной реакции — «происшествия» (*удивляться*, *поражаться*, *восхищаться*) — практически всегда указывают на необычное значение как на причину эмоции.

4. На лексему G могут указывать и другие контексты, выражающие причину, например, предлоги с каузальным значением *по*, *из-за*: *трудовые пенсии по возрасту*; *дискриминация по возрасту* (ср. *Чем отличается дискриминация по возрасту… в современной России?... Почему в объявлениях о приеме на работу чаще всего пишут: до 35 лет?*); *пропустить занятия из-за температуры*; *Из-за ширины и длины коляски могут возникнуть проблемы с заездом в лифт* и т.п. Фактически, эти контексты выражают свернутую предикацию: *пропустить занятия из-за температуры* = пропустить занятия из-за того, что температура повысилась.

Встречается и такой тип контекста каузации, как *проблемы*, *сложности*: *проблемы с заболеваемостью* (значит, скорее всего, заболеваемость высокая); значение (или коннотация) G здесь обусловливается отрицательной аксиологической окрашенностью параметра (низкая заболеваемость в меньшей мере каузирует проблемы, нежели высокая).

5. Еще одна группа примеров выражает, так сказать, «чистое» значение G, не обусловленное ни пространственными метонимическими переходами, ни контекстами каузации: *литье под давлением*, *езда на скорости*, *провод под напряжением*, *держать / соблюдать дистанцию*, *давать нагрузку* и др. Эта группа конструкций, в том числе атрибутивных, представляется наиболее близкой к модельной структуре параметрической информации — к так называемой «параметрической триаде», включающей сведения о параметре, о характеризуемом по нему объекте и о числовом значении параметра ([4]); при этом числовое значение здесь эллиптически опускается. Строго говоря, опущенное числовое значение не обязательно является большим; оно является «достаточным», «подобающим» для обозначаемой ситуации. В определенной мере приведенные примеры относятся к фразеологии; они представляют собой полусвободные сочетания, в которых задействованы предложно-падежные конструкции и лексические функции. Набор таких сочетаний невелик, хотя, в принципе, он может пополняться.

К этому типу близки сентенции *Количество переходит в качество*; «*Большое видится на расстояньи*» (С. Есенин). Здесь тоже играет роль идея достаточно большого значения параметров *количество*, *расстояние*; лишь такое значение обеспечивает выполнение сформулированного онтологического закона.

3. О круге релевантных параметрических имен

При том что контексты, приведенные в пп. 1–2, возможны для ограниченного круга параметрических имен (главным образом, пространственных), и сочетаний типа 5 выявлено ограниченное количество, контексты каузации, указанные в пп. 3–4, допустимы для широкого круга имен параметров. Особенно «сильным» представляется модельный контекст удивления (*удивляться* + имя параметра в дативе). Этот контекст, который можно считать тестовым, теоретически возможен для основной массы параметрических имен. Хотя можно назвать и некоторые ограничения. Например, при подстановке в него возникает напряжение для имен *долгота* и *широта,* так как эти имена обозначают не величины как таковые, а координаты, и предметом удивления может быть не собственно их числовое значение, а расположение некоторого объекта на планете / карте. Странно воспринимается данный контекст и для имен математических операций *квадрат, произведение, частное* и т. п., поскольку они являются зависимыми от своих аргументов-величин (*квадрат расстояния*; *произведение массы на ускорение* и т. п.), и предметом эмоции, скорее, будет значение самих аргументов.

Таким образом, можно утверждать, что полисемия «P — G» возможна для основной массы имен количественных параметров (кроме, быть может (!), имен географических координат и абстрактных математических функций).

В прикладных лексикографических описаниях, на наш взгляд, следует отражать индивидуальную способность к метонимии у пространственных и нек. др. параметров, а также фиксировать лексически и синтаксически несвободные сочетания, выражающие смысл G (типа *литье под давлением*). А способность выражать этот смысл в контекстах каузации считать частью грамматики класса имен параметров.

4. По поводу неколичественных параметров

Любопытно, что аналог рассматриваемой регулярной полисемии проявляется и у некоторых неколичественных параметров. Напомним, что класс параметрических существительные делится на два подкласса: имен количественных параметров и неколичественных параметров (или признаков): *цвет*, *профессия*, *нравы* и др.; формальным критерием отнесения существительного к классу параметров, объединяющему эти два подкласса, может выступать способность употребляться в роли синтаксического объекта при дополнительно распределенных глаголах получения / передачи информации *вычислить*, *указать* и нек.др.: *вычислить скорость*, *указать цвет, определить нравы аборигенов / качество звучания аппаратуры* [6].

Аналог полисемии «P — G» отмечается у аксиологически окрашенных неколичественных параметров, и для них, не имеющих валентности на числовое значение, данная полисемия приобретает формы «P — Magn» (или «P — Bon», «P — AntiBon»); ср. контексты имен неколичественных параметров *нрав,*

качество, оценка [качественная, в отличие от количественной — числового параметра], *характер*: *товары со знаком качества*; *Качество гарантируем* (т. е. хорошее качество, Bon); *О времена! О нравы!* (т. е. ужасные нравы, AntiBon &Magn); *конь с норовом* (Magn); *Наконец-то тебя оценили!* (глагольный коррелят имени *оценка*, Bon); *Девушка с характером* (Magn).

Поведение неколичественных параметров, точный состав этого класса, взаимные переходы лексем P, Bon, Magn, в которых можно усмотреть, например, параллели с параметрическими vs прямыми диатезами предикатов принятия решений (*выбрать директора / выбрать Иванова директором*), подробно описанными Е. В. Падучевой, нуждается в отдельном изучении.

5. Некоторые итоги

Таким образом, можно констатировать неоднородность явлений и ситуаций, сопровождающихся полисемией «параметр — большое числовое значение параметра». В работе кратко охарактеризовано несколько явлений, регулярно представленных в классе параметрических имен и проявляющихся в характерных контекстах. Среди этих явлений: метонимические сдвиги «параметр — ареал с большой вертикальной координатой» и «параметр — ареал, значительный по соответствующему измерению», коннотация большой количественной оценки у квантора общности, свернутая предикация, маскирующая большое значение и проявляющаяся в контекстах каузауции (в т. ч. в контекстах эмоционального восприятия), эллиптическое опущение группы количественного значения в ряде фразеологизированных единиц.

И примеры полисемии «P — G» , приведенные в [1], относятся к разным явлениям: сочетания *прыгнул с высоты* и *ушел в глубину* раскрывают метонимический переход к ареалу с большим значением вертикальной координаты; гоголевское выражение *приметить толщину* [дам города N] (дословно: … *толщины никак нельзя было приметить*) иллюстрирует нестандартное (и притом большое) значение параметра, обнаруживаемое в процессе восприятия; фразеологизированные сочетания *иду на скорости*, *обрабатывать под давлением* обозначают параметрическую ситуацию с тройкой сущностей <объект, параметр, значение>, при эллиптическом опущении (большого) значения.

Представленные в работе случаи полисемии «P — G», видимо, не исчерпывают всех возможностей ее проявления в русском языке, и анализ может быть продолжен.

В целом полисемию «P — G» можно характеризовать как метонимию, поскольку переходы от параметра к его значению и к сущностям с большим значением (как и сама параметризация, т. е. возникновение параметрического значения у мотивирующих слов: *высота-* параметр есть дериват от S0 (*высокий*); *радиус-* параметр дериват от *радиуса-* отрезка [6, 7]) представляют собой преобразования по смежности.

Одной из сфер продолжения анализа может стать более подробное рассмотрение регулярных типов полисемии для неколичественных параметров.

Другим путем дальнейших исследований может стать сопоставительный анализ полисемии параметрической лексики, охватывающий и полисемию данного типа, и прочие явления семантической деривации в других языках. Например, полисемия «P — G» имеет место в английском языке и проявляется в некоторых контекстах, аналогичных русским: *in the depths of the lake* (метонимический сдвиг «параметр — ареал с большой вертикальной координатой»), *the economist would have been shocked by the size of bonuses* …(контекст эмоционального восприятия), *in the depth of my heart* (метонимический сдвиг переносного, метафорического, значения; В.В. Колесов показал, что данный тип деривации в странах христианской культуры сформировался под воздействием священных текстов [3]). Предикаты эмоционального восприятия в разной мере проявляют способность раскрывать смысл G у параметрического имени: предикату *to be shocked* эта способность присуща в гораздо большей мере, чем, например, предикатам *to be surprised* и *to be astonished*; фраза *They* [the children] *were astonished by the depth of the Siberian sea* (речь идет об озере Байкал), по-видимому, представляет собой русизм. В целом сопоставительный и типологический анализ видится перспективным и назревшим.

Литература

1. *Апресян Ю. Д.* Лексикографические портреты (на примере глагола быть) // НТИ. Сер. 2. — 1992, № 3. — С. 20–33.
2. *Апресян Ю. Д.* Фундаментальная классификация предикатов и системная лексикография // Грамматические категории: иерархии, связи, взаимодействие. Материалы международной научной конференции. СПб., 2003. — С. 7–21.
3. *Колесов В. В.* Философия русского слова. — СПб.: ЮНА, 2002.
4. *Семенова С. Ю.* Поиск параметрической информации в тексте: алгоритмический и лексикографический аспекты // Труды Международного семинара Диалог'96 по компьютерной лингвистике и ее приложениям. — М., 1996. — С.227–230.
5. *Семенова С. Ю.* О некоторых свойствах имен пространственных параметров // Логический анализ языка. Языки пространств / Отв. ред. Н. Д. Арутюнова, И. Б. Левонтина — М.: «Языки русской культуры», 2000. — С. 117–126.
6. *Семенова С. Ю.* Параметризация как метод познания и как языковой механизм // Логический анализ языка. Квантификативный аспект языка. — М.: Индрик, 2005. — С. 466–476.
7. *Семенова С. Ю.* Русское имя параметра: метафорические и метонимические процессы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). Т. 1. — М.: Изд-во РГГУ, 2012. — С. 568–577.
8. *Семенова С. Ю.* О спецкурсе по теоретическим и прикладным вопросам изучения русской параметрической лексики // Вестник РГГУ — 2013, № 8 (в печати).
9. *Сепир. Э.* Избранные труды по языкознанию и культурологии: Пер. с англ. / Общ. ред. и вступ. ст. А.Е. Кибрика. — 2-е изд. — М.: Прогресс, 2001.

References

1. *Apresian Ju. D.* (1992), Lexicographic portraits (A Case Study of the Verb byt' [to be]) [Leksikograficheskie portrety (na primere glagola byt')], Nauchno-tekhnicheskaia informatsiia, Seria 2 [Automatic documentation and mathematical linguistics], no.3, pp. 20–33.
2. *Apresian Ju. D.* (2003), A Fundamental Classification of Predicates and Systematic Lexicography [Fundamental'naja klassifikatsija predikatov i sistemnaja leksikografija], Grammaticheskie kategorii:ierarhii, svjazi, vzaimodeistvie. Materialy mezhdunarodnoi nauchnoi konferencii [Grammer Categories: Hierarchies, Links, Interaction. Proceedings of an international Conference], St. Petersburg, pp. 7–21.
3. *Kolesov V. V.* (2002) Philosophy of the Russian word [Filosofija russkogo slova], Juna Publ., St. Petersburg.
4. *Semenova S. Ju.* (1996), Search of the parametric information in the text: the algorithmic and lexicographic problems [Poisk parametrichskoj informatsii v tekste: algoritmicheskij i leksikograficheskij aspecty], Trudy Mezhdunarodnogo seminara Dialog'96 po komp'juternoj linguistike i ee prilozhenijam [Proceedings of the International summit Dialogue'96 on computational linguistics and its applications], Pushchino, Moscow, pp. 227–230.
5. *Semenova S. Ju.* (2000), Some features of the dimensional parametric nouns [O nekotoryh svojstvah imen prostranstvennyh parametrov], in Logicheskij analiz jazyka. Jazyki prostranstv [Logical analysis of language. Languages of spaces], Jazyki russkoj kul'tury, Moscow, pp. 117–126.
6. *Semenova S. Ju.* (2005), Parametrization as the method of investigations and the language process [Parametrizatsija kak metod poznanija i kak jazykovoi mehanizm], in Logicheskij analiz jazyka. Kvantitativnyi aspect yazyka [Logical analysis of language. The quantitative aspect of language], Indrik, Moscow, pp. 466–476.
7. *Semenova S. Ju.* (2012), On metaphor and metonymy of the Russian parametric noun [Russkoe imja parametra: metaforicheskie i metonimicheskie protsessy], Komp'juternaja linguistika i intellectual'nye tehnologii: Po materialam ezhegodnoi Mezhdunarodnoj konferencii «Dialog» [Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue"], Bekasovo, Moscow, no. 11 (18), Vol. 1, pp. 568–577.
8. *Semenova S. Ju.* (2013), On the special course "The Russian parametric words: theory and applications" [O spetskurse po teoreticheskim i prikladnym voprosam izuchenija parametricheskoi leksiki], Vestnik RGGU [Gerald of RSUH], no. 8 (in print).
9. *Sapir, Edward* (2001), Selected works in language and culture [Izbrannye trudy po jazykoznaniju i kul'turologii], Progress, Moscow.

ДИСКУРСИВНОЕ МАРКИРОВАНИЕ НЕТРИВИАЛЬНОГО ЛЕКСИЧЕСКОГО ВЫБОРА

Шилихина К. М. (shilikhina@gmail.com)

Воронежский государственный университет,
Воронеж, Россия

Ключевые слова: дискурсивная единица, вводная конструкция, нетривиальный лексический выбор, номинация, категоризация

DISCURSIVE MARKERS OF NON-TRIVIAL LEXICAL CHOICE

Shilikhina K. M. (shilikhina@gmail.com)

Voronezh State University, Voronezh, Russia

The paper discusses discursive functions of three Russian constructions: "*esli mozhno tak skazat*" [if I can say so], "*esli mozhno tak vyrazit'sya*" [if I can express it this way] and "*s pozvolenija skazat*" [if I'm allowed to call it X]. These constructions play the role of metalinguistic tools that structure the information flow. Functioning as parenthesis, these constructions mark the speaker's attitude towards his/her own speech actions and attract the attention of the addressee to the non-trivial form of expression. These non-trivial forms include unexpected lexical choice, metaphoric nomination and breaking the norms of word formation. By using "*esli mozhno tak skazat*" or "*esli mozhno tak vyrazit'sya*" the speaker can also introduce the process of searching for the most optimal way of expressing an idea.

Non-trivial lexical choice or ungrammatical forms introduced by the constructions "*esli mozhno tak skazat*" ('if I may say so') or "*esli mozhno tak vyrazit'sya*" ('if I may express myself so') are signals of the speaker's stance towards the object or the situation. Another possible goal of unusual verbal behavior is switching from bona fide to non-bona fide mode of communication. Along with the negative evaluation this switch can lead to the ironic interpretation of the utterance. The third construction — "*s pozvolenija skazat*" ('if I am allowed to say so') — functions as a signal of linguistic categorization process. By using it the speaker shows that the object cannot belong to a particular category due to the lack of necessary properties.

Key words: discursive unit, parenthesis, non-trivial lexical choice, nomination, categorization

1. Дискурсивные единицы

В последнее время популярным объектом лингвистического анализа стали дискурсивные слова — единицы, которые ранее считались второстепенными, несамостоятельными и, следовательно, не заслуживающими детального описания. Оказалось, что эти «второстепенные» элементы обеспечивают взаимодействие между участниками коммуникации, высказыванием/текстом и окружающим миром: в частности, дискурсивные слова помогают говорящему структурировать информационный поток и регулировать процесс понимания высказывания/текста (Дискурсивные слова 2003; Кобозева, Захаров 2004; Fraser 2006; Кобозева, Латышева 2008; Борисова 2012).

Основные критерии отграничения случаев метадискурсивного использования языка от «информационного» могут быть сведены к следующим:

- языковые средства, функционирующие на метауровне, не входят в пропозицию высказывания, но показывают, каким образом должно быть интерпретировано это высказывание (логико-семантический критерий);
- дискурсивные слова грамматически не связаны с остальными членами предложения (синтаксический критерий).

Функция структурирования высказывания/текста может выполняться не только отдельными словами, но и словосочетаниями. С точки зрения их дискурсивного использования эти конструкции соответствуют обозначенным выше критериям метадискурсивности. Для обозначения конструкций, выполняющих в высказывании метаязыковую функцию, в данной работе используется термин *дискурсивные единицы*.

В статье рассматриваются дискурсивные функции трех русских конструкций: синонимичных *если можно так сказать*, *если можно так выразиться* и близкой по значению *с позволения сказать*. Материалом для исследования стали данные Национального корпуса русского языка. На момент написания работы количество вхождений для каждой конструкции составило: *если можно так сказать* — 158 вхождений; *если можно так выразиться* — 629 вхождений; *с позволения сказать* — 352 вхождения.

При описании функциональных возможностей этих конструкций грамматики русского языка включают их в группу вводных словосочетаний, с помощью которых говорящий характеризует отношение к способу выражения мысли [Ляпон 1998, Розенталь 2000]. Анализируя конструкцию *так сказать*, Т. В. Шмелева отмечает, что с ее помощью говорящий подчеркивает нестандартность языкового оформления высказывания [Шмелева 1987]. Однако далеко не всегда нетривиальный лексический выбор сопровождается маркером. Очевидно, помимо собственно акцента на необычности маркеры выполняют и некоторые другие функции. Поэтому цель данной работы — не только описать дискурсивную семантику этих конструкций, но и показать, какие дополнительные коммуникативные задачи может решать говорящий, обращая внимание адресата на необычность лексического выбора.

2. *Если можно так сказать, если можно так выразиться* и *с позволения сказать*: сфера действия и дискурсивные функции

Семантика глаголов *сказать* и *выразиться* — *передать свою мысль теми или иными словами* — определяет сферу действия интересующих нас конструкций. Для дискурсивного употребления конструкций *если можно так сказать*, *если можно так выразиться* и *с позволения сказать* характерно, что их сферой действия является не пропозиция высказывания целиком, а лишь ближайшее окружение слева и/или справа. Таким образом, говорящий делает акцент на сделанном им выборе определенных лексических средств. Так, в примере 1 «сферой действия» является словосочетание *фальсификационный ресурс власти*, внутрь которого «вклинивается» дискурсивная единица; в примере 2 сфера действия — только ближайший правый контекст, а именно — номинация *адресатка*. В результате в первом случае говорящий делает акцент на нетривиальности коллокации *фальсификационный ресурс власти*, а во втором подчеркивается необычность морфологической формы слова *адресатка*:

(1) *Фальсификационный же,* ***если можно так выразиться,*** *ресурс власти не превышает 4–5 % от числа всех избирателей, а потому этот барьер в принципе преодолим.* (НКРЯ)

(2) *Так вот Милуша / автор / ее адресат / или /* ***если можно так выразиться*** */ адресатка / значит / мы тоже ее имя скоро увидим / в той части / это нечто вроде приписки / но само по себе довольно выразительной.* (НКРЯ)

Анализ семантических и грамматических свойств словоформ, попадающих в «сферу действия» этих трех маркеров, позволяет дать описание основной функции интересующих нас конструкций. В общем виде ее можно определить следующим образом: вводя в высказывание один из маркеров, говорящий обращает внимание адресата на некоторую «нетривиальность» способа языкового оформления мысли. «Нетривиальность» заключается в том, что лексические средства, с точки зрения самого говорящего, являются неожиданными для адресата либо не соответствуют представлениям говорящего о языковой норме, и поэтому заслуживают привлечения внимания для их правильной интерпретации. Т. В. Шмелева считает, что с помощью маркера *так сказать* говорящий не только акцентирует внимание на необычной форме выражения мысли, но и извиняется за допущенную «языковую вольность» [Шмелева 1987]. Однако, по нашему мнению, вряд ли здесь можно говорить об извинении; скорее, говорящий обращает внимание адресата на другие имплицитные смыслы, скрытые за нетривиальностью формы.

Важность информации, передаваемой маркерами, можно показать, «изъяв» вводные конструкции из высказывания. При сохранении целостности грамматической структуры утраченными оказываются не только акцент на способе номинации, но и отношение говорящего к объекту речи:

(3) *Это было нечто вроде боевого сбора — ребята соревновались на дальность бросания гранаты. И не просто как легкоатлетического снаряда, коим некогда швырялись поборники ГТО. Это были вполне нормальные гранаты — только учебные Ф-1. Так что их бросали по всем правилам* ***солдатского, с позволения сказать, искусства.*** (НКРЯ)

(3а) *Это было нечто вроде боевого сбора — ребята соревновались на дальность бросания гранаты. И не просто как легкоатлетического снаряда, коим некогда швырялись поборники ГТО. Это были вполне нормальные гранаты — только учебные Ф-1. Так что их бросали по всем правилам* ***солдатского искусства.***

Сфера действия конструкции «с позволения сказать» в примере 3 — коллокация *солдатское искусство*. С помощью маркера говорящий указывает на необычность этого словосочетания. Его можно считать нетривиальным, поскольку в НКРЯ оно не встречается ни разу, а при попытке найти его в других источниках поисковик Google выдает результаты только на словосочетание *советское искусство*, что на наш взгляд является еще одним показателем необычности именной группы *солдатское искусство*. Узуальным для носителей русского языка является словосочетание *военное искусство*. В НКРЯ оно встречается 294 раза:

(4) *Точно так же и* ***военное искусство*** *представляет собой высшую степень практического умения и мастерства военачальника, проявляемую при подготовке и ведении вооруженной борьбы.* (НКРЯ)

В примере 3 конструкция *с позволения сказать* находится внутри словосочетания. Таким образом, сфера его действия — это ближайшее окружение слева и справа. Поскольку в сферу действия дискурсивной единицы попадает словоформа в постпозиции, мы можем предположить, что маркер нужен говорящему, чтобы предупредить адресата о нетривиальном выборе лексемы. Однако в корпусе немало случаев, когда сферой действия дискурсивной единицы является только левый контекст, т. е. говорящий как бы предупреждает адресата об уже сделанном лексическом выборе:

(5) *Большинство этих временных союзов распалось; иные — как Алов и Наумов — стали одним режиссером,* ***если можно так сказать.*** (НКРЯ)

(6) *Идея Бога растиражирована,* ***если можно так выразиться.*** (НКРЯ)

(7) *Но самое противное то, что Ломакин за годы и годы приученный к камере, все время СЕБЯ чувствовал мишенью — будто за ним неотступно и неусыпно наблюдал объектив, даже… субъектив,* ***с позволения сказать.*** (НКРЯ)

Очевидно, в таких контекстах маркеры выполняют не предупреждающую функцию, а служат сигналом некоторого «лингвистического эксперимента» со стороны говорящего. Далее мы попытаемся ответить на два вопроса: что именно делает избранный способ именования объекта или действия необычным и к каким изменениям в семантике высказывания приводит появление одного из маркеров рядом с нетривиальной номинацией?

3. Нетривиальный лексический выбор: виды «лингвистических экспериментов»

3.1. Нетривиальная номинация/сочетаемость

Необычная номинация объекта, о котором идет речь, а также нетривиальная лексическая сочетаемость, позволяющая приписать объекту признак, в норме для него не характерный — это один из вариантов нетривиальной номинации. В письменном тексте на такую нетривиальность может указывать не только дискурсивная единица, но и кавычки:

(8) *То есть удельный,* ***если можно так сказать****, уровень «мракобесия» в США/ЕС близок к российскому.* (НКРЯ)

(9) *Словом,* ***если можно так выразиться****, перед нами — «махровый» социум.* (НКРЯ)

В примерах 8 и 9 говорящий подвергает некоторому сомнению и потому оговаривает саму возможность употребления лексики с ярко выраженной негативной коннотацией. Выбор такого способа «упаковки мысли» можно объяснить потребностью выразить оценочное отношение к объекту речи.

Вариантом нетривиальной номинации можно считать использование метафор. Дополнительный акцент на метафорической номинации свидетельствует о том, что говорящий сам воспринимает собственный лексический выбор как неожиданный и считает необходимым обговорить право на такое употребление:

(10) *Школа управляется,* ***если можно так сказать «семьёй»****, все члены которой либо дети директора школы, либо дети его друзей.* (НКРЯ)

(11) *Вот и сегодня возникают то тут, то там, подпевая Кремлю, башенки. Кремль,* ***если можно так выразиться, соло****. Центробанк и «Балчуг» — back-вокал.* (НКРЯ)

3.2. Нарушение словообразовательных и словоизменительных норм

Еще один тип вариант «лингвистического эксперимента» связан с имеющимся у говорящего знанием языковых норм и коммуникативным опытом. Говорящий может это маркировать сознательное нарушение этих норм:

(12) *У Сережи Светлакова шутки более «быдловатые», **если можно так выразиться**.* (НКРЯ)

(13) *Мы решили прикинуться бомжами, проникнуть в самое ядро этой, **с позволения сказать**, движухи и на собственной шкуре испытать все радости и невзгоды жизни рядового привокзального маргинала.* (НКРЯ)

(14) *На мой взгляд / это / пожалуй / одна из самых сложных музык / **если можно так сказать**.* (НКРЯ)

Маркеры в примерах 12–14 служат сигналом того, что говорящий осознает факт нарушения словообразовательной/словоизменительной нормы, но, тем не менее, считает «неправильный» вариант наиболее экономным и удобным способом выражения мысли в данном контексте.

3.3. Уточнение мысли, выбор слова в процессе порождения текста

В нарративном тексте маркер может акцентировать внимание адресата не столько на необычности лексического оформления высказывания, сколько на процессе выбора способа номинации из нескольких возможных:

(15) *Накануне Игр я понимала, у меня уже нет сил, я машинально исполняю программу, делаю что полагается, но **свежесть, или, если можно так сказать, одухотворенность** — исчезли.* (НКРЯ)

Какие коммуникативные задачи решает говорящий с помощью нетривиального лексического выбора, помеченного дискурсивным маркером? Попробуем интерпретировать нетривиальный лексический выбор с точки зрения наличия в дискурсивной семантике высказывания некоторого дополнительного смысла.

4. Цели «лингвистического эксперимента»

Нетривиальный лексический выбор может быть интерпретирован в зависимости от того, какой маркер использует говорящий для привлечения

внимания адресата. Конструкции *если можно так сказать* и *если можно так выразиться* связаны с выражением оценочного отношения к объекту речи и переходом из серьезного в иронический модус коммуникации. Маркер *с позволения сказать* ведет себя особым образом: указывает на несоответствие объекта некоторой категории. Отрицательная оценка объекта становится логическим продолжением этого несоответствия. Проиллюстрируем сказанное на примерах.

4.1. Выражение оценки

Маркеры *если можно так сказать* и *если можно так выразиться* могут использоваться говорящим в тех случаях, когда выбранный способ именования объекта или его свойства одновременно выражает оценочное отношение, как правило, негативное:

(16) *В медицинских школах не затрагивались эти вопросы, и отношение к человеку, который тяжело болен или в опасности смерти, было всегда или очень формальное, или,* ***если можно так выразиться****, пугливое.* (НКРЯ)

Отношение говорящего к тому, о чем идет речь, выражается через прилагательное *пугливое*. Это прилагательное, обладающее отрицательной коннотацией, входит в сферу действия конструкции *если можно так выразиться*. Маркер указывает на то, что негативная оценка ситуации основана на субъективном восприятии ее говорящим и служит своего рода средством смягчения этой оценки.

4.2. Переход в иронический модус коммуникации

Маркер может быть сигналом притворного сомнения говорящего в адекватности выбранной формы. В таком случае выбранный способ номинации может быть интерпретирован и как показатель смены модуса коммуникации с серьезного на иронический. Эта смена, в свою очередь, означает и изменение правил интерпретации сказанного: адресат не должен воспринимать высказывание серьезно:

(17) *Объем груди составляет 167,6 см, ширина плеч — 83,8, объем талии* (***если можно так выразиться***) *— 129,5 см.* (НКРЯ)

(18) [Т. Г. Винокур, жен] *Ну разве это дело / что в современном обществе* [усмехается] *такие /* ***с позволения сказать*** */ нормы общения щас берут верх / как / например / «А ну / давай / бабка* (ну в лучшем случае бабуля) *проходи / чего застряла!* (НКРЯ)

В приведенных примерах ирония возникает как результат несоответствия номинации нашим знаниям об окружающем мире. Так, в примере 18 наше знание о том, что талия — это наиболее узкая часть тела между грудью и тазом, вступает в противоречие с приведенными данными. В примере 19 показателем иронии является не только усмешка говорящего, но и категоризация явно грубого высказывания как нормативного.

4.3. Указание на неадекватную категоризацию

Дискурсивная семантика конструкции *с позволения сказать* связана с категоризацией объектов окружающего мира. Употребляя ее, говорящий отрицает возможность отнесения объекта к некоторой категории, поскольку объект не соответствует представлению говорящего о том, каким этот объект должен быть:

(19) *Если мы зададимся созданием только жесткой структуры «государственный федеральный оператор — государственный региональный оператор», мы рискуем построить,* ***с позволения сказать****, игрушечную железную дорогу вместо федеральной сети железных дорог.* (НКРЯ)

Заметим, что маркер *с позволения сказать* отличается от двух других и по синтаксическому поведению (в подавляющем большинстве случаев он предшествует своей «сфере действия»). Кроме того, в сферу действия этой конструкции практически не попадают метафорические номинации. Конструкция *с позволения сказать* акцентирует внимание не столько на словесном оформлении, сколько на несоответствии объекта категории. Иными словами, это не вполне «лингвистический эксперимент», скорее, это акцент на соответствии некоторого объекта нашим представлениям о категории, к которой мы его относим при помощи номинации.

5. Заключение

Анализируемые конструкции — индикаторы субъективной оценки говорящим собственных речевых действий. Эти действия маркируются, поскольку, по мнению говорящего, являются нетривиальными и требуют дополнительного согласования с контекстом (как в случае с метафорической номинацией) или с адресатом (как в случае нарушения языковой нормы).

При том, что конструкции *если можно так сказать* и *если можно так выразиться* являются практически синонимичными, конструкция *с позволения сказать* функционирует как указание на несоответствие объекта речи той категории, к которой его причисляет говорящий. Именно факт категоризации объекта отличает конструкцию *с позволения сказать* от двух других дискурсивных единиц.

Литература

1. *Борисова Е. Г.* (2012). Роль дискурсивных слов в управлении пониманием текста // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — Вып. 11 (18). — М.: Изд-во РГГУ, 2012. — С. 93–102.
2. *Дискурсивные слова* (2003). — Дискурсивные слова русского языка: контекстное варьирование и семантическое единство / Сост. К. Киселева, Д. Пайар. — М.: Азбуковник, 2003. — 207 с.
3. *Кобозева И. М., Захаров Л. М.* (2004). Для чего нужен звучащий словарь дискурсивных слов русского языка // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог 2004. — М.: Наука.
4. *Кобозева И. М., Латышева Н. С.* (2008). Дискурсивные единицы собственно и фактически как операторы коррекции представления адресата // Язык средств массовой информации как объект междисциплинарного исследования. Материалы II Международной научной конференции 14–16 февраля 2008 г. / Сост. М. Н. Володина. М: МАКС Пресс. — С. 183–187.
5. *Ляпон М. В.* (1998). Вставная конструкция // Русский язык: Энциклопедия. Изд.2, перераб. и доп. М.: Большая Российская энциклопедия: Дрофа. — С. 97.
6. *Розенталь Д. Э.* (2000). Справочник по правописанию и литературной правке. М.: Айрис-пресс.
7. *Шмелева Т. В.* (1987). «Так сказать» и «как говорится» // Служебные слова. Новосибирск: НГУ. — С. 125–132.
8. *Fraser B.* (2006). Towards a Theory of Discourse Markers // Approaches to Discourse Particles / Ed. by K. Fischer. Amsterdam/London/New York: Elsevier. — P. 189–204.

References

1. *Borisova E. G.* Discourse Markers Used for Governing Understanding of texts [Rol' diskursivnyh slov v upravlenii ponimaniem teksta]. Kompjuternaja lingvistika i intellektualnye tehnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii "Dialog 2012". [*Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006"*]. Bekasovo, 2006, pp. 93–102.
2. *Fraser B.* (2006). Towards a Theory of Discourse Markers, in K. Fischer (ed.). *Approaches to Discourse Particles*. Elsevier, Amsterdam/London/New York, pp. 189–204.
3. *Kiseleva K., Paillard D. (Eds.). (2003). Diskursivnyje slova [Discursive Words]. Mosow, Azbukovnik.*
4. *Kobozeva I. M., Latysheva N. S.* (2008). Discursive Units *Sobstvenno* i *Fakticheski* as Operators of Correction [Diskursivnye edinitsy sobstvenno i fakticheski kak operatory korrektsii predstavlenij adresata]. Jazyk Sredstv Massovoj Informatsii kak ob"ekt mezhdistsiplinarnogo issledovanija. Materialy II Mezhdunarodnoj Nauchnoj Konferentsii [Mass Media Language as an Object of Cross-Disciplinary Research. Proc. of the 2nd International Conference]. Moscow, 2008, pp. 183–187.
5. *Kobozeva I. M., Zaharov L. M.* (2004). Why Is a Sounding Dictionary of Russian Discursive Words Necessary [Dlya chego nuzhen slovar' diskursivnyh slov russkogo jazyka]. Kompjuternaja lingvistika i intellektualnye tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2004" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2004"]. Moscow.
6. *Ljapon M. V.* (1998). Parenthetical Construction [Vstavnaja konstruktsija]. Russkij Jazyk: Entsiklopedija. Bolshaja Rossijskaja Entsiklopedija, Drofa, Moscow. — P. 97.
7. *Rozental D. E.* (2000). [Spravochnik po pravopicaniju i literaturnoj pravke]. Iris-Press, Moskow.
8. *Shmeleva T. V.* (1987). "*Tak skazat'*" and "*kak govoritsja*" ["*Tak skazat'*" i "*kak govoritsja*"], Grammatical Words [Sluzhebnye slova], Novosibirsk State University, Novosibirsk, pp. 125–132.

PROCESSING OF QUANTITATIVE EXPRESSIONS WITH UNITS OF MEASUREMENT IN SCIENTIFIC TEXTS AS APPLIED TO BELARUSIAN AND RUSSIAN TEXT-TO-SPEECH SYNTHESIS

Skopinava A. M. (skelena777@gmail.com),
Hetsevich Yu. S. (Yury.Hetsevich@gmail.com),
Lobanov B. M. (Lobanov@newman.bas-net.by)

United Institute of Informatics Problems of the NAS of Belarus, Minsk, Belarus

The article discusses problems of identification, analysis, classification (according to the International System of Units and separately according to word formation peculiarities), and processing of quantitative expressions (QE) with measurement units (MUs) as applied to text-to-speech synthesis by means of the linguistic processor NooJ[1] and specially collected legal, scientific and technical text corpora for the Belarusian and Russian languages. In addition to a general description of algorithms and resources for finding QE in Belarusian and Russian texts, the paper gives an overview of QE with MUs with regard to how their components could be written, i.e. digital descriptors, and MUs proper (five different types). It is shown that QE with MUs can get the correct intonation marking only after they are properly generated, i. e. expanded into orthographical words.

Key words: text-to-speech synthesis, NooJ, units of measurement, quantitative expressions, finite-state automata, generation of an orthographical text, identification, processing, intonation marking, Belarusian, Russian

Introduction

After the Belarusian and Russian NooJ modules [3] were obtained, it became possible to check and update experimental solutions to different linguistic tasks in application to text-to-speech synthesis [1, 2]. Synthesizers which use orthographic texts cope well with voicing orthographic words [7], but abbreviations, acronyms, numbers, symbols, etc. demand preprocessing into real words before they can be voiced.

The main purpose of this article is to describe approaches to identification and transformation of quantitative expressions (QE) with measurement units (MUs) into correct orthographic words in hand-crafted scientific, technical and legal text corpora for Belarusian and Russian; and to prove its importance for correct intonational marking of texts.

[1] http://www.nooj4nlp.net/pages/nooj.html

To give an example, Belarusian sequences like *123 мА* '123 mA' and *120 мА* '120 mA' have to be transformed by the synthesizers into sequences of words with intricate agreement: resp. *сто дваццаць тры міліампер**ы*** and *сто дваццаць міліампер*, because Belarusian (and Russian) numerals are declinable and can influence subsequent words (in our case measurement units), unlike English, where, e. g., a preposition before a numeral does not change anything in the voicing of MUs. For the present we deal with generating QE in the Nominative.

When dealing with QE with MUs, many difficulties arise. First, they are conditioned by a great variety of numeral quantifiers and names of units, both in writing and formation. Creating rules of complex expressions localization for all cases is practically impossible (that is exactly the reason why regular expressions are not the best way to obtain localization rules). In order to simplify this process, it is extremely important to use tools that allow users to easily modify previously-developed rules and add new ones. The international program NooJ is one such tool. It allows implementing sophisticated algorithms of searching for compound text fragments in Belarusian and Russian in the form of visual executable graphs.

Second, an expression with a MU is difficult to recognize and analyze (note a considerable number of digits, words with quantitative meaning with all their possible paradigmatic forms, names of metrological system units) without thoroughly prepared linguistic resources, i.e., dictionaries with all possible word forms, abbreviations, and rules for building derivative forms of measurement units. This is necessary, e. g., for proper treatment of expressions with units of length, written in various ways: *1 м (1 m), 31 метр (31 meters), 25 метраў (25 meters), 44 метры (44 meters)* [4].

Third, QE with MUs are language-dependent: in English *meter* and *mile* are abbreviated as *m*, while in Belarusian and Russian as м; even within largely similar Russian and Belarusian, names of measurement units differ in spelling — *гадзіна*, *час* 'hour'. Therefore, it is essential to make accurate provisions for each language.

Significant results have been achieved by European researchers and developers of the Quantalyze semantic annotation and search service[2], and Numeric Property Searching service in Derwent World Patents Index on STN[3]. However, language orientation is the reason why theoretical or practical results cannot be fully reusable for Belarusian or Russian. We view QE with MUs as combinations where each component requires a specific approach for successful identification.

Searching for and classifying QE with MUs according to the SI

In order to construct and test algorithms, four text corpora were formed for two domains: scientific, technical and legal (two for each language) (Fig. 1) [4]. According to the main graph (Fig. 2) of the obtained algorithms (for Belarusian and Russian they differ in some language-dependent subgraphs), any text fragment is initially checked in the 1st subgraph (Numeral Quantifier) if it has a compound numerical descriptor (Fig. 3).

---

[2] https://www.quantalyze.com/en/

[3] http://www.stn-international.com/numeric_property_searching.html

File Name	
Раздзел 21. Рух гужавых транспартных сродкаў, коннікаў і прагон жывёлы	186.2.1. 12 метраў для аўтамабіля, тралейбуса, прычэпа;
Раздзел 22. Карыстанне знешнімі светлавымі прыборамі і гукавымі сігналамі	186.2.2. 13,5 метра для аўтобуса з двума восямі, 15 метраў
Раздзел 23. Перавозка пасажыраў	для аўтобуса з больш чым двума восямі;
Раздзел 24. Перавозка грузаў	186.2.3. 18,75 метра для сучлененага аўтобуса, сучлененага
Раздзел 25. Буксіроўка механічных транспартных сродкаў	тралейбуса;
Раздзел 26. Асноўныя палажэнні аб допуску транспартных сродкаў да ўдзелу	
Раздзел 27. Абавязкі службовых і іншых асоб па забеспячэнні бяспекі дарожн	

a)

File Name	
Глава 10. Расположение транспортных средств на проезжей части дороги	89.2. автобусам и мотоциклам — не более 90 км/ч;
Глава 11. Скорость движения транспортных средств	89.3. автобусам, легковым и грузовым автомобилям при их
Глава 12. Обгон, встречный разъезд	движении с прицепом, грузовым автомобилям с технически
Глава 13. Проезд перекрестков	допустимой общей массой более 3,5 тонны на автомагистралях —
Глава 14. Пешеходные переходы и остановочные пункты маршрутных тран	не более 90 км/ч, на остальных дорогах — не более 70 км/ч;
Глава 15. Преимущество маршрутных транспортных средств	
Глава 16. Железнодорожные переезды	

b)

File Name	
Chapter 21. Traffic of animal-drawn vehicles, horseback riders and guiding	186.2.1. 12 meters for a motor vehicle, trolleybus, trailer;
Chapter 22. Use of external luminous and audible devices of vehicles	186.2.2. 13.5 meters for a bus with two axles, 15 meters
Chapter 23. Carriage of passengers	for a bus with more than two axles;
Chapter 24. Carriage of goods	186.2.3. 18.75 meters for an articulated bus, articulated
Chapter 25. Towing of power-driven vehicles	trolleybus;
Chapter 26. General provisions about admission of vehicles to participate i	

c)

Fig. 1. Fragments of legal text corpora for (a) Belarusian, (b) Russian, and (c) translated into English

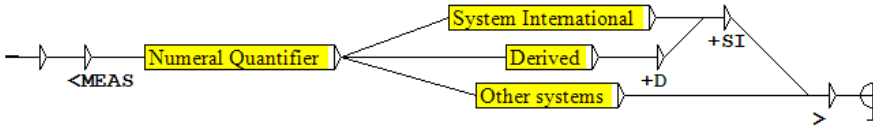


Fig. 2. The main graph of the algorithm for identification of QE with MUs

It should be noted that this subgraph works out not only for prime, decimal and fractional numbers in various forms of writing, but also for compound numerical combinations with exponential parts and periods. Some results of its work can be observed in the form of a concordance (Fig. 4). It should be emphasized that this subgraph is language-independent (Fig. 4c).

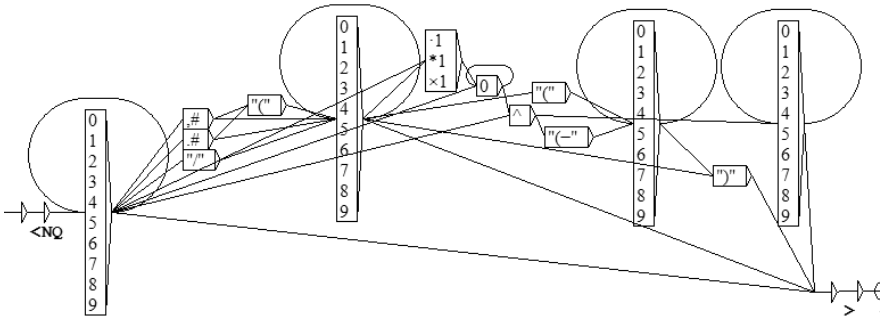


Fig. 3. The subgraph for identification of numbers and compound numerical combinations

Before	Seq.	After
электратэхнічнай камісіяй) IEC	60027	ужываецца пазначэнне Mbit
проста Mb). 1 мегабіт =	1000^2	біт = 10^6 біт = 1000000 біт
Mb). 1 мегабіт = 1000^2 біт =	10^6	біт = 1000000 біт. Дзесятков
Напрыклад: 1/6 = 0,166666... =	0,1(6)	; 1/7 = 0,1428571428... = 0,(14
= 0,1(6); 1/7 = 0,1428571428... =	0,(142857)	.

a)

Before	Seq.	After
автомагистралях - не более	110	км/ч, на
двумя осями; · -	18,75	метра для сочлененного
в среднем составляет	5·10^(−5)	Тл, а на
на экваторе (широта 0°) —	3,1·10^(−5)	Тл. 5. Ом — единица
бомбардировке Хиросимы: около	6·10^13	Дж. Энергия фотона
красного видимого света:	2,61·10^(−19)	Дж.

b)

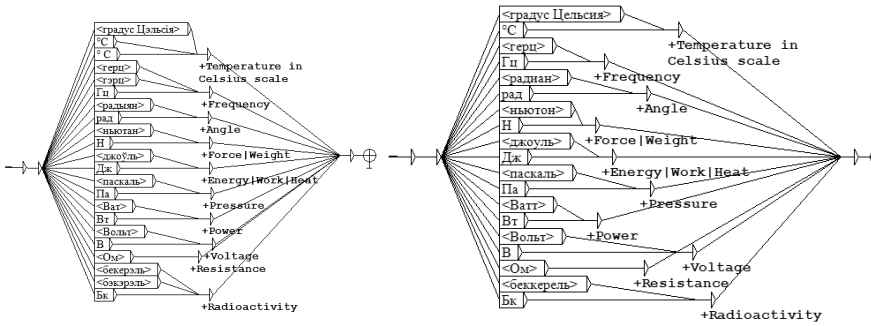
Before	Seq.	After
is equal to	6.24150974×10^18	eV (electronvolts). 1 joule
is equal to	2.3901×10^(−4)	kcal (thermochemical kiloc
defined as exactly	0.0254	m, and the
defined as exactly	453.59237	g. Also a
are equivalent to	1/100	. An integer such

c)

Fig. 4. Results of identifying complex numerical expressions in (a) Belarusian, (b) Russian, and (c) English texts

After the first subgraph has been processed, the algorithm proceeds to other subgraphs, which are connected to its output by means of respective transition lines. The subgraph *System International* identifies units according to the SI, e. g., *кілаграм* 'kilogram'; the subgraph *Derived* — SI derivatives (Fig. 5), such as *герц* 'hertz'; the subgraph *Other systems* — frequently used, but non-systemic units, such as *час* 'hour'. If any of the three subgraphs works out, the sequence of respective transition lines on the way to the main graph's output is indicated by markers. Let us draw up a list of some possible markers: *MEAS, MEAS+SI+…, MEAS+D+SI+…* . They correspond to the above-mentioned subgraphs' respective predestinations. Three dots in the last two markers can be replaced by special markers within a respective subgraph that works out. At the same time names of MUs (or their word forms) correspond to names of respective physical values (or their word forms). Take the word combination дадаць *3,3 молі* 'add 3,3 moles' as an example. The algorithm will recognize the following expression: *3,3 молі* '3,3 moles'. It will receive the following marker: *MEAS+SI+Amount of substance*. The marker enables one to identify exactly which subgraph works out and which unit of measurement is used. The code *MEAS* means that the expression *3,3 молі* '3,3 moles' contains a unit of measurement *молі* 'moles'. The code + *SI* informs

that the MU *молі* 'moles' belongs to the SI units. The code + *Amount of substance* means that *молі* 'moles' are used for measuring amounts of substances. The component *D* of the marker *MEAS+D+SI+…* requires the existence of the second distinct subgraph in order to separate expressions with MUs derived from the SI basic units, i. e., *degree Celsius, hertz, radian, newton, joule, pascal, watt, volt, ohm, becquerel*.



a) b)

Fig. 5. The subgraphs which identify expressions with SI-derived units for (a) Belarusian and (b) Russian

Such a flexible system of markers allows building search queries of different types: to find all expressions with MUs (Fig. 6); to find expressions without derived units *(<MEAS+SI-D>)* (Fig. 7); etc. Table 1 contains the search results in Fig. 6 and Fig. 7 translated into English and listed from top to bottom.

Before	Seq.	After
ашэнне –	1м/<MEAS+Length|Distance+SI>	(бач.),
2-30 кэв,	0,1 Гц/<MEAS+Frequency+D+SI>	-300 кгц
вую масу	8 т/<MEAS+Mass>	, вывед
Зямлі. У	2005 г./<MEAS+Time>	Іран зд
н і ўхілам	74 градусы/<MEAS+Angle>	. Затым

a)

Before	Seq.	After
ература	109 К/<MEAS+Thermodynamic temperature+...	В эт
стью ок.	200 000 л/<MEAS+Volume>	. Желе
а спустя	33 года/<MEAS+Time>	– и ег
вышало	5°/<MEAS+Angle>	), а пот
ре выше	600° С/<MEAS+Temperature in Celsius scal...	, а хал

b)

Fig. 6. Results of identification of QE with MUs in (a) Belarusian and (b) Russian

цыі на ўзроўні 1–	10 м	. Такая дэта
иаса перавышае	3600 кг	. Разліковы
масай меней за	10 кг	, а праз 10–2
– масай парадку	1 кг	, якія змогуц
ала парадку 150–	500 метраў	. У 70–80-х г

a)

– ОД % (вид.),	0,1 К	(ИК), 1д
разрешение –	1м	(вид.), 5
тупая 19 апреля	1904 с	больши
через каждые	30 секунд	трех бро
рез 30, а через	3 секунды	, то прям

b)

Fig. 7. Results of identification of QE with only SI-units of measurement on the request <MEAS+SI-D> in (a) Belarusian and (b) Russian

Table 1. Search results in Fig. 6, Fig. 7 translated into English

	Figure 6	Figure 7
a)	1m <MEAS+Length|Distance+SI> 0,1Hz <MEAS+Frequency+D+SI> 8 t <MEAS+Mass> year 2005 <MEAS+Time> 74 degrees <MEAS+Angle>	10 m 3600 kg 10 kg 1 kg 500 metres
b)	109 K <MEAS+Thermodynamic temperature+SI> 200 000 l <MEAS+Volume> 33 years <MEAS+Time> 5° <MEAS+Angle> 600°C <MEAS+Temperature in Celsius scale+D+SI>	0,1 K 1m 1904 30 seconds 3 seconds

Identification of MUs with metrological prefixes

First of all, the authors created necessary linguistic dictionaries *S* for Belarusian and Russian (Fig. 8). They contain some basic stems of MUs — complete nouns and their abbreviations. Each stem is marked by a respective attribute: either *Base* or *Mbase*. In addition, descriptions of full stems include indicators of respective inflectional classes. The dictionary *S* is obviously a language-dependent linguistic resource, unlike algorithms for identification of MUs with metrological prefixes, which are implemented as language-independent components. The next step was to develop language-dependent linguistic resources (*Fsubmultiple*, *Fmultiple*, *Ssubmultiple*, *Smultiple*) (Fig. 9). For MUs-formation either *multiple* or *submultiple* prefixes can be used. Besides, they can take a shortened (*S*-) or full (*F*-) form (Fig. 10) [5].

```
г,ABBREVIATION+Mbase
га,ABBREVIATION+Mbase
гг,ABBREVIATION+Mbase
гектар,NOUN+FLX=ГЕКТАР+s5+UNAMB+Base
герц,NOUN+FLX=АМПЕР+s2+UNAMB+Base
год,NOUN+FLX=ГОД+sN+UNAMB+Base
град,ABBREVIATION+Mbase
грам,NOUN+FLX=ГРАМ+s3+UNAMB+Base
```

a)

```
ампер,NOUN+FLX=АЛТЫН+s4+UNAMB+Base
А,ABBREVIATION+Mbase
байт,NOUN+FLX=АБАЖУР+s2+UNAMB+Base
бит,NOUN+FLX=АБАЖУР+s2+UNAMB+Base
Б,ABBREVIATION+Mbase
ватт,NOUN+FLX=АЛТЫН+s2+UNAMB+Base
Вт,ABBREVIATION+Mbase
вольт,NOUN+FLX=АЛТЫН+s2+UNAMB+Base
```

b)

Fig. 8. Dictionary resources of basic MUs' stems for (a) Belarusian and (b) *Russian*

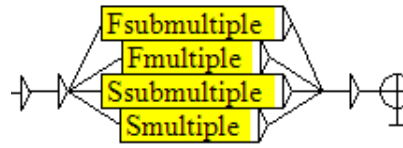


Fig. 9. Classifying metrological prefixes using a NooJ finite-state automaton

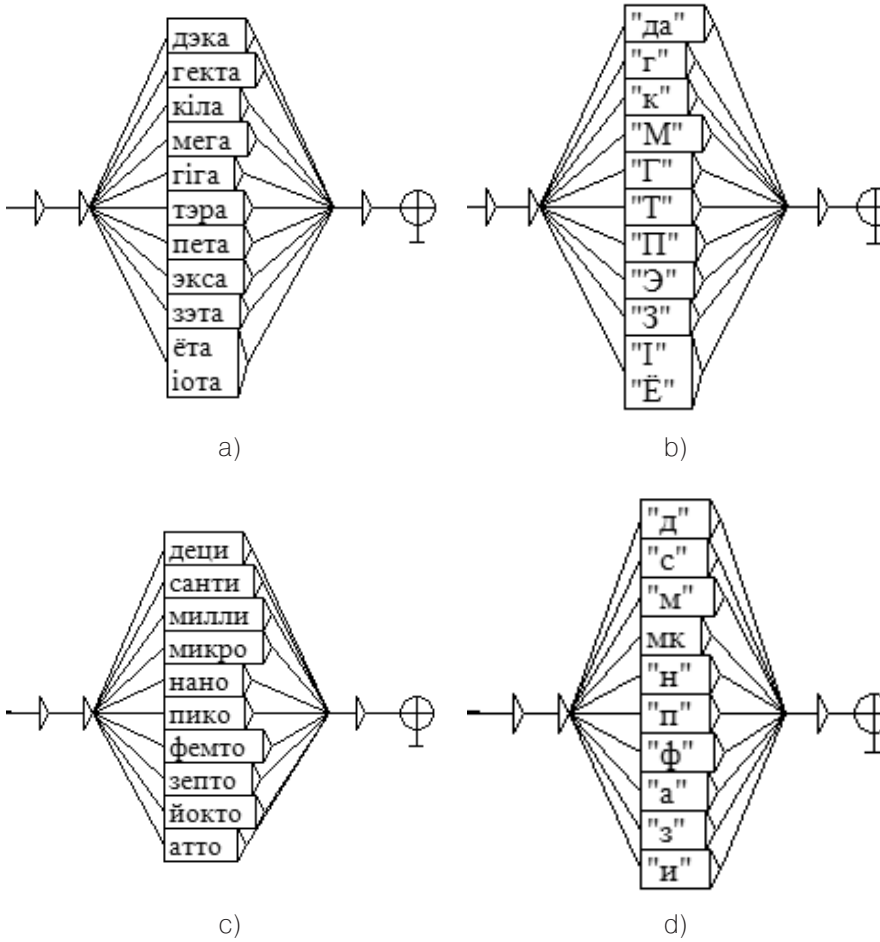


Fig. 10. Graphs for identification of (a) full-stem and (b) shortened-stem multiple prefixes for Belarusian, and (c) full-stem and (d) shortened-stem submultiple prefixes for Russian

The basic principle for the components became the following word-formative classification of MUs:

- MUs with full-form stems and without prefixes (*метр* 'meter', *Герц* 'hertz', *Ом* 'ohm');
- MUs with shortened stems and without prefixes (*Дж* 'J', *га* 'ha');
- MUs with full-form stems and full-form prefixes (*нанафарады* 'nanofarads', *міліампер* 'milliampere');
- MUs with full-form stems and shortened prefixes (*кБайт* 'Kbyte');
- MUs with shortened stems and shortened prefixes (*км* 'km', *дл* 'dL', *гПа* 'hPa').

Depending on word formation peculiarities, 4 morphological language-independent grammars M1-M4 (algorithms) were obtained. They use the dictionary *S* and

linguistic resources *Fsubmultiple, Fmultiple, Ssubmultiple, Smultiple*. For example, the morphological grammar M2 identifies MUs which are formed with the help of multiple and/or submultiple full-form prefixes (Fig. 11).

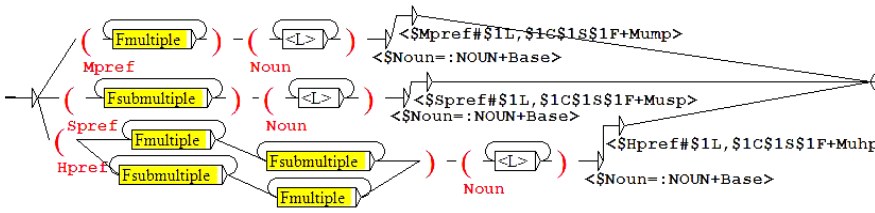


Fig. 11. The morphological grammar M2 which identifies MUs with full-form stems and full-form multiple and/or submultiple prefixes

As a result of its work, MUs may be given one of the following markers:

- *Mump* means that identified MUs have multiple prefixes;
- *Musp* implies that identified MUs have submultiple prefixes;
- *Muhp* denotes MUs which have several prefixes, e. g.: *мікрамегафарад* 'micro-megafarad'. According to the SI, such a way of formation is not common among MUs, so such words require a specified marker, so later they can be extracted from text within a list of mistakes.

Fig. 12 represents operation examples of the above-described morphological component. Note that the obtained morphological components enable the identified MU to inherit all grammatical and inflectional characteristics of initial words. E. g., the word *дэкалітрамі* 'deciliters' (in the Instrumental case) will remain the noun with all its inflectional endings and grammatical features, though the resource dictionary *S* does not contain it (Fig. 13).

Before	Seq.	After
несколько сотен	километров	. Первый вариант
ıгать нескольких	килограммов	. Куски брони пора
цностью в сотни	мегаватт	. Проблема в том
е десятки тысяч	мегагерц	, что соответствуе
іх дисках тысячи	гигабайт	информации, тре
лучения порядка	мегаджоуля	(106 Дж) и кпд

a)

Before	Seq.	After
пашырыць да 2	тэрабайт	! Паскаральнік
;еткі «усяго» 50	кілават	. Астатнія канст
кунд. Таму 425	кілаграмаў	рабочага цела
тую ж мэта 300	кілаграмаў	аргону штогод,
(магутнасцю да	мегавата	) (ілюстрацыя А
ўстаноўкі ў 200	мегават	. Шмат. Але зат

b)

Fig. 12. The resulting concordance of full-stem MUs on the request <NOUN+Mump> as applied to (a) Belarusian and (b) Russian

дэкалітр,NOUN+Meaning=Common
+Animation=Inanimate
+Case=Instrumental
+Gender=Masculine+Number=Plural
+s2+Meas=Base+Mump

a)

наносекунда,NOUN+ProperCommon=Common
+Gender=Feminine+Animation=Inanimate
+Case=Instrumental+Number=Plural
+s4+Meas=Base+Musp

b)

Fig. 13. Examples of annotated word forms for (a) Belarusian and (b) Russian

Finally, the algorithm proceeds to the syntactic grammar *S1* (Fig. 14). It accumulates all the markers from the text *T*, placed by means of the dictionary *S* and morphological grammars *M1-M4*. It works out only for QE with MUs (numerical descriptors in front of them). Numerical descriptors are identified by the inbuilt syntactic component *S1*. Each QE with MUs receives the marker <*MUEXPR*>. It enables users to create concordances of QE with MUs (Fig. 15).

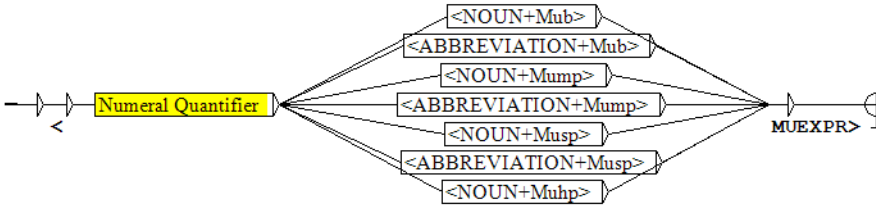


Fig. 14. The main syntactic component S2, which identifies QE with MUs

Before	Seq.	After
адзінку масы - грам (0,001 кг).	31 мкТл	(3,1×10^(-5) Тл) - напружанасць магні
звычайна вар'іруецца зблізку	2,4 мЗв	у год. 1 Н ёсць
на апору з сілай	9.81 Н	. Прыбліжэнне, што 1 кг адпавядае
дамі або нанафарадамі (пішуць	60 000 пф	, а не 60 нф; 2 000 мкф
ёмістасць шара з радыусам	1 сантыметр	, змешчанага ў вакуум. 1 сантыметр
Mbit(альбо проста Mb).	1 мегабіт	= 1000^2 біт = 10^6 біт = 1000000 біт.
святло ў вакууме за (	1 / 299 792 458) секунды	. Метр быў упершыню ўведзены
ны дыяпазон - 40 кэв-3 МЭВ, 2-	200 МЭВ	, 2-30 кэв, 0,1 Гц-300 кгц, 0-50 кгц
ай трубцы тэлевізара - парадку	20 кілаэлектронвольт	. Энэргіі касьмічных прамянёў - ад
энергіі касьмічных прамянёў - ад	1 мегаэлектронвольта	да1000 тэралектронвольтаў.

a)

Before	Seq.	After
организм ток не превышал	1 мА	. На человека токи статического
могут сказать «файл в	100 килобайт	»). При обозначении скоростей тел
ротивление величиной от 1 до	100 МОм	, чтобы протекающий через челове
икромегафарад пикотеравольт	13 йоттайоктограммов	Каждая строка содержит информа
до 64 Мбит/с) и	137,4 МГц	(метровый диапазон, формат АРТ
евонширский изумруд» массой	1383,95 каратов	. Изумруды выращивают искусств
время жизни мюонов - около	2.2 мкс	- осложняет задачу создания мюон
сса которой оказалась равной	22 фемтограммам	(1 фг = 1•10^(-15) г). . Мюоны, как
то они оказались равными:	8.1·10^21 Дж	(уменьшение массы ледников на
: - высота 670 км - наклонение	98,00 град	. Срок активного существования 1 г

b)

Fig. 15. Some results of identification of QE with MUs after processing (a) Belarusian and (b) Russian texts by means of the obtained morphological M1–M4 and syntactic S1 grammars

Generation of orthographical words from QE with MUs

In text-to-speech synthesis tasks it is important to develop algorithms not only for identification of definite expressions but also for their processing and transformation into orthographical word sequences. With this aim, grammars in the form of visual finite-state automata for Belarusian and Russian were worked out. As a result, for each language a ramified algorithmic complex of 21 graphs and subgraphs was obtained. Fig. 16 represents the structure of the main graph. Since QE with MUs consist of 2 components (numbers and nouns), it is required to work out separate graphs for their generation. This algorithm contains of graphs of 2 types. Those, which have names starting with *a_*, generate numbers from 0 to 999,999,999,999. All the rest, in particular the ones with *b_*, are intended for generation of nouns which denote MU.

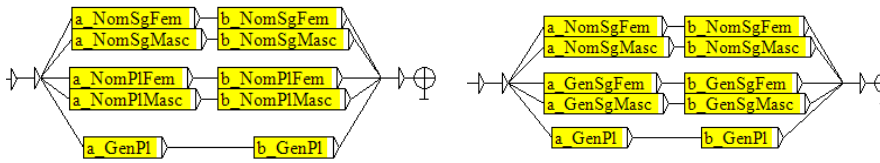


Fig. 16. The main graph of the algorithm which generates orthographical words from QE with MUs for (a) Belarusian and (b) Russian

QE with MUs pass from input to output by means of one of 5 ways in accordance with peculiarities of the inflection of nouns after numerals, in particular for the first 3 ways:

1. After number *1* (including numbers with *1* as a final digit) nouns take endings of the Nominative Singular *(NomSg)*. QE will proceed to one of the top branches, depending on the gender of nouns, in particular *Masculine (Masc)* or *Feminine (Fem)*.
2. After numbers *2, 3, 4* (including numbers with *2, 3* or *4* as a final digit) nouns take the Nominative plural *(NomPl)* in Belarusian, whereas in the Russian these numbers require nouns in the Genetive singular *(GenSg)*. Depending on the gender, QE will move to branches 3 or 4.
3. Numbers from 5 to 19 and round numbers (including numbers with them as final digits) require nouns in the Genitive plural *(GenPl)* in both languages. QEs will follow the 5th branch.

As an example, let us stop on the first branching of the algorithm, in particular the graph *a_GenPl* (Fig. 17). It generates any whole number from 0 to 999,999,999,999, which demands the Genitive plural form.

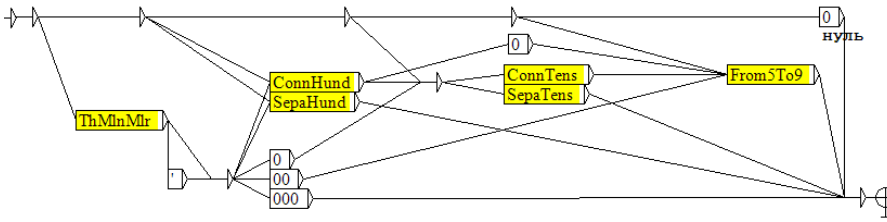


Fig. 17. The subgraph a_GenPl for Belarusian

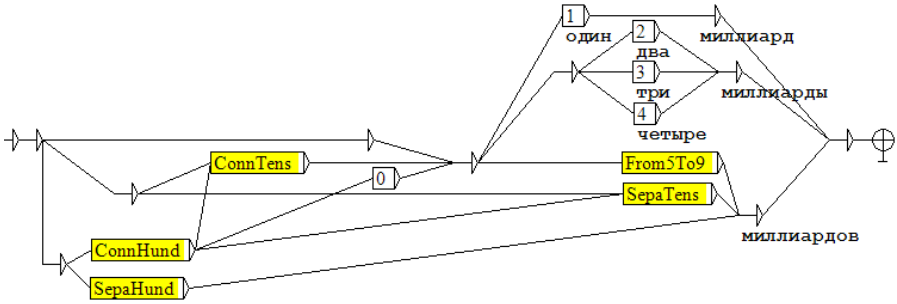
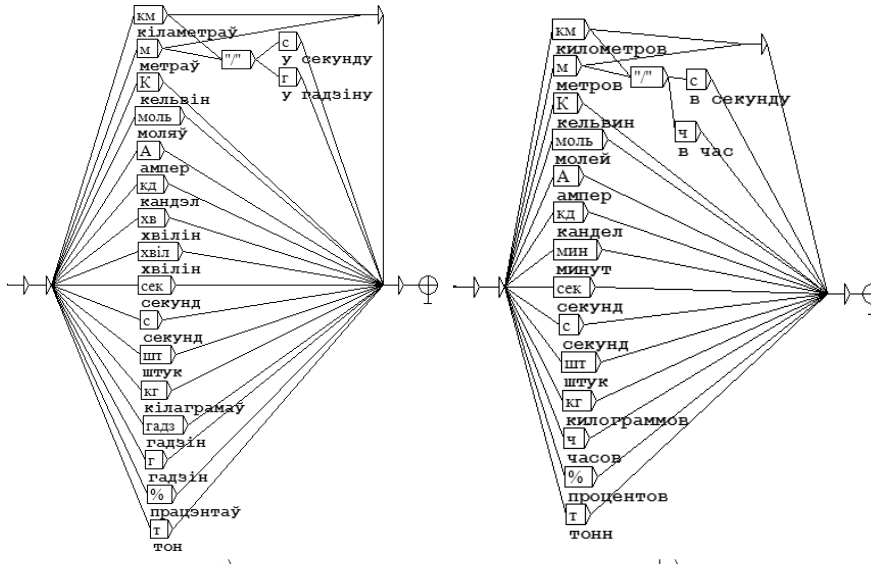


Fig. 18. The subgraph Mlr for Russian

The structure of the algorithm for generation of numbers resembles Russian dolls. At first the graph for numbers of the first triad (from 0 to 999) was obtained. It includes the inbuilt subgraph *ThMlnMlr* for the class of thousands or numbers with 2 triads (from 1,000 to 999,999). Inside of this subgraph the other one (*MlnMlr)* was placed for the class of millions or three-triads numbers; at last, the subgraph *Mlr* (Fig.18) for the class of billions or numbers with 4 triads (from 1,000,000,000 to 999,999,999,999) was worked out. Depending on research goals, the algorithm can be expanded by further triads. After generating numbers the algorithm proceeds to processing nouns which denote MUs. Concerning the last branch of the algorithm, it happens with the help of the graph *b_GenPl* (Fig. 19).

For the present, this subgraph can generate basic SI units and some frequently used ones. Thanks to the visuality of finite-state automata, the algorithm can be easily and rapidly improved by adding more MUs. In order to add a new unit, three case endings (mind the gender and number) should be added to the respective graphs. Variations of written forms should also be taken into account. For example, for the noun градус (shortened гр, ° — three variants; in English *degree*, shortened *deg*, °), one should add 3 respective word forms (градус, градусы, градусаў for Belarusian; and градус, градусы, градусов for Russian) for each variant into the following graphs: *b_NomSgMasc, b_NomPlMasc, b_GenPl* for Belarusian; *b_NomSgMasc, b_GenSgMasc, b_GenPl* for Russian.



a) b)

Fig. 19. The subgraph b_GenPl for (a) Belarusian and (b) Russian

Thus, language-dependent complexes of grammars for generation of orthographic words from QE with MUs have been obtained. Fig. 20 demonstrates some results of their operation.

700001г/семсот тысяч адна гадзіна
0 с/нуль секунд
777'700т/семсот семдзесят сем тысяч семсот тон
888'808хв/восемсот восемдзесят восем тысяч восемсот восем хвілін
2220020 хвіл/два мільёны дзвесце дваццаць тысяч дваццаць хвілін
444'014моль/чатырыста сорак чатыры тысячы чатырнаццаць моляў

a)

10120202 мин/десять миллионов сто двадцать тысяч двести две минуты
70000000071 шт/семьдесят миллиардов семьдесят одна штука
81234999 А/восемьдесят один миллион двести тридцать четыре тысячи девятьсот девяноста девять ампер
8600км/ч/восемь тысяч шестьсот километров в час
90673 м/с/девяноста тысяч шестьсот семьдесят три метра в секунду

b)

Fig. 20. Generation of orthographic words from QE with MU for (a) Belarusian and (b) Russian with the help of the developed algorithms

Variety of ways to express QE with MU in Belarusian and Russian texts

Since the practical goal is to identify MUs and generate expressions with them, a question inevitably arises: which ways of written forms should be taken into consideration? Thus, it is required to make a certain sample of QE in order to cover all the variety of ways of their expression in writing.

In order to attain this aim, a hand-crafted corpus (about 1 mln words) of Russian scientific texts (in the format NooJ) was used. With the help of statistical computations 8 texts with highest amounts of MUs were selected, in particular texts on: *military equipment (ME), geography (Geo), history (His), space travel science (STS), physics (Phy), mineralogy (Min), transportation communications (TC), botany (Bot)*. After that a special formula in the form of a regular expression was created (Fig. 21), starting with the simple (<WF><NB>).

Further the right and left contexts were analyzed, so the formula was constantly updated. Table 2 demonstrates what happens when the formula is applied to the selected texts, whereas Table 3 contains generalized data for them.

```
(от <NB> до <NB>)|(от <NB> до <NB><NB>)|(от <NB><NB> до <NB>)|(от <NB>,<NB> до <NB>,<NB>)|
(от <NB>,<NB> до <NB>)|(от <NB>.<NB> до <NB>.<NB>)|(от <NB> до <NB>,<NB>)|
(от <NB>–<NB> до <NB>–<NB>)|(от <NB>×<NB>-<NB> до <NB>×<NB>-<NB>)|
(от <NB>×<NB>–<NB> до <NB>×<NB>–<NB>)|(от <NB> до почти <NB>)|(<WF>|"~"|"=")(<NB>|(<NB><NB>)|
(<NB><NB><NB><NB><NB>)|(<NB><NB><NB><NB><NB>)|(<NB><NB><NB>)|(<NB>×<NB>)|(<NB>,<NB>)|
(<NB>–<NB>)|(<NB>—<NB>)|(<NB>-<NB>)|(<NB>.<NB>-<NB>.<NB>)|(<NB>.<NB>)|(<NB>–<NB>,<NB>)|
(<NB>.<NB>.<NB>)|(<NB>,<NB>–<NB>,<NB>)|(<NB>,<NB>–<NB>)|(<NB>,<NB>-<NB>)|(<NB>-<NB>,<NB>)|
(<NB>"/"<NB>)|(<NB>":"<NB>)|(<NB>":"<NB>-<NB>)|(<NB>":"<NB>":"<NB>)|(<NB>,<NB>×<NB>–<NB>)|
(<NB>×<NB>–<NB>)|(<NB>´<NB>)|(<NB>,<NB>´<NB>,<NB>)|(<NB>´<NB>,<NB>)|(<NB>·<NB>)|
(<NB>,<NB>·<NB>–<NB>)| (<NB>.<NB>•<NB>)|(<NB>•<NB>–<NB>)|(<NB>.<NB>•<NB>–<NB>)|
(<NB>.<NB>·<NB>)|(<NB>,<NB>·<NB>)|(<NB>·<NB>–<NB>)|(">"<NB>)|(<NB>">>"<NB>)|
(<NB>|(<NB><NB>)|(<NB>,<NB>)|(<NB>–<NB>)|(<NB>—<NB>)|(<NB>-<NB>)|(<NB>.<NB>)|
(<NB>,<NB>–<NB>,<NB>)|(<NB>,<NB>–<NB>)|(<NB>-<NB>,<NB>)|(<NB>":"<NB>)|
(<NB>,<NB>×<NB>–<NB>)|(<NB>,<NB>·<NB>)|(<NB>·<NB>–<NB>)|(<NB>x<NB>)|(<NB>x<NB>')|
(<NB>x<NB>x<NB>)|(<NB>.<NB>x<NB>.<NB>)|(<NB>,<NB>×<NB>)|(<NB>¸<NB><NB>)|
```

Fig. 21. The regular expression for a search for QE with MUs

Table 2. Excerpts: variety of QE, revealed with the help of the obtained formula

Formula's constituents	Examples of QE, found with the help of a certain constituent
(от <NB> до <NB>)	находится в пределах от 40 до 347 г/л. Несмотря лежит в диапазоне от 2000 до 3000 А, и следовательно удаленной на расстояние от 160 до 640 км от восточного мере удаления "Пионера-10" от 40 до 60 а.е. величина
(от <NB> до <NB><NB>)	находятся в диапазоне от 4000 до 14 000 МГц и даже вариации) с периодами от 10 до 10 000 лет, сосредоточенными в
(от <NB><NB> до <NB>)	см (частотный интервал от 600 000 до 1000 МГц) относится к
(от <NB>,<NB> до <NB>,<NB>)	семь камней массой от 94,45 до 4,39 кар каждый. Кроме между рельсами пути от 1,52 до 1,68 м) типичны для кристалла со скоростями от 0,01 до 0,001 скорости света; их
(от <NB>,<NB> до <NB>)	видов токсичность составляет от 0,01 до 1 %. Практически это означает
(от <NB>.<NB> до <NB>.<NB>)	диапазоне длин волн от 1.3 до 1.7 мкм. На подложке

Formula's constituents	Examples of QE, found with the help of a certain constituent
(от <NB> до <NB>, <NB>)	отношениях ионных радиусов от 1 до 0,732 (рис. 4,а). При С-диапазоне и от 12 до 12,7 ГГц в Q
(от <NB> – <NB> до <NB> – <NB>)	выемчатые. Их длина от 1–2 до 30–40 см. Самые длинные длиной волны l от 10–3 до 10–8 м. Этот диапазон
(от <NB> ×<NB> – <NB> до <NB> × <NB> – <NB>)	с удельным сопротивлением от 5×10-8 до 8×10-5 Ом×м. Композиционные
(от <NB> × <NB> – <NB> до <NB> × <NB> – <NB>)	в разных материалах: от 3×10–6 до 2×10–5 см. Магнитный поток
(от <NB> до почти <NB>)	током (при этом от 50 до почти 100 % его энергии превращается
(<WF>|»~»|»=»)<NB>	Мировом Океане составляет около 550 млрд. тонн в излучения с l ~ 10 Å не существует до цели L = 1000 км, получим ограничение затем разгоняются до энергии 5 МэВ на линейном крупного «суперматерика» Го... Около 160 млн. лет назад
(<WF>|»~»|»=») (<NB><NB>)	отравлений растениями страдают примерно 15 000 человек. Для домашних ядерных взрывов суммарной силой 10 000 Мт в центральных и в среднем на 386 063 км от центра В 1990 она насчитывала приблизительно 900 000 верующих, в основном установки. В 9.50 на высоте 15 800 м Волков - первым в

Table 3. Distribution of various ways of expression of QE in different texts

Text	*Phy*	*STS*	*Geo*	*ME*	*Min*	*Bot*	*TC*	*His*
Number of variations, *a*	51	23	22	19	18	16	14	9
Number of numeral expressions, *b*	2841	2245	2765	9961	3668	1407	2066	4198

Intonation marking in sentences which contain QE with MUs

Text-to-speech synthesis requires an automatic procedure of building current contours of melody, sound intensity, phoneme and pause duration, which is based on the analysis of certain properties of sentences according to rule-based prosodic marking. Prosodic marking of sentences implies their division into syntagmas, marking emphatically highlighted words, indicating syntagmas with accent units, and creating a melodic contour of each syntagma in accordance with certain rules. Solutions to these problems by means of in-depth syntactic analysis are thoroughly discussed in [6, 8]. Texts, when being synthesized, are first reduced to a normalized orthographic form. Next they undergo a complete syntactic analysis, performed by the parser ЭТАП-3 'ETAP-3'. The parser (1) divides texts into separate sentences; (2) for each sentence it builds treelike syntactic structures; (3) using special rules, which can be applied to ready syntactic structures, it sets boundaries among speech syntagmas

and emphatically highlighted components. The system Мультыфон 'Multiphone' [7] processes this information and, depending on syntactic types, determines a melodic contour and duration of pauses between syntagmas. Prosodic and intonation marking of sentences which contain QE with MUs can be carried out by the method proposed in the [6, 8]. However, before syntactic analysis of such sentences it is required to generate QE into orthographic words.

Indeed, Fig. 22 gives an example of syntactic analysis of the following sentence: *Расстояние до Марса 55764878 км* 'The distance to Mars is 55764878 km'. As a result of processing this sentence by the system ETAP-3 in accordance with the rules discussed in [6], at the output of the system the following information for synthesis is received: *Расстоя́ние до <EMP t="*4">Ма́рса</EMP> 55764878 км*. Thus, the system suggests no additional partitioning of this sentence into syntagmas. After *55764878 км* '55764878 km' is generated into orthographical words, we have the following result: *Расстояние до Марса пятьдесят пять миллионов семьсот шестьдесят четыре тысячи восемьсот семьдесят восемь километров* 'The distance to Mars is fifty-five million seven hundred sixty-four thousand eight hundred seventy-eight kilometers'. The syntactic analysis of this sentence can be observed in Fig. 23.

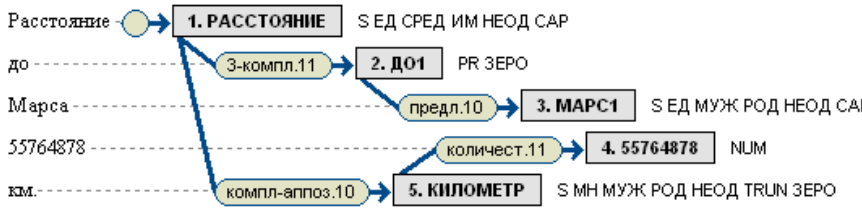


Fig. 22. Syntactic analysis of the sentence *Расстояние до Марса 55 764 878 км*

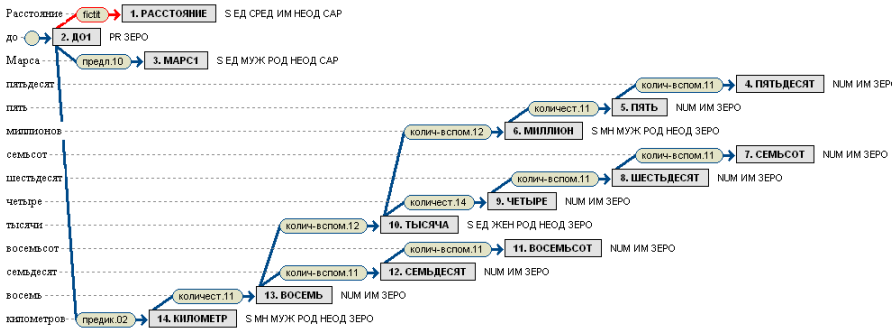


Fig. 23. The syntactic analysis of the sentence *Расстояние до Марса пятьдесят пять миллионов семьсот шестьдесят четыре тысячи восемьсот семьдесят восемь километров*

At the output of the system ETAP-3 the following information for the text-to-speech synthesizer MULTIPHONE is obtained: *Расстоя`ние <EMPt=”*16”> до </EMP> <EMPt=»*4»> Ма`рса</EMP> пятьдеся`т пя`ть <EMPt=»*4»> миллио`нов </EMP> семьсо`т шестьдеся`т четы`ре <EMPt=”*4”> ты`сячи </EMP> восемьсо`т се`мьдесят во`семь <EMPt=»*4»> киломе`тров </EMP>*. According to this information the Multiphone forms 4 syntagmas with melodic contours C01, C3, C3_1, P4 (emphatically highlighted words are indicated with the «+» sign).

1 *C01* *расстоя+ние/доЪма+рса/*
2 *C3* *пядеся=т пя+ть/миллио+нов/*
3 *C3_1* *семьсо=т шездеся=т четы=ре ты+сячи/*
4 *P4* *восемьсо=т се=мьдесят во=семь киломе+тров/*

Fig. 24 demonstrates another example of syntactic analysis, in particular for the following sentence: *Производительность компьютера 65,5 Мбит/с была достигнута* “The computer performance 65,5 Mbit/s was achieved”. The algorithm identifies *65,5 Мбит/с* ‘65,5 Mbit/s’. After processing into orthographical words it gets the following form: *Производительность компьютера шестьдесят пять целых и пять десятых мегабит в секунду была достигнута* ‘The computer performance sixty-five point five megabits per second was achieved’.

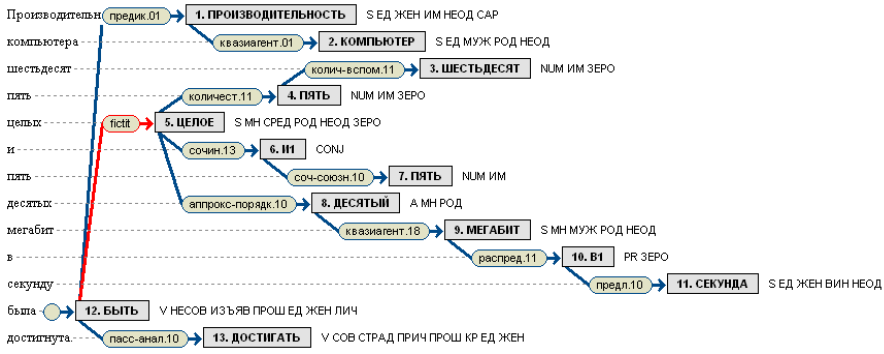
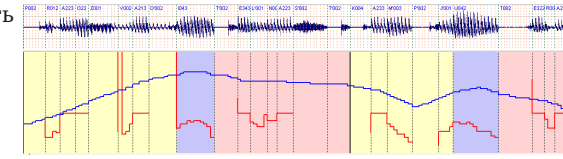


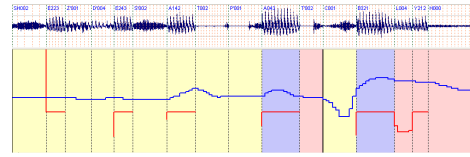
Fig. 24. The syntactic analysis of the sentence *Производительность компьютера шестьдесят пять целых и пять десятых мегабит в секунду была достигнута*

According to the data obtained by the ETAP-3, the MULTIPHONE forms 4 syntagmas (Fig. 25):

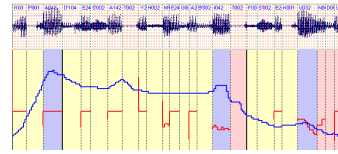
1 C4 производи+тельность компью+тера /



2 C1 шестьдеся=т пя+ть/ це+лых/



3 C3 и= пя+ть/деся=тых мегаби+т/вЪсеку+нду/



4 P4 была= дости+гнута/

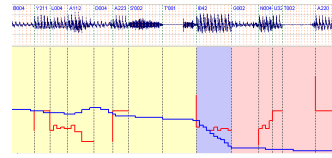


Fig. 25. Syntagmas and melodic contours for *«Производительность компьютера шестьдесят пять целых и пять десятых мегабит в секунду была достигнута»*

Conclusion

It can be concluded that the main goal of this research — to develop appropriate algorithms which *identify* quantitative expressions with various MUs and *generate orthographic texts* for the Belarusian and Russian languages for scientific, technical and legal text corpora — has been achieved. The results can be applied in any branches of science connected with information retrieval systems and text-to-speech synthesis. The resulting algorithms are created in the form of finite-state automata through a set of syntactic grammars within the powerful linguistic processor NooJ, which helps to build up formal grammars without requirements for special knowledge of programming. The automata demonstrate how the algorithms work and indicate how they can be further updated in order to improve their accuracy. Future work includes:

- disambiguation, e. g., in such cases when algorithms "confuse" some units (the same initial letter г for год 'year', грам 'gram', гадзі́на 'hour';
- developing algorithms that will identify numeral quantifiers expressed not only by numbers (mathematical objects), but also by numerals (parts of speech).

References

1. *Hetsevich Yu. S., Hetsevich S. A.* (2012), Overview of Belarusian and Russian dictionaries and their adaptation for NooJ, Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011, Dubrovnik, pp. 29–40.

2. *Hetsevich Yu. S., Hetsevich S. A., Lobanov B. M.* (2012), Belarusian and Russian linguistic modules processing for the system NooJ as applied to text-to-speech synthesis, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'juternaja Lingvistika i Intellektual'nye Tehnologii: po Materialam Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 198–212.
3. *Hetsevich Yu. S., Hetsevich S. A., Lobanov B. M., Yakubovich Ya.* (2012), Belarusian module for NooJ, available at: http://www.nooj4nlp.net/pages/belarusian.html
4. *Hetsevich Yu. S., Skopinava A. M.* (2012), Identification of Expressions with Units of Measurement in Scientific, Technical and Legal Texts in Belarusian and Russian [Ідэнтыфікацыя выразаў з адзінкамі вымярэння ў навукова-тэхнічных і прававых тэкстах на беларускай і рускай мовах], Development of information and the state system of scientific and technical information (DISTI-2012): Reports of the XI International Conference [Razvitie Informatizatsii i Gosudarstvennoj Sistemy Nauchno-Tehnicheskoj Informatsii (RINTI-2012): Doklady XI Mezhdunarodnoj Konferentsii], Minsk, pp. 260–265.
5. *Hetsevich Yu. S., Skopinava A. M.* (2013), Components for Identification of Quantitive Expressions with Measurement Units in Belarusian and Russian Texts [Кампаненты ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння ў тэкстах на беларускай і рускай мовах], Open Semantic Technologies for Intelligent Systems (OSTIS–2013): Proceedings of the III International scientific and technical conference [Otkrytye Semanticheskie Tehnologii Proektirovanija Intellektual'nyh Sistem (OSTIS–2013): Materialy III Mezhdunarodnoj Nauchno-Tehnicheskoj Konferentsii], Minsk, pp. 319–328.
6. *Iomdin L. L., Lobanov B. M., Hetsevich Yu. S.* (2011), The talking ETAP. Using the ETAP parser in Russian speech synthesis [Govorjashhij "ÈTAP". Opyt Ispol›zovanija Sintaksicheskogo Analizatora Sistemy ÈTAP v Russkom Rechevom Sinteze], Proceedings of the International Conference "Computational Linguistics and Intellectual Technologies" (Dialog'2011) [Trudy Mezhdunarodnoj Konferentsii "Komp›juternaja Lingvistika i Intellektual›nye Tehnologii" (Dialog'2011)], Bekasovo, pp. 269–279.
7. *Lobanov B. M., Tsirulnik L. I.* (2008), Computer speech synthesis and cloning [Komp›juternyj sintez i klonirovanie rechi], Belarusian Science [Belorusskaja Nauka] Publ., Minsk.
8. *Lobanov B. M., Iomdin L. L.* (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 èlementov predlozhenija i ih rol› v zadachah sinteza rechi po tekstu], Proceedings of the International Conference "Computational Linguistics and Intellectual Technologies" (Dialog'2009) [Trudy Mezhdunarodnoj Konferentsii "Komp›juternaja Lingvistika i Intellektual›nye Tehnologii" (Dialog'2009)], Bekasovo, pp. 339–348.

ПАДЕЖНАЯ НЕОДНОЗНАЧНОСТЬ ПРИ ВОСПРИЯТИИ (ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА)

Слюсарь Н. А. (slioussar@gmail.com)

Утрехтский лингвистический институт, Утрехт, Нидерланды; СПбГУ, Санкт-Петербург, Россия

Череповская Н. В. (ajmi@yandex.ru)

СПбГУ, Санкт-Петербург, Россия

Ключевые слова: морфологическая неоднозначность, падеж, ошибки в согласовании с интерференцией, русский язык, восприятие

PROCESSING OF CASE MORPHOLOGY: EVIDENCE FROM RUSSIAN

Slioussar N. A. (slioussar@gmail.com')

Utrecht Institute of Linguistics OTS, Utrecht, Netherlands; Saint-Petersburg State University, Saint-Petersburg, Russia

Cherepovskaia N. V. (ajmi@yandex.ru)

Saint-Petersburg State University, Saint-Petersburg, Russia

Many studies discuss how morphological ambiguity influences processing. In particular, it is well known that attraction errors in subject-verb agreement are produced more often and cause smaller delay in comprehension if the form of the intervening noun coincides with the Nominative case form. This is the case in the German example *die Stellungnahme gegen die Demonstrationen waren...* 'the position against the demonstrations (Acc. Pl=Nom.Pl) were' as opposed to *die Stellungnahme zu den Demonstrationen waren...* 'the position on the demonstrations (Dat.Pl≠Nom.Pl) were'. However, the explanation of this phenomenon is a matter of debate. How are such errors produced or missed in comprehension, how are ambiguous forms represented so that they can influence this process?.. We offer a novel perspective on this problem by looking at novel data. We conducted two self-paced reading experiments exploring how Russian adjective forms ambiguous for case influence processing of case errors on the following nouns. We compare sentences containing errors like *fil'my bez izvestnyh akterah* 'movie.$_{NOM.PL}$ without famous.$_{GEN.PL=PREP.PL}$ actor.$_{PREP.PL}$' and *fil'my bez izvestnyh akteram* 'movie.$_{NOM.PL}$ without famous.$_{GEN.PL≠DAT.PL}$ actor.$_{DAT.PL}$' to grammatically correct sentences. Errors of the first type are detected later and their effect is less pronounced. The results help answering several questions that arise in connection with attraction errors in subject-verb agreement.

Keywords: morphological ambiguity, case, attraction errors, agreement, Russian, comprehension

1. Introduction

This paper addresses the question how morphologically ambiguous forms are processed and influence the processing of other words. Our study focuses on case ambiguity in Russian adjectives, but we will start with another piece of data because it has already been studied experimentally in several languages. So we will use it to establish the necessary background and will refer to it later when interpreting the results of our experiments.

In the last two decades, many production and comprehension studies analyzed so-called agreement attraction errors. The data came from a variety of languages including Russian (e.g. Bock, Miller 1991; Eberhard et al. 2005; Fayol et al. 1994; Franck et al. 2002, 2006; Lorimor et al. 2008; Vigliocco et al. 1995, 1996; Wagers et al. 2009; Wilson, Nicol 1999; Yanovich, Fedorova 2006). However, almost all studies focused primarily on number agreement between the subject and the predicate. A classical English example is given in (1).

(1) *The key to the cabinets are rusty.*

The term *attraction* is used to describe the following phenomenon. The verb *are* erroneously agrees not with the head of the subject NP *key*, but with an intervening noun, or attractor, *cabinets*. Such errors frequently occur naturally and are produced in high numbers in experimental conditions. Compared to them, agreement errors without attraction, like (2), are very rare. It was also demonstrated that people tend to overlook the same agreement errors that they produce more often (Pearlmutter et al. 1999, a.o.). This tendency can be traced in reading times, in grammaticality judgment accuracy and in ERP data.

(2) *The key (to the cabinet) are rusty.*

Various syntactic, semantic and morphological factors affecting production and perception of agreement attraction errors were examined, which sheds light on the workings of the mental grammar. In particular, it was noted that in the languages with case morphology, like German, they are produced more often and cause smaller delay in comprehension if the form of the intervening noun coincides with the Nominative case form, as in (3a) compared to (3b) (e.g. Hartsuiker et al. 2003).

(3) a. *Die* *Stellungnahme* *gegen* *die* *Demonstrationen* *waren…*
ART.NOM.SG position against ART.ACC.PL(=NOM.PL) demonstrations were'

b. *die* *Stellungnahme* *zu* *den* *Demonstrationen* *waren…*
ART.NOM.SG position on ART.DAT.PL(≠NOM.PL) demonstrations were'

Intuitively, we can make the following conclusion: although we know on some level that the intervening noun is not Nominative, it can be mistaken for the subject. But how exactly this happens is a matter of debate. Our self-paced reading study capitalizing on particular morphological characteristics of Russian offers a novel view on this question.

In Russian, some adjective forms are ambiguous between different cases: Gen. Sg, Dat.Sg, Instr.Sg and Prep.Sg for Feminine forms, and Gen.Pl and Prep.Pl for all genders. Rusakova (2001, 2009 etc.) who studied naturally occurring errors in Russian noted several examples like (4). We decided to study such errors in detail and so far conducted two comprehension experiments[1].

(4) *v* *teh* *razmerov*
in those.PREP.PL(=GEN.PL) size.GEN.PL

2. Experiment 1

2.1. Method

27 native speakers of Russian, aged 18–26, took part in our first self-paced reading experiment. The materials consisted of 33 sets of target sentences and 108 fillers. All target sentences contained a subject noun with a PP modifier ('N P Adj/Part N') and a verb with an object or a modifier. NPs inside these PPs were in Gen.Pl and Prep. Pl (where the adjective form is ambiguous) and in Dat.Pl used as a control condition. In every target set, the noun inside the PP was in the correct form in one sentence and in a wrong form in two others. An example is given in (5a–c).

(5) a. *Neudachi* *v* *proshlyh* *sezonah* *zastavili*
failure.NOM.PL in previous.PREP.PL season.PREP.PL make.PST.PL
komandu *potrudit'sja.*
team.ACC.SG work.INF

b. *Neudachi* *v* *proshlyh* *sezonov…*
failure.NOM.PL in previous.PREP.PL(=GEN.PL) season.GEN.PL

c. *Neudachi* *v* *proshlyh* *sezonam…*
failure.NOM.PL in previous.PREP.PL(≠DAT.PL) season.DAT.PL

The resulting experimental conditions are shown in Table 1. Let us note that conditions C2 and C4 contain the errors we are interested in: the preposition requires case A, the adjective form is ambiguous between cases A and B and the noun appears in case B.

[1] It would also be very interesting to study them in production, but experimental techniques used to induce subject-predicate agreement errors are not applicable to this case, and we could not find a suitable alternative.

Table 1. Experimental conditions C1–C9[2]

	Prepositions taking Genitive: 11 sets	**Prepositions taking Prepositional: 11 sets**	**Prepositions taking Dative: 11 sets**
Nouns in Genitive	C1: correct form	C4: wrong form, as in (5b)	C7: wrong form
Nouns in Prepositional	C2: wrong form	C5: correct form, as in (5a)	C8: wrong form
Nouns in Dative	C3: wrong form	C6: wrong form, as in (5c)	C9: correct form

The information in the following paragraphs is also applicable to our Experiment 2, so we will not repeat it in section 3. Filler sentences contained no errors. Every subject saw one sentence from each target each set, so we had three experimental lists in Experiment 1 and six lists in Experiment 2. The number of target sentences in different conditions was balanced across lists. Every list started with five filler sentences, and then target and filler sentences were mixed pseudo-randomly (at most two target sentences with errors appeared in a row).

The experiment was run on a PC using *Presentation* software. Target and filler sentences appeared one by one and were masked. Every key press revealed a new word in a sentence and masked the previously revealed word, and RTs were measured. Comprehension questions with a choice of two answers were asked after 50% randomly selected sentences to ensure that the participants were reading properly.

We analyzed participants' question-answering accuracy and reading times. The raw reading times (per word) that exceeded 1,500 ms were adjusted to this threshold. In total, about 0,4% of the data was adjusted in Experiment 1 and about 0,6% in Experiment 2. As for question-answering accuracy, given that no participant made more than five mistakes, a breakdown of RTs into correct and incorrect question trials was not done.

2.2. Results

We compared average RTs per region in C1–C3, C4–C6 and C7–C9 (see diagrams in Fig. 1). All target sentences were 7 words long, so there were 7 regions in every sentence. There were no significant differences in regions 1–3 (before the nouns in a wrong case appeared) and in regions 6–7. I.e. the effects of violations were local, confined to regions 4–5. Average RTs in these regions are given in Table 2.

[2] Initially, we had 12 sets in every group, but one had to be removed due to a minor mistake in the procedure, and two sets in two other groups were removed to keep materials balanced.

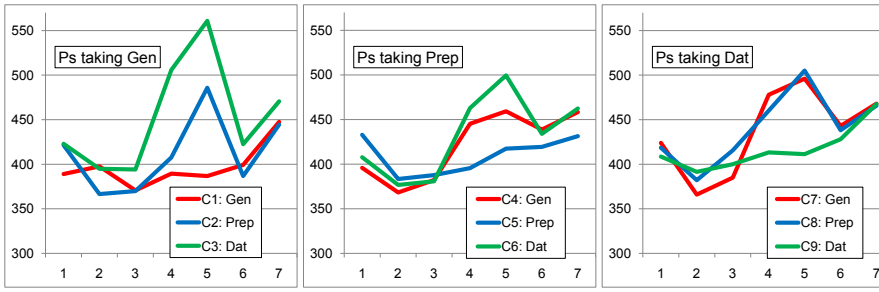


Fig. 1. Average RTs per region (in ms) in different experimental conditions

Table 2. Average RTs (in ms) in regions 4–5 in conditions C1–C9

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Region 4	389.5	407.5	505.7	445.3	395.6	462.9	478.0	460.2	413.4
Region 5	386.7	485.8	560.9	459.3	417.4	499.5	496.0	505.1	411.3

In the sentences with prepositions selecting Genitive, the difference between C1 and C3 is significant both in region 4 ($F1(1,52)=7{,}19$, $p=0{,}01$; $F2(1,20)=6{,}12$, $p=0{,}02$) and region 5 ($F1(1,52)=12{,}26$, $p<0{,}01$; $F2(1,20)=15{,}55$, $p<0{,}01$). The difference between C1 and C2 is significant only in region 5 ($F1(1,52)=6{,}89$, $p=0{,}01$; $F2(1,20)=10{,}180$, $p<0{,}01$). In the sentences with prepositions selecting Prepositional, the difference between C5 and C4 never reaches significance, the difference between C5 and C6 is significant in region 5 ($F1(1,52)=4{,}14$, $p=0{,}05$; $F2(1,20)=5{,}81$, $p=0{,}03$). In the sentences with prepositions selecting Dative all errors are processed similarly. The differences between conditions are not significant in region 4, but reach significance in region 5 ($F1(1,52)=7{,}74$, $p<0{,}01$; $F2(1,20)=13{,}34$, $p<0{,}01$ for C9 vs. C7; $F1(1,52)=9{,}40$, $p<0{,}01$; $F2(1,20)=9{,}23$, $p<0{,}01$ for C9 vs. C8).

To conclude, in C2 and C4, where the adjective form is ambiguous between cases A and B and the wrong noun appears in case B, the slow-down associated with the errors is delayed and less pronounced in comparison with the other cases. Several hypotheses explaining this effect can be suggested. According to the first one, we forget what the case on the noun should be, try to recover it from the adjective and can make a mistake if the adjective is ambiguous. However, this hypothesis is undermined by the fact that the distance between the preposition and the noun is too short. According to the second hypothesis, it is possible to build a local syntactic structure, say, an NP, in C2 and C4, and the violation is discovered only at a later stage, when we embed this NP in a PP, while otherwise, it is visible immediately. However, this does not explain parallel mistakes in production. So we favor the third hypothesis that will be elaborated below: the phenomenon is similar to subject-predicate agreement attraction discussed in the introduction.

To compare these hypotheses, Experiment 2 analyzes how the effect we observed depends on the linear distance between the adjective and the noun. If forgetting or locality are at stake, this effect should increase or decrease, respectively, while agreement attraction phenomena are known to be independent from linear distance (e.g. Bock, Miller 1991). Finally, let us note that it is unclear why all effects are more pronounced in the sentences with Genitive.

3. Experiment 2

3.1. Method

In our second experiment, we used the same methodology as in the first one. 36 native speakers of Russian, aged 17–34, took part in it. The materials consisted of 36 sets of target sentences and 108 fillers. There were six sentences in every set, so we had six experimental lists. As before, all target sentences contained a subject noun with a PP modifier and a verb with an object or a modifier. But this time, three sentences in every set had three words inside the PP ('P Adj/Part N') and the other three had six words: the adjective or participle was followed by a three word long modifier. An example is given in (6a–b). The prepositions required Genitive case in 18 sets and Prepositional case in the other 18 sets (this time, we did not include prepositions taking Dative). In every set, the noun inside the PP was in the correct form in two sentences and in a wrong form in four others. Genitive, Prepositional and Dative case were used, as before.

(6) a. *Listja* *na* *peshehodnyh* *dorozhkah / dorozhek / dorozhkam*
leaf.NOM.PL on pedestrian.PREP.PL(=GEN.PL) path.PREP.PL path.GEN.PL path.DAT.PL

radujut *zolotistym* *tsvetom.*
gladden.PRS.3PL golden.INSTR.SG colour. INSTR.SG

b. *Listja* *na* *idushchih* *vdol'* *krutogo* *berega*
leaf.NOM.PL on going.PREP.PL(=GEN.PL) along steep.GEN.SG bank.GEN.SG

dorozhkah / dorozhek / dorozhkam…
path.PREP.PL path.GEN.PL path.DAT.PL

The resulting experimental conditions are shown in Table 3. Conditions C2, C4, C8 and C10 contain the errors we are interested in: the preposition requires case A, the adjective form is ambiguous between cases A and B and the noun appears in case B.

Table 3. Experimental conditions C1-C12

	Prepositions taking Genitive: 18 sets		**Prepositions taking Prepositional: 18 sets**	
	'Short' conditions	**'Long' conditions**	**'Short' conditions, as in (6a)**	**'Long' conditions, as in (6b)**
Nouns in Genitive	C1: correct form	C7: correct form	C4: wrong form	C10: wrong form
Nouns in Prepositional	C2: wrong form	C8: wrong form	C5: correct form	C11: correct form
Nouns in Dative	C3: wrong form	C9: wrong form	C6: wrong form	C12: wrong form

3.2. Results

We compared average RTs per region in C1–C3, C4–C6, C7–C9 and C10–C12 (see diagrams in Fig. 2). The effects of the violations were local, as in Experiment 1. Target sentences contained 7 words in the short conditions and 10 words in the long conditions. In the short conditions, significant differences were confined to region 4 (where the noun in the wrong case appears) and 5. Average RTs in these regions are given in Table 4. In the long conditions, there were significant differences only in region 8 (following the region where the noun in the wrong case appears). Average RTs in this region are given in Table 5.

Table 4. Average RTs (in ms) in regions 4–5 in conditions C1–C6

	C1	**C2**	**C3**	**C4**	**C5**	**C6**
Region 4	365.4	419.9	439.8	381.7	372.6	402.5
Region 5	384.7	476.0	511.5	450.0	397.6	475.3

Table 5. Average RTs (in ms) in regions 4–5 in conditions C7–C12

	C7	**C8**	**C9**	**C10**	**C11**	**C12**
Region 8	366.6	417.8	459.4	417.4	394.3	431.3

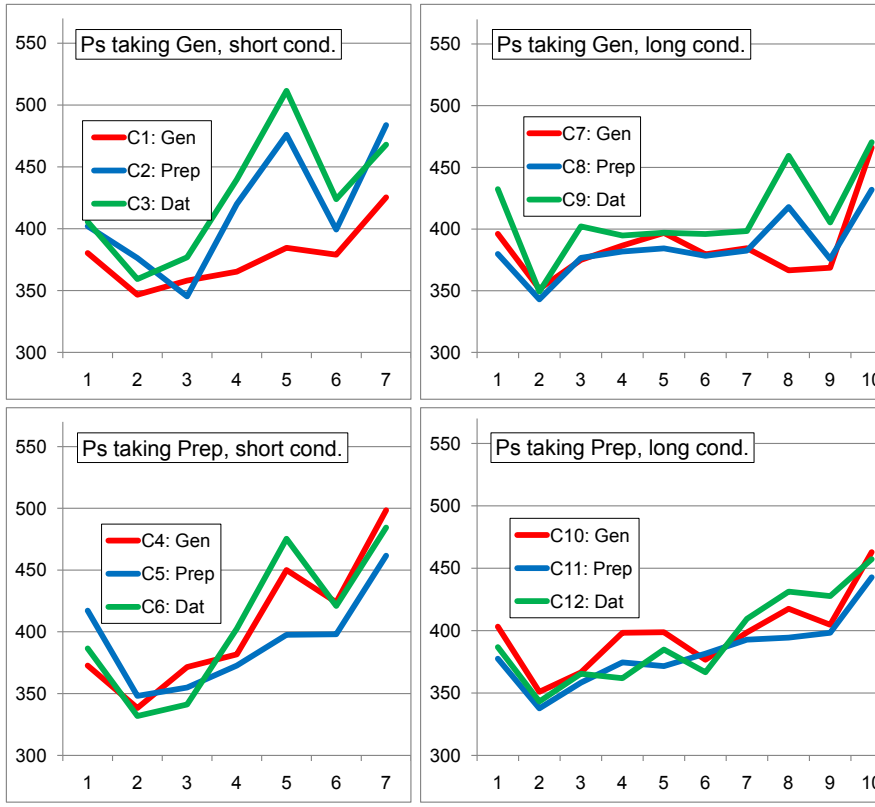


Fig. 2. Average RTs per region (in ms) in different experimental conditions

In the short conditions, the results were the same as in Experiment 1. In the sentences with prepositions selecting Genitive, the difference between C1 and C3 is significant both in region 4 (F1(1,70)=4,01, p=0,05; F2(1,34)=6,96, p=0,01) and region 5 (F1(1,70)=9,15, p<0,01; F2(1,34)=10,05, p<0,01). The difference between C1 and C2 is significant only in region 5 (F1(1,70)=7,67, p=0,01; F2(1,34)=8,11, p=0,01). In region 4, it approaches significance (F1(1,70)=3,06, p=0,08; F2(1,34)=4,11, p=0,05). In the sentences with prepositions selecting Prepositional, the difference between C5 and C4 never reaches significance, while the difference between C5 and C6 is significant in region 5 (F1(1,70)=4,53, p=0,04; F2(1,34)=5,65, p=0,02).

Now let us turn to the long conditions. In the sentences with prepositions selecting Genitive, in region 8 the difference between C7 and C9 is significant (F1(1,70)=10,92, p<0,01; F2(1,34)=11,12, p<0,01) and the difference between C7 and C8 approaches significance (F1(1,70)=3,24, p=0,07; F2(1,34)=4,93, p=0,03). In the sentences with prepositions selecting Prepositional, there are no significant differences in any region.

4. General discussion and conclusions

In total, the effects of all violations are less pronounced and delayed in the long conditions. This is expected: numerous studies demonstrate that readers' ability to detect errors degrades when the syntactic complexity increases. But the difference between two types of errors is visible both in the short and in the long conditions. The errors our study focuses on (the preposition requires case A, the adjective form is ambiguous between cases A and B and the noun appears in case B) are detected later and cause smaller delays than the other errors. For the sentences with prepositions selecting Genitive, this can be proved statistically in Experiments 1 and 2 both in short and in long conditions. In the sentences with prepositions selecting Prepositional, no differences reached significance in the long conditions, but average RTs show the same tendency as in the short conditions in Experiments 1 and 2: in regions 7–9, they are longer in the sentences with Dative nouns than in the sentences with Genitive nouns.

The fact that the observed effect does not depend on linear distance supports the hypothesis that it is similar to agreement attraction. Notably, only one of the existing approaches to attraction, the one advocated by Wagers et al. (2009), can be extended to our case. According to this approach, when a wrong form is produced or encountered (a wrong number on the verb or a wrong case in ours), the speaker or reader comes back to recheck the structure, and certain things may interfere with this process (an attractor noun or an adjective or participle ambiguous for case). Most other authors assume a different mechanism of agreement attraction: the subject NPs erroneously inherits its number or other features from a dependent NP rather than from its head. However, this mechanism is inapplicable to the structures we study.

One of the most important questions is how ambiguous forms are represented so that errors become possible. In comprehension, morphological ambiguity should be resolved by the time the verb or the noun in the wrong form appears. In production, we should know from the very start which case the ambiguous form bears. The fact that errors arise in production and go unnoticed in comprehension nevertheless suggests that morphologically ambiguous forms are deeply interconnected and potentially share some syncretic representation, as it was first suggested by Jakobson (1936). Some of the modern morphological theories adopted this idea, the others did not. Evidently, experimental data can be used to support the former approach.

Finally, in Experiment 1 all effects were more pronounced in the sentences with Genitive. This was also the case in Experiment 2, both in the long and in the short conditions. Thus, this can hardly be accidental, but so far, we have no explanation for this finding.

References

1. *Bock J. K., Miller C. A.* (1991), Broken agreement, *Cognitive Psychology*, Vol. 23, pp. 45–93.
2. *Eberhard K. M., Cutting J. C., Bock J. K.* (2005), Making syntax of sense: Number agreement in sentence production, *Psychological Review*, Vol. 112, pp. 531–559.
3. *Fayol M., Largy P., Lemaire P. (1994), When cognitive overload enhances subject-verb agreement errors: A study in French written language, Quarterly Journal of Experimental Psychology, Vol. 47, pp. 437–464.*
4. *Franck J., Lassi G., Frauenfelder U., Rizzi L.* (2006), Agreement and movement: A syntactic analysis of attraction, *Cognition*, Vol. 101, pp. 173–216.
5. *Franck J., Vigliocco G., Nicol J.* (2002), Attraction in sentence production: The role of syntactic structure, *Language and Cognitive Processes*, Vol. 17, pp. 371–404.
6. *Hartsuiker R., Schriefers H., Bock K., Kikstra G.* (2003). Morphophonological influences on the construction of subject-verb agreement, *Memory and Cognition*, Vol. 31, pp. 1316–1326.
7. *Jakobson R.* (1936), Beitrag zur allgemeinen Kasuslehre. Gesamtbedeutungen der russischen Kasus, *Travaux du Cercle linguistique de Prague,* Vol. 6.
8. *Lorimor H., Bock K., Zalkind E., Sheyman A., Beard R.* (2008), Agreement and attraction in Russian, *Language and Cognitive Processes,* Vol. 23, pp. 769–799.
9. *Pearlmutter N. J., Garnsey S. M., Bock K.* (1999), Agreement processes in sentence comprehension, *Journal of Memory and Language*, Vol. 41, pp. 427–456.
10. *Rusakova, M.* (2009), *Rechevaja realizatsija grammaticheskih èlementov russkogo jazyka* [Speech realization of some grammatical features of Russian]. Habilitation dissertation, St.Petersburg State University.
11. *Vigliocco G., Butterworth B., Garrett, M.* (1996), Subject-verb agreement in Spanish and English: Differences in the role of conceptual constraints, *Cognition,* Vol. 61, pp. 261–298.
12. *Vigliocco G., Butterworth B., Semenza C.* (1995), Constructing subject-verb agreement in speech: The role of semantic and morphological factors, *Journal of Memory and Language,* Vol. 34, pp. 186–215.
13. *Wagers M. W., Lau E. F., Phillips, C.* (2009), Agreement attraction in comprehension: Representations and processes, *Journal of Memory and Language,* Vol. 61, pp. 206–223.
14. *Wilson R., Nicol J.* (1999), Agreement and case marking in Russian: A psycholinguistic investigation of agreement errors in production, in *Proceedings of FASL 8*, Michigan Slavic Publications, Ann Arbor, MI, pp. 314–327.
15. *Yanovich I., Fedorova O.* Subject-verb agreement errors in Russian: Head noun gender effect. *Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2006"* [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006"]. Bekasovo, 2006.

КАКИЕ «СИТУАЦИИ» ОБОЗНАЧАЮТСЯ РУССКИМИ ГЛАГОЛАМИ «ОТЛИЧИТЬ — ОТЛИЧАТЬ»[1]

Соколова Е. Г. (minegot@rambler.ru)
Российский государственный гуманитарный университет, Москва, Россия

Кононенко И. С. (irina_k@cn.ru)
Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

Ключевые слова: глагол, лексическое значение, дискурсивный контекст, интерпретация

WHAT KIND OF "SITUATIONS" UNDERLIE THE RUSSIAN VERBS «OTLICHIT' — OTLICHAT'» (DISCRIMINATE)

Sokolova E. G. (minegot@rambler.ru)
Russian State University for the Humanities, Moscow, Russia

Kononenko I. S. (irina_k@cn.ru)
Institute of Informatics Systems SB RAS, Novosibirsk, Russia

In the paper the method of Discourse Contexts is introduced to describe the semantics and use of one Russian verb pair that corresponds to situation of discrimination. Discrimination implies comparison resulting in differentiation and singling out. Discourse context is understood as complex entity including: (i) abstract Immanent Situation which underlies the use of verbs and represents a configuration of essential elements such as an idea of discrimination and entities involved including subject of discrimination and features of discrimination; (ii) Entity Situation in which the essential elements are classified according to concrete verb; (iii) Grammar Constituent. The analysis of the material of Russian National Corpus gives five types of discourse contexts for verbs *otlichit' — otlichat',* which are presented and exemplified in the paper. Discourse contexts are shown to help catch different meanings and explain semantic peculiarities of *otlichit' — otlichat'.*

Key words: verb, lexical meaning, discourse context, semantic interpretation

---

[1] Работа выполнена при финансовой поддержке Президиума СО РАН (Интеграционный проект № 15/10).

1. Введение

При описании значений глаголов традиционно рассматривается некоторая ментальная ситуация, выражаемая данным глаголом, например, в [1:625] читаем: «Попытаемся выявить главные черты ментальной операции сравнения». При описании ментальной ситуации привлекается наиболее типичная картина, которая извлекается из контекстов употребления глагола. Его значение представляется в виде толкования на ЕЯ- описания «картинки» обозначаемой ситуации. Пример толкования из [2]: «*X похож на Y-а Z-ом* = ‘объект X имеет такое же свойство Z, какое есть у объекта Y, поэтому, когда человек воспринимает объект X или думает о нем, в его сознании присутствует образ объекта Y». Однако ситуации действительности, к которым описываемые глаголы также применимы, разнообразны и не укладываются в «картинку», что убедительно показано в [3]. Замена предметных понятий в толковании на «взаимодействия» в [3] увеличивает референциальную силу толкования, но уменьшает пригодность его для практического анализа.

Формализованные толкования отечественных семантических словарей наследуют еще одну проблему, свойственную толкованиям: раздельное описание значений слова. Так, в МАС [4] для глаголов *отличить* и *отличать* выделяются четыре значения, которые а) не увязаны между собой (общее) и в то же время б) не учитывают специфику ситуаций, лежащих за выделенными значениями (различия), см. табл. 1 в р. 4.

В словарных ресурсах, ориентированных на машинное использование, контекст глагола представляется в виде модели управления или фрейма, но каждая лексема-значение также описывается отдельно. Остается неясным, что лежит в основе группы семантически связанных лексем, на чем основаны сходства и различия их синтаксического поведения. Этот недостаток в определенной мере преодолевается в рамках подхода FrameNet [5], в котором выделяемые типовые контексты представляются в виде фреймов, а фреймы выстраиваются в иерархию, которая должна отражать сходства описываемых ситуаций [6], однако принципы, по которым строится эта иерархия, неясны.

В статье представлен метод **Дискурсивных Контекстов** (ДК), к которому мы пришли в результате описания фреймов для двух русских предикатов, связанных с ситуацией сравнения — *различить(-ать)* и *отличить(-ать)*. Ниже мы рассматриваем метод ДК на примере видовой пары *отличить(-ать)* для контекстов только активных форм глагола, в том числе инфинитива с модальными словами, например, «*позволяет отличить*»[2]. Метод основан на корпусных данных НКРЯ и их ручной обработке. Материал извлекался из НКРЯ путем сплошной выборки (по 200 контекстов для каждого глагола) и целенаправленного поиска контекстов для неочевидных гипотез[3].

---

[2] В предположении, что для остальных глагольных форм ГС можно описать с помощью регулярных соответствий (которое, однако, требует проверки).

[3] К сожалению, формат статьи не позволяет представить более подробное описание глагола *отличить(-ать)*, описание глагола *различить(-ать)*, неизбежно вовлекаемый

2. Метод дискурсивных контекстов

Мы исходим из положения о том, что разнообразные языковые средства создают «значение» высказывания совместно, соответственно, и анализ значения предиката как центра высказывания должен быть многофакторным.

ДК — это сочетание сущностных и грамматических составляющих, создающее определенный контекст, в котором может быть использован конкретный глагол. По методике ДК, конкретный токен глагола в тексте ассоциирован с а) Имманентной Ситуацией **ИС[]**, которая задает абстрактную структуру обозначаемой глаголом ситуации, б) Сущностной Составляющей **СС[]**, которая задает общее значение глагола, в) Грамматической Составляющей **ГС[]**, которая описывает синтаксические функции участвующих сущностей. Роль ГС состоит в реализации сущностных элементов и передаче коммуникативных, устоявшихся в культуре элементов значения. **ДКi[]** — дискурсивные контексты — обобщают часто повторяющиеся сочетания свойств контекста и приблизительно соответствуют по статусу традиционным «значениям» глагола.

Мы полагаем, что различие в значениях двух видовых пар русских глаголов — *отличить(-ать)* и *различить(-ать)* — в первую очередь связано со значением приставок[4], так как глаголы однокоренные и в остальном имеют одинаковый морфемный состав. Мы предлагаем абстрактную структуру ИС[], выделенную нами интуитивно, как общую базу для описания значений этих глаголов. Она представляет идею, а не семантически конкретный процесс или отношение и задает исходные точки, на которых «вырастает» семантика конкретных глаголов и их конкретные «значения».

2.1. Имманентная ситуация «размежевание» по признаку

ДК для глаголов *отличить(-ать)* базируются на ИС[размежевание], которой обозначается операция сравнения с целью диверсификации или выделения, базирующаяся на опыте восприятия — чувственного или с применением специальных познавательных средств (приборов, систем).

ИС[размежевание] включает: идею (операцию) размежевания I (Idea), сущностную составляющую, представленную Множеством размежевания M (Multitude) и сущностями E (Entities), вовлеченными в I, субъекта размежевания S (Subject), обозначающего сущность, осуществляющую размежевание, и признак размежевания F (Feature). Эти понятия обнаруживаются в контекстах:

(1) *Среди них* (M) *я* (S) *сразу отличил запах цветов шиповника* (E).

в материалах НКРЯ диахронический аспект и сравнение семантики двух рассматриваемых видовых пар (в том числе контексты, в которых они являются квазисинонимами).

[4] М. А. Кронгауз пишет в [7], что «разные приставки практически никогда не используются для образования синонимических глаголов».

(2) *Самих скинов* (E), *кроме бритых голов* (F), *можно отличить ещё и по высоким «омоновским» ботинкам* (F).

(3) *Взрослого самца* (E1) *от самки* (E2) *отличают по окраске* (F).

(4) *И одновременно и постоянно Ленина* (E1) *отличала безжалостность* (F), *резкость* (F), *грубость по отношению к политическим противникам* (F).

(5) *Система* (S) *отличает человека* (E1) *от мелких и средних животных* (E2) <…>.

Из пяти перечисленных составляющих ИС[размежевание] свойства четырех (I, M, E и F) формируют контекстные значения глагола в данном дискурсе: лексический предикат определяется характером операции I, собственно значение в традиционном понимании определяется свойствами M, E и F. Субъект распознавания представляет биологические сущности, обладающие сознанием или функцией распознавания, от человека (например, автор в (1)) и животных до микроорганизмов и биовеществ, а также интеллектуальных систем, наделенных функциями распознавания (5). Человек как S не только обладает различительными функциями, но и действует, поэтому в конкретном дискурсе операция размежевания I может представлять собой спектр от ментального процесса до материального (см. ниже п. 3.3). Отсутствие эксплицитного S в дискурсе может означать взгляды множества людей (обобщенного субъекта), что реализуется как обобщенно- или неопределенно-личная форма глагола или безличная конструкция (2), (3).

2.2. Сущностная составляющая для глаголов *отличить(-ать)*

Для рассматриваемой пары глаголов ИС[размежевание] конкретизируется как СС[отличить(-ать)], которая демонстрирует следующие особенности элементов:

- I — сравнение/выделение E1, противопоставленных E2 по признаку F;

- E — сущности, непосредственно вовлеченные в операцию размежевания. Особенность СС[отличить(-ать)] состоит в том, что частью его семантики является наличие наряду с E1 ее контрагента — сущности E2, в противоположность СС[различить(-ать)], для которой в процесс втягивается только E1. Это можно продемонстрировать на возвратной форме: *они* (E1 и E2) *различаются (*друг от друга)* и *они отличаются друг от друга*. При отсутствии противопоставления предлогом, глагол *отличить(-ать)* требует выражения противопоставления пустой «рамкой» процесса «*друг от друга*». Таким образом, в отличие от СС[различить(-ать)], в СС[отличить(-ать)] M и погружение в него E1 — вторично.

- M формируется одним из двух способов:

- имплицитное М: а) как целое (род) на некоторой ступени таксономической иерархии: в (1) это «запахи цветов», в (3) — «конкретнй отряд птиц» (в данном случае попугаев); б) как совокупность экземпляров, объединенных ситуативно по контекстным «осям» (пространственным, временным, социальным и пр.): в (2) это «молодые люди на улице», в (4) — «современные Ленину политики»;

- М как контаминация Е2 и М, т. е. М′ — «остаток» М после выделения Е1.

- F (Feature) — признаки размежевания:

- Fi — характерный для М внутренний признак, например, пол в (3), который, дифференцируя элементы М, позволяет выделить Е1,

- Fe — «внешние» признаки, воспринимаемые органами чувств или устанавливаемые в результате опыта, измерений, эксперимента, теста (цвет, одежда, манера поведения, ситуация и т. д.), по которым можно отличить Е1 от Е2 или отождествить Е1 в среде М. В (3) внешним признаком является «окраска», в (2) — «бритые головы», «высокие «омоновские» ботинки».

2.3. Грамматическая составляющая дискурсивных контекстов для глаголов *отличить(-ать)*

В соответствии с корпусными данными, ГС[*отличить(-ать)*] для активной формы глагола включает следующие основные синтаксические позиции, реализующие элементы СС[отличить(-ать)]:

А. **«X-Acc»** — обязательная позиция **прямого дополнения** у глаголов *отличить(-ать),* средство реализации сущности Е1 (см. примеры (1–5)). Верно и обратное, Е1 всегда выражается прямым дополнением. Аналогом прямого дополнения является сентенциальное дополнение, которое возможно в этой позиции только в виде косвенного вопроса, что, как отмечено в [8], характерно для класса глаголов восприятия:

(6) *Знаете, как мы* (S) *отличаем,* <u>*чьи окопы нащупали*</u> (Е1)*?*

В. **«*От* Y-Gen»** — средство реализации контрагента операции сравнения, как Е2 в (3) и (5) или М′ в (7). М′ выражается родовым именем М с прилагательным, обозначающим «остаточность, невыделенность» (*обычные, остальные, другие, прочие, предыдущие, аналогичные*):

(7) *Есть только одна тонкость, но именно она* (Fe) *существенно отличает* <u>*бонсаи*</u> (Е1) *от* <u>*обычных растений*</u> (М′).

C. **«*По* Z-Dat»** является способом реализации признака Fe: (2), (3), (8).

(8) *Отличить их* (E1) *легко* *по милиционеру, стоящему рядом и проверяющему содержание их документов* (Fe) .

D. **«Локативное» дополнение «*среди/из* W-Gen», «*в* W-Loc», «*между* W-Instr»** реализует M: (1), (8), (9)

(9) *Нужно два помощника местным, а бригадир* (S) *изо всей шефской массы* (M) *только нас двоих* (E1) *и отличает.*

E. **«Инструментальное» дополнение «V-Instr».** Ожидаемая реализация в этой позиции сущности СРЕДСТВО как инструмента у активной формы глагола не подтверждается корпусными данными. В этой позиции реализуется сущность ОРГАН, которая является результатом расщепления S:

(10) *<…>Наполеон* (S) *простым глазом* (ОРГАН) *мог в нашем войске* (M) *отличать конного* (E1) *от пешего* (E2).

F. **Инструментальное дополнение «*С помощью* V-Gen»** представляет
- СРЕДСТВО (предмет):

(11) *Адам и Ева научились отличать добро* (E1) *от зла* (E2) *с помощью яблока с древа Познания* (СРЕДСТВО).

- МЕТОД (правила), который, в отличие от признака, не имеет непосредственного отношения к объекту размежевания:

(12) *Отличить подделку* (E1) *можно только с помощью химико-фармацевтического анализа* (МЕТОД).

G. **Подлежащее** — позиция Актора и грамматической темы при нейтральном порядке слов:
- Эксплицитный S. Примеры (1), (5). Верно и обратное: эксплицитный S у активной формы глагола реализуется только как подлежащее.
- ОРГАН как неотъемлемая часть S, при этом сам S часто не выражен:

(13) *Меньшевичку* (E1) *опытный глаз* (ОРГАН) *тотчас отличит от большевички* (E2).

- Fe. При этом, если признак не является старой, уже введенной, информацией (ср. (14)), обязательна инверсия подлежащего, так как темой текста остается носитель признака E1, выражаемый всегда прямым дополнением.

(14) *Способность входить с людьми в общение* (Fe) *и в самом деле его* (E1) *отличала.*

- МЕТОД (только в конструкции с модальным глаголом «позволять»:

(15) *Визуальная диагностика* (МЕТОД) *не позволяет отличить эти минералы* (E) *друг от друга.*

- СРЕДСТВО

(16) *В свою очередь каждый из этих приборов* (СРЕДСТВО) *анализирует массу сходных раздражителей и отличает их* (E) *друг от друга.*

ГС[*отличить(-ать)*] включает ряд дополнительных элементов, из которых отметим а) «ресурсоемкость» (*легко, с трудом, без труда*), предполагающий наличие субъекта (в т. ч. обобщенного субъекта) — пример (9); б) «степень» (*резко, существенно, заметно),* характерный для контекста без выраженного субъекта (см. п. 3.4) — пример (7).

3. Дискурсивные контексты

Анализ конкретных контекстов в НКРЯ позволил выделить следующие типы ДК для глаголов *отличить(-ать)*:

3.1. ДК1 «Сравнение»

Минимальный ДК1 состоит из S, E1 и E2. S — объект, наделенный сознанием. S могут представлять также неопределенно-личная и обобщенно-личная формы (*мы*, *-ют*), модальные контексты (*следует, можно*), фиксирующие общепринятую точку зрения или общее положение вещей.

M имплицитно. Оно должно «охватывать» E1 и E2. Его роль — определение условий, в которых возникает основание для сравнения: родовая однородность (17), похожесть (18) E1 и E2. E1 и E2 а) противопоставлены по видовому признаку и принадлежат одному вышестоящему в онтологической иерархии понятию — роду: *музыкальные инструменты* в (17); б) видовые, не принадлежащие одному роду, но внешне похожие в определенных условиях (18); в) экземпляры одного вида (19).

(17) *Человек, закончивший музыкальную школу,* (S) *должен отличать гобой* (E1) *от фагота* (E2).

(18) *Некоторых морских коньков* (E1)*, например, с трудом можно отличить от кораллов* (E2).

(19) *С той самой, которая* (S) *не может отличить Бальзака* (E1) *от Флобера* (E2).

Комбинаторное сравнение. Если Fe определяется не для пары сущностей E1 и E2, а для множества однородных элементов M, позицию прямого дополнения занимает произвольный элемент E (M), а идея противопоставления элементов по признаку выражается пустой рамкой:

(20) *Отличить* <u>*один дом* (E1) *от другого* (E2)</u> *можно, пожалуй, только по номеру да ещё по оформлению маленького кусочка земли перед фасадом* (Fe).

3.2. ДК2 «Выделение»

S — существо, обладающее сознанием и осознающее свой выбор. E1 — вид или экземпляр. S «знает» E1 по некоторому Fe, ортогональному иерархии M. Например, в (22) Fi формирует множество пассажиров M, а Fe — профессия E1, которая не идентифицирует E1 внешне среди других людей. По остальным параметрам все E (пассажиры) в вагоне в данной ситуации идентичны. Fe может быть сформулирован и выражен в виде текста или маркера (знака, помечающего E1), или представлять собой экспертное, «индивидуальное, образное» знание, полученное S в результате опыта.

M — однородное, дискретное и представлено как целое локативным дополнением (21) или как E2 — «от M´» (22):

(21) *Она* (S) *уже почти безошибочно умела отличить шпиона* (E1) *в уличной толпе* (M).

(22) *Наметанный глаз сразу отличил их* (E1) *от обычных пассажиров* (M´).

3.3. ДК3 «Выделение как материальный процесс»

Если в ДК1 и ДК2 глагол имеет скорее ментальный тип, то ДК3 — это материальная ипостась I, в которой действует S. Обязательно присутствуют: S — лицо, демонстрирующее или совершающее выделение и E1 — социально значимая сущность, обычно человек, но может быть и объект. M обычно имплицитно.

Сфера: социальная

3.3.1. ДК3.1 «поведенческий»

S ведет себя особым образом по отношению к E1 (лицо) по сравнению с другими «E», потому что E1 обладает привлекательным для S признаком Fi. Соответственно, предполагается импликативная часть: установление особых отношений между S и E1, изменение социального статуса E1.

(23) *Теперь она знала наверняка, что Господь* (S) *отличил её* (E1) *из огромного людского множества* (M)…

(24) *И во все времена власть* (S) *особо отличала тех, кто подавляет внутреннего врага* (E1)

(25) *Росси* (S) *сразу отличил Митю* (E1) *за недюжинные способности, прямодушный ум* (Fi) *и, узнав про мечту его выкупить свою невесту, стал ему передавать доходную работу.*

Синоним: (зависит от Fi) «*уважать* (за личностные качества)», «*ценить* (за профессиональные, деловые качества)».

Синоним: *выделить*

3.3.2. ДК3.2 «Дарование отличительного признака»

S — лицо, E1 — лицо или социально важный экземплярный объект, Fe — награда: чин, имя, звание и т. п.:

(26) *Сандецкий* (S) *благоволил к П. и отличил его* (E1) *чином генерал-лейтенанта* (Fe) *и орденом* (Fe).

(27) *Я* (S) *отличил сей остров* (E1) *наименованием острова великого князя Александра* (Fe).

(28) *Прибавьте к тому всю раздражительность поэта, его природную наклонность к роскоши, <…>, которыми* (Fe)*, наперекор обществу, природа* (S) *любит отличать своего собственного аристократа* (E1)*!*

Синоним: *~наградить*

3.4. ДК4 «Авторская констатация выделенности»

I как процесс отсутствует, соответственно, отсутствует и S как участник процесса I. Есть известный автору текста Fe (или параметр), свойственный E1 и не свойственный другим E. Сам Автор присутствует в тексте в роли наблюдателя (N) — «закадровой» фигуры, констатирующей факт отличия E1 от других «E». Таким образом, глагол представляет отношение между E1 и признаком, по которому E1 выделяется из остальных E. Контекст образуется только формами несовершенного вида, характерного для «отношений».

E1 часто вводится в текст в виде названия, описания, предшествующего ситуации, обозначенной глаголом *отличить* (часто в предыдущей клаузе того же предложения): (29), (30). Возможна также интродукция в контексте *отличить,* с нарушением характерной в этом случае инверсии подлежащего (31).

(29) *Шопенгауэр сказал, что чувство сострадания к товарищам по жизни — это единственное качество, которое* (Fe) *отличает человека* (E1) *от других участников биологического цикла* (M′).

(30) *Специально для китайского рынка разработана трехобъемная версия модели 307: ее* (E1) *отличает самый большой среди одноклассников багажник объемом 506 л.* (Fe)

(31) *Сдержанность формы* (Fe) *и намеренный уход от специальных эффектов* (Fe) *отличают хронограф Mercedes* (E1).

Синоним: *выделять*
Сфера: распространен в научной, рекламной и «технологической» сферах.

4. Заключение и выводы

Анализ конфигураций ДК позволил выявить следующие значения исследуемых глаголов:

ДК1 — сравнение E1 с противопоставленным ему E2;
ДК2 — выделение E1 из M с E1, противопоставленным M’;
ДК3.1 — выделение E1 поведением S
ДК3.2 — выделение E1 действием S
ДК4 — констатация отношения между выделительным признаком Fe и E1.

Табл. 1. Сравнение значений МАС и ДК для глаголов *отличить(-ать)*

	Значения МАС *отличить*	**ДК**	**Значения МАС *отличать***
1	Распознать какой-л. предмет, явление и т. п. среди других	ДК1 и ДК2	Распознать какой-л. предмет, явление и т. п. среди других
2	Наградой выделить из числа других.	ДК3.2	Наградой выделить из числа других.
3		ДК4	Служить отличительным признаком, являться отличительной, характерной особенностью кого-, чего-л.
4		ДК3.1 — несов. и сов. в.[5]	Проявлять особое внимание, интерес к кому-л., оказывать предпочтение кому-л.

[5] Точка зрения МАС относительно этого значения, указанного только для несовершенного вида, материалом не подтвердилась, ср. пример (23).

Таким образом, методом ДК все значения МАС обнаружены, 1 — уточнено, разделено на два; 2 уточнено; 3 уточнено, 4 — уточнено. Также объяснено употребление «друг друга» в комбинаторной разновидности ДК1.

Метод ДК предложен нами как метод многофакторного анализа семантики предикатов для преодоления коллапса, в котором оказалась объяснительная лексическая семантика. Он потенциально ориентирован на такие приложения, как системы генерации текстов из неязыковых данных. Метод ДК основывается на семантике сущностей, а не процесса и его ролей, позволяет увидеть анатомию значения предиката: его абстрактную структуру (ИС), специфику данного глагола как целой языковой единицы (СС) и отдельные значения глагола в виде ДК, сложившихся в языковой практике. А также учесть коммуникативную организацию предложения и связать значения со сферами употребления, способствует созданию «когнитивного метаязыка» для описания ИС. Мы сделали только первый шаг, во многом интуитивно, но нам кажется, что современный уровень развития лингвистики позволяет разрабатывать методы анализа лексической семантики высокой точности. Продолжение исследования предполагает распространение метода ДК на другие предикаты для уточнения его возможностей в компьютерной лингвистике.

Литература

1. *Урысон Е. В.* (2004) Ситуация сравнения и ее выражение в языке // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции «Диалог 2004». — М.: Наука, 2004. с. 624–637.
2. *Апресян Ю. Д.* и др. (1997) Новый объяснительный словарь синонимов русского языка. Вып. 1. М., 1997.
3. *Кошелев А. Д.* (1996) Референциальный подход к анализу языковых значений //Московский лингвистический альманах, выпуск 1. — М.: Школа «Языки русской культуры», 1996. с. 82–185.
4. *Малый академический словарь* (МАС). Доступ: http://mas-dict.narod.ru/
5. *Framenet* [Electronic resource].
   Available at: https://framenet.icsi.berkeley.edu/fndrupal/about
6. *Baker, Collin F., Fillmore, Charles J., and Lowe, John B.* (1998): The Berkeley FrameNet project. in Proceedings of the COLING-ACL'98, Montreal, Canada. 1998. pp. 86–90.
7. *Кронгауз М. А.* (1998) Приставки и глаголы в русском языке: семантическая грамматика. М. Ж. Школа «ЯЗЫКИ РУССКОЙ КУЛЬТУРЫ». 1998.
8. *Падучева Е. В.* Прямая и косвенная диатеза ментального глагола: корпусное исследование // Труды международной конференции «Корпусная лингвистика — 2008». — СПб, 2008, с. 303–317.

References

1. *Apresjan Ju. D.* et al. (1997), New Explanatory Dictionary of Russian Synonyms [Novyj ob”jasnitel’nyj slovar’ sinonimov russkogo jazyka] Vol.1, Studia philologica, Moscow.
2. *Baker, Collin F., Fillmore, Charles J., and Lowe, John B.* (1998), The Berkeley FrameNet project. In Proceedings of the COLING-ACL’98, Montreal, Canada. pp. 86–90.
3. *Framenet* [Electronic resource]. Available at: https://framenet.icsi.berkeley.edu/fndrupal/about
4. *FrameNet project.* (1998), In Proceedings of the COLING-ACL’98, Montreal, Canada. pp. 86–90.
5. *Koshelev A. D.* (1996), Referential approach to the analysis of language meanings [Referencial’nyj podhod k analizu jazykovyh znachenij]. Moskovskij lingvisticheskij al’manah [Moscow linguistic collection]. Moscow, Vol. 1. pp. 82–185.
6. *Krongauz M. A.* (1998), Prefixes and Verbs in Russian: Semantic Grammar [Pristavki i glagoly v russkom jazyke: semanticheskaja grammatika], Shkola «Jazyki russkoj kul’tury», Moscow.
7. *Small Academic Dictionary of Russian Language* [Malyj akademicheskij slovar’]. Available at: http://mas-dict.narod.ru/
8. *Paducheva E. V.* (2008), Direct and indirect diathesis of mental verb: corpus study [Prjamaja i kosvennaja diateza mental’nogo glagola: korpusnoe issledovanie], Proc. of Int. Conf. “Corpus linguistics — 2008” [Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika — 2008»]. SPb, pp. 303–317.
9. *Uryson E. V.* (2004), Situation of comparison and its expression in language [Situacija sravnenija i ee vyrazhenie v jazyke]. Proc. of Int. Conf. “Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ‚Dialog 2004‘” [Komp’juternaja lingvistika i intellektual’nye tehnologii: trudy Mezhdunarodnoj konferencii «Dialog 2004»]. Moscow. pp. 624–637.

БАЗА ДАННЫХ «ЯЗЫКИ МИРА» И ЕЕ ПРИМЕНЕНИЯ. СОВРЕМЕННОЕ СОСТОЯНИЕ

Соловьев В. Д. (maki.solovyev@mail.ru)
КФУ, Казань, Россия

Поляков В. Н. (pvn-65@mail.ru)
НИТУ МИСиС, Москва, Россия

Ключевые слова: лингвистические базы данных, типология, квантитативные методы, ареальная лингвистика, языки мира

DATABASE “LANGUAGES OF THE WORLD” AND IT’S APPLICATION. STATE OF THE ART

Solovyev V. D. (maki.solovyev@mail.ru)
Kazan Federal University, Kazan, Russia

Polyakov V. N. (pvn-65@mail.ru)
National University of Science and Technology “MISiS”, Moscow, Russia

The article is dedicated to the largest digital resource in the world that contains a uniform description of language grammars — typological database “Languages of the World” (“Jazyki Mira”). There is information on the contents of the database, the programs for data procession. The database “Languages of the world” has three main areas of application: it can be used for quantitative researches, as a reference linguistic resource and for educational purposes. We give examples of database application in scientific researches in typology and areal linguistics. The examples demonstrate new opportunities of studying such questions as stability of grammatical features, liability to borrowing, typological and areal classification of languages. “Languages of the World” is compared with another famous typological database WALS.

Keywords: linguistic databases, typology, quantitative methods, areal linguistics, languages of the world, Jazyki Mira

1. Introduction

At the turn of the century there appeared various digital linguistic resources aimed at supporting of linguistic researches. An important place among them belongs to typological databases (TDB) that contain the descriptions of formalized grammar features of the languages. The development of this area began with small databases (DB) dedicated to a rather limited number of features, which contained the description of a small number of languages. Examples of such databases and a general review of TDB application can be found in [14, 16].

The new stages of TDB development began with the appearance of The World Atlas of Language Structures (WALS) [6] and database “Languages of the World” («Jazyki Mira»). The latter was created in Institute of Linguistics of Russian Academy of Science (IL RAS) on the base of a series of monographs of the same name (16 volumes). The first publications on this database are [7, 12]. The database is available in the Internet at http://dblang2008.narod.ru/, www.dblang.ru.

WALS and “Languages of the World” can be called big typological databases; each contains over 1 million bits of information. WALS describes over 2,500 languages by 142 features (128 of them are grammatical ones), and each of them has one of a few meanings: from 2 to 9. “Languages of the World” has the descriptions of 315 languages by 3,821 binary features. Both databases embrace all parts of grammar.

Examples of features: free word order, presence of ergative and absolutive cases, presence of exactly 5 monophthongs, etc. The set of features was formed as a result of systematic study of language grammars with the initial development of a formalized model of grammar description, and it was replenished after “Languages of the World” monograph had been written. The aim of the development of the set of features was the most detailed and precise description of grammar. The set of features is open, and it can be broadened when new languages are added.

TDB were initially created as reference books with a user-friendly interface, which helped quickly find the necessary information. But it soon turned out that TDB give us essentially new opportunities to study grammars of the languages by applying mathematical (including statistical) and computational methods. Many phenomena, which were until now regarded only on the qualitative level and on the base of separate examples, can now be studies by quantitative methods and with use of huge arrays of information. An important aspect of such studies is their objective character based on the application of strict mathematical methods. There are several types of the questions, which can be answered with help of TDB.

1. How homogeneous is this or that language areal? Can it be considered a language union? TDB allow applying of quantitative methods in areal linguistics for the estimation of the level of language proximity.
2. How were linguistic features spread during the spreading of the humanity and linguistic evolution? J. Nichols’ [10] conducted her pioneer researches in this direction on a very limited data access. Modern TDB can help define more exactly many aspects of humanity settlement.
3. Linguistic dynamics: what is the speed of grammar changing? What parts of grammar change faster?

4. What grammar features are easier to borrow during linguistic contacts?
5. Typological classification of the languages.

The article contains the description of the DB "Languages of the World" and of the program instruments it uses, it also gives examples of its application.

2. Structure and software of the database "Languages of the World"

2.1. Composition and structure

DB "Languages of the World" presents the following language families: Austro-Asiatic, Austronesian, Altaic, Afroasiatic, Indo-European, Kartvelian, North Caucasian, Sino-Tibetan, Uralic, Hurro-Urartian, Chukchi-Kamchatkan, Eskimo-Aleut and several isolate languages. The wide range of linguistic families, presented in the DB, justifies the name "Languages of the World". The database is constantly expanded as new monographs of the series are published. This work is conducted in the sector of areal linguistic of IL RAS under the guidance of A. A. Kibrik[1]. There are 10 more volumes planned for publication.

The DB has a genetic reference, which was developed in IL. In general, it corresponds to the classification from [2]. It contains 4 levels: families, branches, groups, subgroups.

The languages are described by a list of features and categories, which was called "Abstract model" in [7], and includes 3,821 features. The description of each language, i.e. a set of meanings of the features, is called its abstract. All languages' abstracts can be found at the web-site of the project: www.dblang.ru. The features are organized in a hierarchy. The top level of the hierarchy: 1.1. Phonemic structure. 1.2. Prosodic phenomena. 1.3. Phonetically motivated processes. 1.4. Syllable. 2.1. Phonological structure. 2.2. Phonological oppositions of morphological categories. 2.3. Phonologically motivated alternations. 3.0. Morphological type of the language. 3.1. Criteria of definition of parts of speech. 3.2. Nominal classifications. 3.3. Number. 3.4. Case meanings. 3.5. Verbal categories. 3.6. Deictic categories. 3.7. Parts of speech. 4.0. Paradigms. 5.1. Word form structure. 5.2. Word formation. 5.3. Simple sentence. 5.4. Composite sentence.

The abstract of a language contains about 300–350 features. 50 languages can be considered poorly described: their abstracts contain less than 200 features. The Russian language is obviously over-described: 536 features.

While using the DB we found mistakes in the data. An expertise was conducted for 30 randomly chosen languages in order to reveal them. On average, less than 3% of feature values were wrong. These mistakes have a different character and are mainly connected to the indistinctness of defining linguistic categories and subjectivism of the researches who described the language. We believe that at the current

[1] http://iling-ran.ru/beta/departments/typol_compar/areal

level of linguistic data formalization it is impossible to eliminate all disagreements in different experts’ interpretations. The database WALS also contains mistakes and contradictions, but they do not influence the results of statistical calculations, as the latter proceed big data arrays, and the mistakes are leveled.

The comparison of WALS and “Languages of the World”, conducted in [13], included building of phylogenetic trees for the same set of languages. It revealed a more serious problem of WALS resource when it is used for statistical calculations: a big number of gaps in the data. On average, languages in WALS are described on less than one third of the features. As a result, due to the lack of data, non-relative languages groundlessly drew closer. “Languages of the World” has a great advantage, as the languages (except for small number of the little-studied ones) are completely described, i.e. by all features.

2.2. Software

The software of the DB “Languages of the World” consists of a nucleus and research tools. The software of the DB “Languages of the World” solves the following tasks:

1) formation and management of the model and abstracts of the database;
2) search for information;
3) binary comparison of abstracts.

The module of binary comparison of abstracts shows lists of common features for the given pair, and also a list of features that are present only in one of the two languages.

The DB “Languages of the World” exists in form of a Web-version, Windows-version and Excel-version. The Windows-version of the DB is a 32-bit application, written in Delphi Pascal (version 7). Borland Database Engine is used as DBMS. The workspace is: Windows 95/98/2000/NT/XP. The volume of installation: 17.4 Mb. The volume of the program and the DB: 18.8 Mb.

The Excel-version gives easy-to-use opportunities for statistical calculations with help of in-circuit tools. Except the nucleus tools, some research tools were created for quantitative investigations. They are:

- Similarity program, for calculation of the level of language proximity;
- LangFam program, for calculation of language portraits of families of languages and revelation of genetic markers.

Standard phylogenetic algorithms, programs for multidimensional scaling and principal component analysis can be applied.

The easiest way to calculate the level of language proximity is Hamming’s metrics (number of unmatched features). Besides Hamming’s metrics, Similarity program provides the calculation of a few other studied measures of language proximity. Moreover, Similarity is an adjustable program. It allows varying different parameters, e.g. choosing groups of features, according to which the calculation of the distance between languages will be implemented. This program helped revealing metrics of calculation of language proximity that describe genetic trees with a high level of precision (up to 80% of match with traditional views) [5].

LangFam program was written in VBA language; it is designed for calculation the frequency of features by all families of languages of all genealogical level that are present in the DB, and by all DB in general. LangFam program helped revealing such phenomenon in the development of the languages as typological shift. It main point is the following: during the linguistic evolution and contacts the feature space is partially "polarized" (most rare features become even more rare, and most widely spread features — even more spread) [11].

The database is constantly replenished with new information and renewed. The version of 2013 is written in C# with use of ASP.NET library and, thus, it requires Microsoft.NET Framework 2.0 and higher. There is a possibility of uploading abstracts from text files. The total volume of installation version is 99 Mb[2]. The program gives a more user-friendly interface for viewing of the main data of the base, it includes annotations of features, examples and references to the source article about the language in the encyclopedia (quantized into pdf). It has more powerful search facilities than the previous version. It also includes "Glossary", which gives a definition of all terms of the language description model; genetic reference; geographic reference, which contains the name of the area where the language is used and geographic coordinates of its center (according to UNESCO's atlas); English translation of features; English names of the language; language code according to ISO 639-2 (Ethnologue, www.ethnologue.com).

3. Examples of application in scientific researches

3.1. Typology

3.1.1. Typological classifications of languages

Evidently, the first serious attempt to classify the languages by their structure is morphological classification, developed in the early 19-th century in Schlegel's, Humboldt's and Schleicher's works. This classification still remains meaningful, but it takes in consideration only one aspect of the linguistic structure: the way morphemes are joined, so, languages, which belong to one class, according to this classification, can radically differ from each other in other aspects.

Other existing typological classifications of languages are also based on one or a few features. It remains unclear, whether holistic classification of languages is possible. It divides all languages into several groups, so that languages within one group are typological homogeneous, and there are sufficient typological differences between the languages from different groups, and they differ in a wide range of features that embraces all main levels of the language.

---

[2] The increase of the volume is due to the big number of graphic materials in the articles of the encyclopedia

With the appearance of big typological databases, like WALS and “Languages of the World”, it becomes possible to build classifications that consider hundreds and thousands of features at the same time.

For the first experiment we chose 27 languages that represent all families and isolate languages from our DB. We apply the well-known phylogenetic algorithm NeighborNet, which was developed in bioinformatics. The results are presents in Pic. 1, where close position of the languages means shorter distance between them, i.e. means bigger typological similarity between them.

In general, the arrangement of the languages in pic. 1 is rather even. Nevertheless, in pic. 1 we can see, though not very clear, 5 main clusters of languages by typological similarity: Indo-European, Uralic-Altaic, Caucasian (probably, with Chukchi and Ket), Far Eastern (several isolate languages) and Afroasiatic. Noteworthy, typological proximity correlates well with linguistic kinship and areal proximity. Thus, Indo-European languages proved to be typologically close, despite the fact that they dispersed from Proto-Indo-European at least 6 thousand years ago and are now spread over a big territory. Nevertheless, during this time they have not acquired such features as vowel harmony (which is characteristic of Turkic languages), incorporation (characteristic of Chukchi-Kamchatkan languages), etc. As a result, typologically, modern Indo-European distinctively differ from Turkic languages, for example. The differences between Proto-Indo-European and Proto-Turkic languages have not been smoothed during this time. Caucasian languages proved to be typologically close, despite their attribution to three different families. This indicates the importance for typological proximity of not only common origin, but also of long-term contacts (several thousand years for Caucasus).

Separate common features can be found in non-relative languages that are located far from each other. Thus, “qualitativeness” (way of action, №2122 in the DB) is found in Aleut and Ethiopian, but it is absent in other languages of the region. Such cases of parallel evolution are rare and they do not influence the general image.

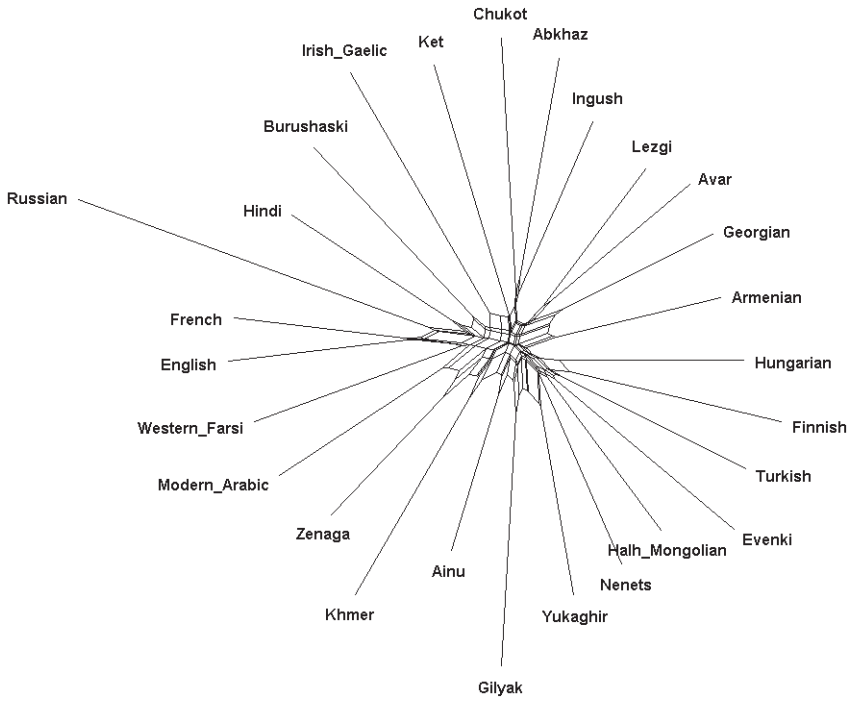


Fig. 1. Languages of Africa and Eurasia according to the data of "Languages of the World"

We shall note that such diagrams do not have an absolute character. Some languages are very strangely positioned in diagrams of this type. For example, the Irish language (Celtic branch of Indo-European family) is placed between Ket and Burushaski. In fact, these languages are not typologically close, they are not related and are geographically very far from each other. The possible reason of such placement if insufficiency of the given method of graphic representation of information for absolutely precise reflection of typological proximity between all pairs of languages. From the mathematical point of view, the DB "Languages of the World" represent languages as points in 3821-dimentional space of features. At the same time, the diagram built by NeighborNet is equivalent to 1-dimentional representation (in a circle). Obviously, when 3821-dimentional space is rolled into 1-dimentional space, there can be distortions.

With help of Similarity program one can find out that Irish is still closer to English (the distance is 285) and Persian (280) that to Burushaski (301). Thus, as well as other tools of computer linguistics (like ancient texts recognition), phylogenetic algorithms require certain post-editing. Nevertheless, these phylogenetic algorithms are more and more widely applied in comparative linguistics, as they allow quickly receiving a rather good result. This method can be compared with Greenberg's method of mass comparison. Articles with use of typological databases and phylogenetic algorithms are published in leading journals, such as Language and Science. A number

of works note that in cases when questions of linguistic kinship have been reliably defined, it turns out that the results of phylogenetic algorithms coincide with the stated ones in 80% of cases.

We should note, even Indo-European languages have not been completely studied from the point of view of their evolution tee reconstruction. For example, [2] enumerates 136 modern Indo-European languages. If their evolution tree was completely studied (i.e. was binary), it would contain 135 tops, which would conform to protolanguages. But the tree presented in [2] contains only 26 tops, i.e. less than a fifth part.

3.1.2. Stability of grammatical features

Let us study the question of the stability of grammatical features. The key idea in the estimation of stability consists in comparison of the prevalence of a feature among related and non-related languages. The biggest part of researches are based on this idea and specify it. The first quantitative researches of stability were conducted by Nichols [10]. She suggested several variants of stability measures. Unfortunately, she did not have a big typological database, which prevented a wide verification (with a big number of languages and features) and spread of her approach.

In [15] there are examples of defining 4 measures of stability of grammatical features. The first one was suggested by Nichols (measure 3 in [10]), the second one was suggested by Wichmann and coauthors [17], the third measure was suggested by Maslova [9], the fourth measure was suggested by one of the authors of the present article, and it is the only measure that realizes the idea of calculation of the number of changes of a feature values during the evolution. Philogenetic algorithms of evolution trees reconstruction are often used for it.

The comparison of these measures on the material of the DB “Languages of the World” showed that there is good correlation between the first and the fourth measures, and also between the second and the third. It is shown that a generalized measure of stability received on the basis of all four measures, in most cases coincide with the qualitative evaluation, previously published in typological literature.

The comparison of measure 2 for WALS and “Languages of the World” was conducted in [15, 1]. There were chosen 23 features of WALS (or, to be more precise, values of features) that match or are very similar in WALS and in “Languages of the World”. In most cases data on the stability of features, calculated by both bases, match or are very close. Reasons for the cases when a considerable mismatch takes place require separate study.

3.1.3. Borrowing of features

TDB allow to systematically studying the inclinations of features and groups of features to borrowings. In [3] B. Comrie used WALS to evaluate the number of matching and mismatching features for three languages: Egyptian, Maltese and Spanish, considering that Maltese is related to Egyptian, but it was in contact with Spanish for a long time. We studied groups of three languages with an analogous structure for “Languages of the World”. For the group of Hungarian, Romanian and Khanty with a contact situation Hungarian-Romanian and genetic relationship Hungarian-Khanty we looked at the features common for Hungarian and Romanian and

different for Khanty. One can suppose that they (at least, part of them) were borrowed from the Romanian language to Hungarian. It turned out that among phonological features they made 20.8%, among morphological — 21.6%, and among syntactical — 19.6%. We studied two more groups with the Hungarian language: Hungarian-Slovak-Khanty and Hungarian-German-Khanty. The averaging of the results gave us the following data: among phonological features the percentage of presumable borrowed one in Hungarian made 20.9%, among morphological features — 19.8%, among syntactical features — 17.4%. We received similar results for other regions and language families. For the group of Finnish, Swedish, Komi-Zyrian the corresponding numbers are: 21.7%, 15.6%, 14.6%; for the group of Polish, German, Macedonian: 15.5%, 13.9%, 7.5%, for the group of Tatar, Mari, Turkish: 19.3%, 16.6%, 24.3%.

It is generally accepted that phonology is easier to borrow than morphology and syntax. We showed that, despite it being true, the difference (especially between phonology and morphology) is not great at all, and in some certain situations this regularity may not be followed at all.

3.2. Areal linguistics

During the past years areal linguistics has become one of the most dynamically developing branches of linguistics. A. E. Kibrik [8] noted that studying of areal connection is especially important for the explanatory approach, which will allow linguistics to move closely to understanding of the essence of the language. But one of the central notions of areal linguistics — concept of language union (LU) is being criticized (e.g. L. Campbell in [16]), due to vagueness and indistinctness of the definition of LU. Use of TDB allows making the definition more precise.

We shall introduce a quantitative measure of convergence of languages in a regional community. Let us assume that there are two groups of languages G1 and G2 from two families S1 and S2 in the studied region. We will calculate an average distance d(G1, G2) between languages from groups G1 and G2 and an average distance d(S1, S2) between languages from families S1 and S2. We will call the difference $R = d(S1, S2) - d(G1, G2)$ measure of convergence of languages in the region. When $R = 0$ (and $R < 0$) there are no grounds to postulate any presence of LC in the region. The bigger R is, the stronger the LC is, i.e. the languages of the region drew closer as a result of borrowings. Similar calculations can be made for languages from several families.

For testing we shall apply the suggested method to the classic LC — Balkan. Balkan LC traditionally includes South Slavonic and Balkan-Romanic languages, and also Albanian and Greek. Unfortunately, Albanian and Greek are not presented in “Languages of the World”. So, we apply the approach to the remaining two groups that belong to the Slavonic and Romanic branches of the Indo-European family.

Calculations show that the average distance between Balkan-Romanic and South Slavonic languages equal 237, while the distance between Romanic and Slavonic languages in general equal 243. As it was presumed, the distance between all Romanic and all Slavonic languages is bigger. Although the difference $243 - 237 = 6$ does not seem so big, it still shows higher typological proximity between Balkan-Romanic and

South Slavonic languages, which cannot be explained by their relationship, as their genealogical proximity is the same as between Romanic and Slavonic languages.

Similar results were received for Volga region, which is inhabited by speakers of Turkic and Finnish-Ugric languages. However, further broader studies may require us to specify the suggested simple formula.

4. Conclusion

The aim of this article was not only to present results, which are still very far from final, but to outline a general way one which one can receive new results of quantitative character in typology and areal linguistic with use of typological databases. There are a lot of methodological problems to be solved on this way.

TDB help receive objective numerical estimations of such characteristics of grammatical features as level of stability, inclination to borrowings. These data can be used in researches on the language evolution. Although it is still impossible to give statistically reliable temporal estimations of feature stability, probably, the most stable grammatical features appeared dozens of thousands of years ago, which allows getting deep inside the history of languages.

Educational program “Databases for Typological and Comparative Researches” was developed on the material of the DB “Languages of the World”. This program was tough as an optional course at the philological faculty of Moscow State University (Department of Theoretical and Applied Linguistics, Moscow) and at linguistics department of South Ural State University (Chelyabinsk).

References

1. *Belyaev O.* (2009) Stability of language features: a comparison of the WALS and JM typological databases. Proceedings of the Int. Conf. “Cognitive Modeling in Linguistics”, FCCL, available at: http://fccl.ksu.ru/conf_CML_2008/jm-wals-stab-2.doc
2. *Burlak C. A., Starostin C. A.* (2005) Vvedenie v lingvisticheskuju komparativistiku [Introduction to comparativistics]. Moscow, Academija.
3. *Comrie B.* (2009) Maltese and the World Atlas of Language Structures. Introducing Maltese Linguistics. Amsterdam, Philadelphia, Benjamins, pp. 3–11.
4. *Everaert M., Musgrave S., Dimitriadis A.* (Eds.) (2009) The Use of Databases in Cross-Linguistic Studies, Berlin, Mouton de Gruyter.
5. *Gusareva U.* Measures of similarity as basis of quantitative researches in the field of historical linguistics. Proceedings of the Int. Conf. “Cognitive Modeling in Linguistics’XI”, Vol. 2, Kazan’, 2009, pp. 391–409.
6. *Haspelmath M., Dryer M., Gil D., Comrie B.* (Eds.) (2005) The World Atlas of Language Structures, Oxford, Oxford University Press, 2005.
7. *Jaroslavtceva E. I.* (2005) Database “Language of the World” and its applications [Kompjuternaja baza dannyh “Jazyki mira” i ee vozmozhnye prilogenija], Linguistic science doctor thesis, Moscow, IJAz RAN.

8. *Kibrik A. E.* (2003) Konstanty i peremennye jazyka [Constants and variables in language]. Sankt-Peterburg, Aleteja.
9. *Maslova E.* (2004) Dynamics of typological distribution and stability of language types [Dinamica tipologicheskih raspredelenij i stabil'nost' lazykovyh tipov], Voprosy jazykoznanija [Problems in linguistics], no. 5, pp. 3–16.
10. *Nichols J.* (1992) Linguistic Diversity in Space and Time, Chicago, London, The University of Chicago Press.
11. *Polyakov V. N., Yaroslavtseva E. I.* (2008) Quantitative laws of typological shift in the Eurasian languages (based on "Languages of the world" data base) [Kvantitativnye zakony tipologicheskogo sdviga i jazykah Evrazii]. Uchenie zapiski Kazanskogo Gosudartsvennogo Universiteta. Seriya Gumanitarnie Nauki. [Scientific notes of Kazan State University. Humanity Series], Vol. 150, Book 2, pp. 97–118.
12. *Polyakov V. N., Solovyev V. D.* (2006), Komp'uternye modeli i metody v typologii i komparativistike [Computer models and methods in typology and comparativistics]. Kazan, Kazan University.
13. *Polyakov V., Solovyev V., Wichmann S., Belyaev O.* (2009) Using WALS and Jazyki mira, Language Typology, Vol. 13, pp. 135–165.
14. *Solovyev V. D.* (2010) Typological databases: perspectives of using [Tipologicheskie basy dannyh: perspectivy ispol'zovanija], Voprosy jazykoznanija [Problems in linguistics], no. 1, pp. 94–110.
15. *Solovyev V. D., Faskhutdinov R. F.* (2009) Evaluation method for stability of grammar features [Metodica otcenki stabil'nosti grammaticheskih svoystv], Izvestija RAN. Serija jazyka I literatury [RAS News. Language and Literature series], Vol. 68, № 4, pp. 44–57.
16. *Vinogradova V. A., Novikov A. I., Jaroslavtceva E. I.* (2003) Database "Languages of the World" as the tool for linguistic researches [Baza dannyh "Jazyki mira" kak instrument lingvisticheskih issledovanii], Voprosy jazykoznanija [Problems in linguistics], no. 3.
17. *Wichmann S., Holman E.* (2009) Assessing temporal stability for linguistic typological features, München, LINCOM Europa.

ГРАММАТИКА ГЛАГОЛА И ДИАЛЕКТНОЕ ВАРЬИРОВАНИЕ[1]

Татевосов С. Г. (tatevosov@gmail.com)

МГУ имени М. В. Ломоносова, Москва, Россия

GRAMMAR OF THE VERB AND DIALECTAL VARIATION

Tatevosov S. G. (tatevosov@gmail.com)

Lomonosov Moscow State University

The paper argues for a theory that accounts for the hierarchical structure of Russian verb. The theory assumes that possible derivations of verb stems are constrained by aspectual selectional characteristics of prefixes or by their position with respect to the "secondary imperfective" morpheme. Accordingly, two groups of prefixes can be identified, selectionally restricted and positionally restricted. The paper focuses on dialectal variation that determines class membership of individual prefixes and shows that this variation is conditioned by the same selectional and positional constraints. In that way, the dialectal variation provides further support for the proposed theory of the structure of Russian verb.

Keywords: derivational morphology, structure of Russian verb, prefixation, dialectal variation

1. Множественная префиксация

За пределами России русский глагол знаменит множественной префиксацией (Babko-Malaya 1999, Ramchand 2004, Romanova 2004, 2006, Svenonius 2004, 2008), однако в своем отечестве это явление пока не привело к интересным пророчествам. Значительная работа, проделанная на префиксально-семантическом направлении в последние 30 лет (Кронгауз, Пайар (ред.) 1997, Кронгауз 1998, Кронгауз (ред.) 2001, Добрушина, Пайар (ред.) 2001, Janda 2008, Andersen

[1] Исследование проведено при поддержке РФФИ, грант № 11-06-00489а. Автор признателен двум рецензентам «Диалога» за содержательные комментарии и замечания к первой версии статьи.

et al. 2012, Janda 2012), как кажется, имела мало последствий для понимания того, как устроены общие ограничения на структуру русской глагольной основы, в частности, на размещение в ней префиксального материала. Возможно, виной тому общий тренд последних лет: мы должны интересоваться «ядерными», частотными, хорошо представленными в корпусах явлениями. Именно они показывают нам, как устроены несущие конструкции языка, именно здесь проявляются «правила» и «законы». То, что низкочастотно, «периферийно», «маргинально» и в целом находится на задворках языковой системы, можно отложить в сторону. Там хаос, беспорядок и неустроенность, там бывает всякое, там место для «исключений» и «окказиональных образований».

Мы, однако, отвергаем такой взгляд на мироздание и придерживаемся полярно противоположной установки. Редкие и периферийные явления — это именно то место, где можно увидеть содержание законов и смысл правил наиболее отчетливо. Как действуют законы, хорошо заметно при столкновении с беззаконием. Сила и мощь правил проясняется, когда мы видим их нарушения. Чтобы понять, где проходит граница между языком и неязыком, надо приблизиться к ней, то есть выйти на окраину.

Предмет этой статьи — русские мультипрефиксальные глаголы, объяснение ограничений на их образование и диалектное варьирование, которое наблюдается в этой сфере. На окончательное решение перечисленных проблем мы не претендуем: жанр заметок, в котором написана статья, допускает, как кажется, некоторую вольность в этом отношении.

2. Теория глагольной основы

Хорошо известно, что с точки зрения дистрибуции префиксы распадаются на две большие группы — внутренние и внешние. Рассмотрим, например, префикс *о-*. Для него верно следующее обобщение: если *о-* присутствует одновременно с каким-то еще префиксом, он всегда ближе к исходной основе (а именно: находится с ней в непосредственном контакте), а другой префикс присоединяется поверх него[2]:

---

[2] Отметим, что существование глаголов типа *обескровить* не нарушает это обобщение. *Без-* — не глагольный префикс, а предлог, который образует предложную группу с именной основой *кровь*. Предложная группа целиком инкорпорируется в глагольную структуру, и *о-* в ней оказывается первым собственно глагольным префиксом: [$_{PP}$ без кров'-] → #[$_{V}$ [$_{PP}$ без кров'-] -и-] → [**о-** [$_{V}$ [$_{PP}$ без кров'-] -и-]] → *обескрови-(ть)* (# помечает шаг деривации, который не реализуется в виде самостоятельного глагола). Увлекательнейший вопрос, почему в русском языке инкорпорацию в глагольную структуру допускают только предложные группы с предлогом *без-* (и они же могут инкорпорироваться в адъективную структуру, ср. [$_{A}$ [$_{PP}$ без кров'] -н] -ый), мы пока оставляем без ответа. (Мы признательны рецензенту этой статьи, который указал нам на то, что для адъективной инкорпорации предложной группы в русском языке есть больше возможностей, чем для глагольной, ср. *над-врат-н-ый, по-дорож-н-ый, при-дорож-н-ый, на-ветренн-ый, при-бреж-н-ый, под-каблуч-н-ый.*)

(1) a. *Кто* [[**пере-**[**о-стекл**]]**-я**]**-л** *балкон, подскажите, пожалуйста, контакты альпинистов для снятия стекол с балкона* [tsj.ru].
b. *о-пере-стеклял

(2) a. *Луч заметался,* [**за-**[[**о-пис**]**-ыва**]]**-л** *круги и дуги* [zhurnal.lib.ru].
b. *о-за-писывал

(3) a. *Сейчас сижу и жду, когда уже будет ясно, что я беременна… Не знаю точно, но вроде низ живота*[**под-**[**о-пух**]] [questions.cafemam.ru].
b. *о-под-пух

(4) a. *Так как* [**до-**[**о-прыска**]]**-ть** *сад было надо, заехали в магазин химии* [eftel.ru].
b. *о-до-прыскать

Префиксы, которые присоединяются поверх внутренних, соответственно, являются внешними. В [Татевосов 2009, в печати] обосновывается, что к этой группе относятся кумулятивный *на-* (как в *наварить варенья*), делимитативный *по-* (*пописать статью*), инцептивный *за-* (*запеть песню*), дистрибутивный *пере-* (*перестрелять предателей*), комплетивный *до-* (*доделать дело*), репетитивный *пере-* (*перечитать роман*), аттенуативный *под-* (*подзаработать немного денег*).

Список, конечно, не является исчерпывающим: выявление полного инвентаря внешних префиксов — эмпирическая задача на будущее. В литературе упоминаются разнообразные претенденты на эту роль, например, терминативный, или финитивный, *от-*, аттенуативный *при-*, пердуративный *про-*. И действительно: мы систематически встречаем глагольные основы, где эти префиксы имеют характерную «внешнюю» дистрибуцию.

(5) *Где тут записывают? — Тут, Вась, не записывают. Тут, Вась,* [**от-**[[**за-пис**]**-ыва**]]**-ли** *уже* [detectivebooks.ru].

(6) *В конце-концов, американцы перестали сопротивляться, кто-то* ***[при-[под-[за-кры]]]-л*** *глаза, кому-то сказали, Саакашвили сказали не возражать, и все; и вот мы взяли и вступили* [charter97.org].

(7) *Весь часовой концерт Александр Панайотов* [**про-**[[**от-кры**]**-ва**]]**-л** *рот под фонограмму* [web2edu.ru].

Не менее интересный сюжет — дистрибуция префиксов в составе так называемых циркумфиксальных глаголов, которые, как предполагает традиционная русистика, образуются одновременным присоединением префикса и постфикса *-ся*. В такой морфологической конфигурации на левой периферии основы мы можем наблюдать префиксы, которые в иных обстоятельствах проявляют свойства внутренних. Это, например, *раз-* и *об-* в (9)–(10).

(8) *— Выключил: хочу найти кого-нибудь в темноте! — А ну не прячься! Свет тут выключает! Ишь! [**Раз-[[вы-ключ]-а]-л-ся**]!!* [pandorahearts.ucoz.com]

(9) *Департамент образования Москвы* [**об-**[[**за-куп**]-**а**]-**л-ся**] *эплами* [forum.strogi.net].

Не касаясь далее циркумфиксальных глаголов, в деривационную историю которых внедряется показатель *-ся* (они нуждаются в отдельном исследовании), перейдем к основным обобщениям, определяющим дистрибуцию внешних префиксов.

3. Аспектуальная селекция и ограничение на позицию

В Татевосов 2009, в печати, мы сформулировали следующие эмпирические обобщения:

(10) *Ограничения на дистрибуцию префиксов*
a. ***Селективное ограничение.*** *Возможность присоединения префикса к основе может быть ограничена ее формальной (им)перфективностью.*
b. ***Позиционное ограничение.*** *Возможность присоединения префикса к основе может быть ограничена взаиморасположением позиции присоединения и позиции «показателя вторичного имперфектива» -ыва-.*

Обобщения в (10a–b) описывают формальные ограничения на структуру глагольной основы. Существенный результат исследования, изложенного в Татевосов 2009, в печати, состоит в том, что именно формальные ограничения играют в деривации русской глагольной основы ключевую роль, которая была серьезно недооценена нашими предшественниками (см., например, Beliakov, Guiraund-Weber 1997). Это не отрицает существования семантических ограничений и не умаляет их важности. Тем не менее весьма значительная часть невозможных русских глаголов невозможна именно в силу того, что при их деривации нарушаются (10a–b), которые не имеют к семантике никакого отношения.

Ограничения (10a–b) логически независимы, и каждый конкретный префикс может подчиняться либо одному, либо другому, либо никакому, либо сразу обоим. Соответственно, внешние префиксы не образуют единого гомогенного естественного класса, а распадаются на группы. В Татевосов 2009, в печати выделяются, в частности, селективно-ограниченные и позиционно-ограниченные внешние префиксы (далее СО-префиксы и ПО-префиксы).

К первым относятся кумулятивный *на-*, делимитативный *по-*, инцептивный *за-*, дистрибутивный *пере-*, а также пердуративный *про-*. Их дистрибуция описывается обобщением в (11), которое представляет собой частный случай (10a):

(11) *Селективно-ограниченные внешние префиксы присоединяются к формально-имперфективной основе.*

(11) предсказывает, что селективно-ограниченные префиксы способны соединяться с первичной имперфективной основой, как в (12), и вторичной имперфективной основой, полученной в результате имперфективации либо непроизводного перфективного глагола, как в (13), либо приставочного перфективного глагола, как в (14).

(12) *[на-[вари]^{I}]^{P}-ть (варенья)*
[по-[сиде]^{I}]^{P}-ть (в кресле)
[за-[би]^{I}]^{P}-ть (в барабан)
[пере-[лови]^{I}]^{P}-ть (всех мышей)
[про-[служи]^{I}]^{P}-ть (три года в армии)

(13) *[на-[[да]^{P}-ва]^{I}]^{P}-ть (пощечин)*
[по-[[реш]^{P}-а]^{I}]^{P}-ть (задачи)
[за-[[ощущ]^{P}-а]^{I}]^{P}-ть (запахи)
[пере-[[брос]^{P}-а]^{I}]^{P}-ть (камни в воду)
[про-[[обман]^{P}-ыва]^{I}]^{P}-ть (друга всю жизнь)

(14) a. — *Черт бы их побрал,* ***[на-[[от-кры]^{P}-ва]^{I}]^{P}-ли*** *детских садов целый город, а мебели не дают* [А. С. Макаренко. Педагогическая поэма].
b. *Поэтому запустил программу, записывающую действия на экране, открыл PSP, и* немного [**по-**[[**за-пис**]^{P}-**ыва**]^{I}]^{P}**-л**, *что и как* [nova-forum.com].
c. *Мальчик вдруг заиграл,* [**за-**[[**за-би**]^{P}-**ва**]^{I}]^{P}-**л** *и ваще стал звездой всей Европы* [jebers.forum24.ru].
d. *Долго искал утилитки для подобной работы,* [**пере-**[[**с-праш**]^{P}-**ива**]^{I}]^{P}-**л** всех *друзей и в итоге нашёл pssh и shmix* [michael.mindmix.ru].
e. *В эту зиму в -37 день* [**про-**[[**за-вод**]^{P}-**и**]^{I}]^{P}-**л**, *потом в сервис погнал авто* [tourerv.ru].

Попытка присоединить селективно-ограниченный префикс к формально-перфективной основе, будь то исходной, как в (15a) (ср. (13)), или производной, как в (15b) (ср. (14)), приводит к неграмматичности.

(15) a. **[на-[да]^{P}]^{P}-ть*
#[по-[реши]^{P}]^{P}-ть
**[за-[ощути]^{P}]^{P}-ть*
#[пере-[броси]^{P}]^{P}-ть
**[про-[обману]^{P}]^{P}-ть*

b. **[на-[от-кры]^{P}]^{P}-ть*
#[по-[за-писа]^{P}]^{P}-ть
**[за-[за-би]^{P}]^{P}-ть*
#[пере-[с-проси]^{P}]^{P}-ть
**[про-[за-вес]^{P}]^{P}-ти*

Предостережем читателя от неправильного прочтения (11). (11) формулирует необходимое, но не достаточное условие для присоединения префикса. Если X — селективно-ограниченный префикс, то основа, к которой он присоединяется, должна быть имперфективной. Обратное неверно: если основа имперфективна, то это не значит, что к ней присоединяется каждый из СО-префиксов. (11) устанавливает границы возможного, но внутри них для любого префикса и для любой основы могут найтись десятки причин, препятствующих гармоничному союзу.

Другая группа внешних префиксов — позиционно-ограниченные. Они подчиняются ограничению в (16), которое выступает частным случаем (10b).

(16) *Позиционно-ограниченные префиксы присоединяются не выше, чем показатель вторичного имперфектива -ыва-.*

Префиксы этой группы — комплетивный *до-*, репетитивный *пере-*, а также аттенуативные *под-* и *при-*. Ограничение в (16) провозглашает невозможным их присоединение к основе, уже содержащей показатель вторичного имперфектива. С точки зрения формальной (им)перфективности от основы ничего не требуется — все эти префиксы чувствуют себя уютно и в перфективном и в имперфективном контексте.

Чтобы увидеть действие (11) наиболее отчетливо, необходимо взять основу, которая допускает и вторичную имперфективацию с помощью *-ыва-*, и присоединение префикса. Это, например, основа *откры-*. Она может имперфективироваться, поскольку сама представляет собой приставочный перфектив. Она комбинируется с ПО-префиксами, поскольку ПО-префиксы не требуют, чтобы основа была имперфективной. Эти возможности показаны в (17); группу ПО-префиксов представляет аттенуативный *при-*.

(17) a. *от-кры-* → *откры-**ва***-
b. *от-кры-* → ***при**-откры-*

Возьмем теперь глагол *приоткрывать*. Априори можно вообразить два способа, которыми этот глагол возник: перфективацией *открыва-* с помощью *при-* и имперфективация *приоткры* с помощью *-ыва-*:

(18) a. *от-кры-* → *откры-**ва***- *→ ***при**-открыва-*
b. *от-кры-* → ***при**-откры-* → *приоткры-**ва***-

В (18a) последний шаг деривации — префиксация, а значит, глагол *приоткрывать* перфективен. В (18b) деривация завершается вторичной имперфективацией, создающей имперфективный глагол. Мы знаем, что в действительности *приоткрывать* — глагол несовершенного вида, а значит, деривация в (18a) невозможна. Поскольку и *открывать*, (17a), и *приоткрыть*, (17b), в русском языке существуют, источник деривационной коллизии следует искать в последнем шаге, то есть в переходе «откры-ва- → **при**-открыва-» в (18a). Проблема не может состоять в несочетаемости аттенуативного *при-* с имперфективной основой (ср. *приглушить*, *припухнуть*, *притормозить*), а значит,

мы имеем дело с аутентичным морфосинтаксическим ограничением, не выводимым из каких-либо других. Если в качестве такого ограничения предложить (16), (18a–b) и аналогичные примеры получают убедительное объяснение. В (19)–(21) приводится несколько дополнительных примеров, которые показывают, что если в основе присутствуют одновременно и ПО-префикс и показатель вторичного имперфектива, последний присоединяется поверх первого.

(19) *Учащийся, работая с любым предметным материалом… [[**пере-[от-кры]**]^{P}]^{P}-**ва**]^{I}-**ет** процесс возникновения того или другого знания, переоткрывает открытие, некогда сделанное в истории* [wiki.ippk.ru].

(20) *Отработанные навыки без проблем позволяют балансировать на одной ноге, пока другая держит и* [[**до-[от-кры**]^{P}]^{P}-**ва**]^{I}-**ет** *дверь* [mindfactor.livejournal.com].

(21) *Проверка выявила ещё один нюанс: после запуска двигателя пневмоклапан* [[**под-[от-кры**]^{P}]^{P} -**ва**]^{I}-**ет** *заслонку на 5,5 мм* [2126.ru].

В (19)–(21) представлены лично-числовые формы трех глаголов, и все они имеют временную референцию к настоящему. Это показывает, что перед нами глаголы НСВ, а значит присоединение ПО-префиксов происходит до имперфективации — в полном соответствии с (16)[3].

4. Диалектное варьирование

Итак, мы утверждаем, что расположение префиксов в основе определяется двумя структурными факторами — аспектуальной селекцией и позицией префикса по отношению к имперфективу. Эти факторы создают пространство из четырех логических возможностей, которые показаны в (22).

(22)

	Нуждаются в имперфективной основе	Не имеют селективных ограничений
Присоединяются не выше *-ыва-*	Некоторые лексические префиксы	ПО-префиксы *(до-, пере-, под-, при-)*
Не имеют ограничений на позицию	СО-префиксы *(на-, пере-, по-, за-, про-, от-)*	—

[3] Отметим принципиально важный момент: бессмысленно говорить о членстве префикса в каком-то из выделенных классов в абсолютных терминах — «Префикс *на-* является внешним селективно-ограниченным». Префикс входит в тот или иной класс в определенном значении. Кумулятивный *на-* состоит в классе СО-префиксов, тогда как *на-* в составе *написать* выступает как внутренний префикс, ср. *донаписать*, *перенаписать* и т. п.

Проработанный выше материал позволяет утверждать, что эмпирически реальны по крайней мере две группы внешних префиксов — СО-префиксы и ПО-префиксы, которые реализуют две из четырех возможностей. Третья закрепляется за некоторыми внутренними префиксами, которые присоединяются до *-ыва-* и нуждаются в имперфективной основе. (Таков, например, прификс *у-* в том значении, которое АГ-80 характеризует как «уменьшить(ся) с помощью действия, названного мотивирующим глаголом»: *уварить(ся)*, *ужарить(ся)*, *ужать(ся)*, *урезать*, *усохнуть*, *ушить*, *укипеть*.)

Возникает резонный вопрос: а что с четвертой возможностью? Есть ли в русском языке префиксы, которые способны появляться, во-первых, и до и после *-ыва-*, а во-вторых, и в перфективном и в имперфективном контексте? Простой ответ состоит в том, что теория допускает их существование, но не требует этого. И если в правом нижнем углу в (22) стоит прочерк, ничего неприятного для теории в этом нет. Неприятности начались бы, если бы обнаружились префиксы, дистрибуция которых не описывается ни в терминах аспектуальной селекции, ни в терминах позиции в иерархической структуре по отношению к *-ыва-*.

Однако на вопрос о четвертой возможности есть и более интересный ответ. В том диалекте русского языка, на котором говорит автором этих заметок и значительное количество других носителей, правый нижний угол пуст. Но есть и другие диалекты[4].

Анна А. Зализняк (2003) полагает, что «кумулятивный способ действия может образовываться от глаголов как совершенного, так и несовершенного вида». В качестве кумулятивов, образованных от глаголов СВ, она упоминает два — *накупить* и *напридумать*; в АГ-80 мы находим также *напустить*, *нарассказать* (разг.), *насочинить* (разг.). Чтобы двигаться дальше, разведем два случая: *накупить* и *напустить*, с одной стороны, и *напридумать*, *нарассказать* и *насочинить*, с другой. Первые признаются в качестве кумулятивов всеми носителями русского языка, в том числе и теми, кто говорит на диалекте, устроенном в соответствии с (22). Назовем его диалектом S. Вторые для носителей диалекта S находятся по ту сторону грамматичности.

Каков статус глаголов *накупить* и *напустить* в диалекте S и не следует ли из-за них изъять кумулятивный *на-* из списка СО-префксов? Ответ был бы положительным в следующем случае: основы СВ, которые допускают соединение с кумулятивным *на-*, образуют естественный класс (пусть даже и количественно ограниченный). Видя, что *купить* и *пустить* — непроизводные перфективы на *-ить*, мы проверяем на предмет этой возможности прочие основы такого же класса. Обнаруживается, что новые кумулятивы от глаголов СВ на *-ить* в список не добавляются — **наощутить*, **нарешить* и т. п. невозможны. Статус *накупить* и *напустить* проясняется — это лексические исключения. Список СО-префиксов для диалекта S остается нетронутым.

---

[4] Мы, разумеется, менее всего имеем в виду географию расселения носителей как определяющее свойство диалекта. Говорить о диалектах мы можем всякий раз, когда сталкиваемся с близкими, но не тождественными состояниями русской грамматики, при условии, что к каждому состоянию прилагается осмысленное количество носителей. См. также ниже.

Есть, однако, значительное количество носителей русского языка, для которых грамматичны не только *накупить* и *напустить*, но и *напридумать*, *нарассказать*, *насочинить*, *наоткрыть*, *назабить,* невозможные в диалекте S. Они образуют отдельный диалект (назовем его диалектом N), который поставляет нам предложения типа (23):

(23) *Для тебя понаоткрывали да понаформулировали законы физики али информатики, которые ты бы никогда не* [**на-**[**от-кры**]^{P}]^{P}-**л** [avanturist.org].

Диалект N отличается от диалекта S ровно по одному параметру: характеристиками кумулятивного *на-*. В диалекте N *на-* не входит в группу СО-префиксов и отправляется в правый нижний угол (22), туда, где находятся единицы, не имеющие ни селективного, ни позиционного ограничения (назовем их *свободными префиксами*). Четверная возможность, отсутствующая в диалекте S, успешно реализуется в диалекте N.

Есть ли еще кандидаты на членство в классе свободных префиксов? В этом месте перед нами возникает другой пример диалектного варьирования. Имеется устойчивая группа носителей, для которых комплетивный *до-* отличается от других единиц класса ПО-префиксов: он способен присоединяться не только ниже, но и выше показателя вторичного имперфектива. В этом диалекте (далее — диалекте D), глаголы типа *доподписывать* или *дораскрывать* являются двувидовыми и допускают обе деривации в (24):

(24) a. *писа* → ***под****-писа-* → ***до****-подписа-* → *доподпис-****ыва***-
b. *писа* → ***под****-писа-* → *подпис-****ыва***- → ***до****-подписыва-*

(25)–(26) иллюстрируют перфективные употребления глаголов *доподписывать* и *дораскрывать*, принадлежащие диалекту D и невозможные в диалекте S:

(25) *В оставшийся час до похода на почту с американкой* [**до-**[[**под-пис**]^{P}-**ыва**]^{I}]^{P}-**л** *адреса на открытках — успел!* [journals.ru]

(26) *Сегодня* [**до-**[[**рас-кры**]^{P}-**ва**]^{I}]^{P}-**л** *виноград и обрезал всё черноту на розах. Аж, спина болит от полевых работ* [forumhouse.ru].

Таким образом, в диалекте D четвертая возможность также реализуется, хотя и иначе, чем в N: в свободные префиксы зачисляется комплетивный *до-*.

Принципиально важно следующее. Диалекты N и D, как и диалект S, находятся внутри того пространства возможностей, которые предсказывает теория. Ни в N, ни в D не обнаруживается префиксов, которые проявляют какие-то неожиданные свойства и/или подчиняются принципиально иным ограничениями. В теории, оперирующей обобщениями (11) и (16), переход от S к N и D и обратно сводится к простому параметрическому варьированию; диалект полностью определяется тем, находится ли интересующий нас префикс под юрисдикцией соответствующего ограничения:

(27) ***Диалект N.*** *В отличие от других СО-префиксов, префикс на- не подчиняется селективному ограничению в (11), то есть не требует формальной имперфективности основы.*

(28) ***Диалект D.*** *В отличие от других ПО-префиксов, префикс до- не соблюдает позиционное ограничение в (16), то есть не исключает присоединения поверх показателя вторичного имперфектива.*

У нас нет полной информации о том, как соотносятся диалекты N и D. Как указывает В. И. Беликов (личное сообщение), диалект N — это «архаичная, но все еще живая в старшем поколении норма». Проведя собственное мини-исследование блогосферы, он установил, в частности, что пропорция глагола *нарассказать* в паре *нарассказывать/нарассказать* монотонно возрастает с увеличением возраста блогеров — от 11,4 % среди блогеров 12–21 года до 37 % среди блогеров 50–69 лет. Согласно его же данным, способности носителей диалекта N присоединять кумулятивный *на-* к основе СВ заметно варьируют в зависимости от лексического контекста: например, аналогичная пропорция для *напридумать* среди блогеров старшего возраста не превышает 15 %, а для *насочинить* — 2–3 %. Составить по этим данным полный социолингвистический портрет блогера-носителя диалекта N, конечно, трудно (в том числе и потому, что в общей массе блогеров есть носители как N, так и S, а их пропорция неизвестна). Ясно, однако, что некоторые тенденции прослеживаются вполне отчетливо.

По нашей предварительной оценке, диалект D несколько более многочислен и значительно более молод, чем диалект N: среди его носителей преобладает люди младших поколений. Но чтобы оценить их количество и качества, требуется подробное корпусное исследование, дополненное анкетированием носителей. Как бы то ни было, поскольку параметры варьирования в (27) и (28) полностью независимы, естественно ожидать, что диалекты N и D пересекаются. Согласно нашим данным, это ожидание полностью подтверждается: имеется устойчивая, хотя и не очень большая, группа носителей смешанного диалекта N&D, одновременно обладающего свойствами в (27) и (28). Префиксальная система в N&D выглядит так, как показано в (29).

(29)

	Нуждаются в имперфективной основе	Не имеют селективных ограничений
Присоединяются не выше *-ыва-*	Некоторые лексические префиксы	ПО-префиксы *(пере-, под-, при-)*
Не имеют ограничений на позицию	СО-префиксы *(пере-, по-, за-, про-)*	Свободные префиксы *(на-, до-)*

В диалекте N&D свободные префиксы представлены двумя единицами — комплетивным *до-* и кумулятивным *на-*. Ясно, что (29) так же хорошо соответствует предсказаниями теории, как и (22). Принципиально важно осознать, что диалектное варьирование, которое отражают (22) и (29), обладает двумя критически важными свойствами: во-первых, оно затрагивает характеристики отдельных лексических единиц; во-вторых, оно происходит в пределах уже установленных в теории грамматических ограничений. Такое варьирование — важный эмпирическое свидетельство в пользу теории.

Последний эмпирический урок, который дает исследование множественной префиксации, приводит нас к пониманию того, что диалект — это не только и не столько отличие говора орловских крестьян от говора рязанских помещиков. Диалекты — это в первую очередь различные состояния одной и той же грамматической системы. Переход между состояниями происходит путем простого изменения простых параметров, в нашем случае — путем включения и выключения конкретных грамматических ограничений для конкретных лексических единиц.

5. Заключение

В предложенной нами теории обосновывается, что префиксация в русском языке чувствительна к двум ограничениям — селективному и позиционному. Исходя из этих ограничений, следует признать эмпирически реальными три группы внешних префиксов. Селективно-ограниченные и позиционно-ограниченные префиксы присущи, по-видимому, всем диалектам русского языка. Наличие свободных префиксов, не подчиняющихся ни одному из ограничений, представляет собой параметр диалектного варьирования.

Литература

1. *АГ* (1980). Русская грамматика. М.: Наука
2. *Добрушина Е. Р., Пайар Д.* (ред.). Русские приставки: многозначность и семантическое единство. М.: Русские словари, 255–270.
3. *Зализняк Анна А.* 2003. Способ действия // Энциклопедия «Кругосвет». http://www.krugosvet.ru
4. *Зализняк Анна А., Шмелев А. Д.* 2000. Введение в русскую аспектологию. М.: Языки русской культуры.
5. *Исаченко А. В.* 1960. Грамматический строй русского языка в сопоставлении со словацким. Т. 2. Морфология. Братислава.
6. *Кронгауз М. А.* 1998. Приставки и глаголы в русском языке: семантическая грамматика. М.
7. *Кронгауз М. А.* (ред.). 2001. Глагольные префиксы и префиксальные глаголы. Московский лингвистический журнал №5–1. М.: РГГУ.
8. *Кронгауз М. А., Пайар Д.* (ред.). 1997. Глагольная префиксация в русском языке. М.: Русские словари.
9. *Татевосов С. Г.* 2009. Множественная префиксация и анатомия русского глагола // Киселева К. Л., Плунгян В. А., Рахилина Е. В., Татевосов С. Г. (ред.) Корпусные исследования по русской грамматике. М.: Языки славянских культур.
10. *Andresen A., Janda L. A., Kuznetsova J., Lyashevskaya O., Makarova A., Nesset T., Sokolova S.* (2012), Russian ‘purely aspectual’ prefixes: Not so ‘empty’ after all? Scando-Slavica, Vol. 58, No. 2, pp. 231–291.
11. *Beliakov V., Guiraund-Weber M.* (1997). О некоторых свойствах вторичных глагольных приставок // Russian Linguistics. 1997. V. 21, N 2.

References

1. *AG* (1980), Russian grammar [Russkaja grammatika]. Moscow: Nauka
2. *Dobrushina E. R., Paillard D.* (ed.) (1997), Russian prefixes: ambiguity and semantic uniformity [Russkie pristavki: mnogoznachnost' i semanticheskoe edinstvo]. Moscow: Russkie slovari, pp. 255–270.
3. *Zaliznjak An. A.* (2003), Aktionsart [Sposob dejstvija], Encyclopedia «Krugosvet», http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/SPOSOB_DESTVIYA.html
4. *Zaliznjak An. A., Shmelev A. D.* (2000), Introduction to the study of Russian aspect [Vvedenie v russkuju aspektologiju]. Moscow: Jazyki russkoj kul'tury.
5. *Isachenko A. V.* (1960), Grammar of Russian as compared to Slovak [Grammaticheskij stroj russkogo jazyka v sopostavlenii so slovackim]. Vol 2: Morphology [Morfologija]. Bratislava.
6. *Krongauz M. A.* (1998), Prefixes and verbs in Russian: a semantic grammar [Pristavki i glagoly v russkom jazyke: semanticheskaja grammatika]. Moscow.
7. *Krongauz M. A.* (ed.). (2001), Verbal prefixes and prefixed verbs [Glagol'nye prefiksy i prefiksal'nye glagoly], Moscow journal of linguistics [Moskovskij lingvisticheskij zhurnal]. Vol. 5, № 1. Moscow: RGGU.
8. *Krongauz M. A., D.Pajar (ed.). (1997), Glagol'naja prefiksacija v russkom jazyke. Moscow: Russkie slovari.*
9. *Tatevosov S. G.* (2009), Multiple prefixation and the anatomy of Russian verb [Mnozhestvennaja prefiksacija i anatomija russkogo glagola], in Kiseleva K. L., Plungjan V. A., Rahilina E. V., Tatevosov S. G. (ed.) Korpusnye issledovanija po russkoj grammatike. Moscow: Jazyki slavjanskih kul'tur.
10. *Andresen A., Janda L. A., Kuznetsova J., Lyashevskaya O., Makarova A., Nesset T., Sokolova S.* (2012), Russian 'purely aspectual' prefixes: Not so 'empty' after all? Scando-Slavica, Vol. 58, No. 2, pp. 231–291.
11. *Beliakov V.* Guiraund-Weber M. (1997), On the secondary verbal prefixes [O nekotoryh svojstvah vtorichnyh glagol'nyh pristavok], Russian Linguistics, Vol. 21, No 2.
12. *Janda L. A.* (2008), Semantic Motivations for Aspectual Clusters of Russian Verbs. Ms, University of Tromso.
13. *Janda L. A.* (2012), Russian prefixes as verbal classifiers [Russkie pristavki kak sistema glagol'nyx klassifikatorov], Voprosy jazykoznanija, № 6, pp. 3–47.
14. *Ramchand, G.* (2004). Time and the event: The semantics of Russian prefixes. *Nordlyd* 32–2. *Special issue on Slavic prefixes*, 323–366.
15. *Romanova, E.* (2004). Superlexical vs. lexical prefixes. *Nordlyd* 32–2. *Special issue on Slavic prefixes*, 255–278.
16. *Romanova, E.* (2006). Constructing Perfectivity in Russian. Ph.D. dissertation. University of Tromsø.
17. *Svenonius, P.* (2004). Slavic prefixes inside and outside VP. *Nordlyd* 32–2. *Special issue on Slavic prefixes*, 323–361.
18. *Svenonius, P.* (2008). Russian prefixes are phrasal. Ms., University of Tromsø.

НЕОТРИЦАЕМЫЕ ПРЕДИКАТЫ: НАРЕЧИЕ *ВПОРУ*

Урысон Е. В. (uryson@gmail.com)

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

Ключевые слова: аналитическое толкование, сфера действия отрицания, языковая аномалия, коммуникативные постулаты Грайса

THE RUSSIAN ADVERB *VPORU* 'SUITING BEST': A PREDICATE WHICH DOES NOT COMBINE WITH NEGATION

Uryson E. V. (uryson@gmail.com)

V. V. Vinigradov Russian Language Institute RAS, Moscow, Russia

The object of the paper is the Russian adverb VPORU 'suiting best'. In the 19 century the meaning of this word was less rich, so it was used in more types of contexts than now. At presence the adverb VPORU is freely used in three types of contexts: (a) *Pidzhak emu vporu* 'The coat is the right size for him'; (b) *Ej zamuzh vporu (a ona v kukly igraet)* 'She should marry (but not play with dolls)'; (c) *Zdes' tak temno — vporu na chetveren'kakh polzti* 'It is so dark here — one might as well crawl on his fours'. The adverb VPORU in (a) freely combines with negation in the 19 cent. language, but not in the the present day Russian. The reason is that in the 19 cent. language the meaning of the word VPORU in (a) is 'suiting' but not 'suiting best'. The latter meaning consists of two predicates. It is demonstrated that negation of such sense breaks a Grice maxim. So, Grice maxims being applied to a meaning of an anomalous word combination can explain the reason of its anomaly. The adverb VPORU in (b) and (c) does not combine with negation. Contexts (b) are similar to (a). As for (c) the meaning of VPORU here has a rich modal frame. Being in the scope of negation the assertion of VPORU contradicts this modal frame; this reason of an anomaly of a word combination has been described by Ju. D. Apresjan [1978/1995].

Keywords: meaning structure, scope of negation, anomaly, Grice maxims

1. Постановка задачи

Объект предлагаемой работы — наречие *впору* в современном русском языке. Слово *впору* заинтересовало нас тем, что оно плохо сочетается с частицей *не*. Нормально: *Эта куртка мне <ему> впору*, однако по крайней мере для некоторых носителей русского языка плохо: **Эта куртка мне <ему> не впору* (ниже мы вернемся к этому утверждению); нормально: *Здесь такая скука — впору запить* или *Да, впору запить*, или *Что, впору запить?* однако невозможно: **Здесь скучно, но не впору запить* или **Нет, не впору запить*; нормально также *Ему жениться впору (а он все балбесничает)*, однако в современном языке невозможно: **Ему жениться не впору*. Слово *впору* можно назвать неотрицаемым предикатом. Задача работы — дать интерпретацию этому явлению.

Нужно сказать, что в логике отрицание «неверно, что» можно присоединить, вообще говоря, к любому утверждению. Так, в логике нормальны утверждения: *Неверно, что эта куртка ему впору; Неверно, что при таких обстоятельствах впору запить; Неверно, что ему впору жениться*. Подобные высказывания правильны и с точки зрения естественного (русского) языка — они не являются аграмматичными, хотя вряд ли встретятся в обычной речи. Однако в естественном языке (мы будем говорить о русском) логическому отрицанию «неверно, что» соответствует отрицание *не*. Русское отрицание (и, видимо, вообще отрицание в естественном языке), в отличие от логического отрицания, присоединяется не к любому предикату.

Выделяется целый класс предикатов, которые не взаимодействуют с отрицанием, — это союзы. Безусловно невозможно сочетание *не* с сочинительным союзом: **не и, *не но, *не а, *не или, * не а то* и т. п. Как правило, невозможно сочетание отрицания с подчинительным союзом **P, не если Q; *P, не хотя Q; *P, не когда Q* и т. п.; иногда соответствующий смысл можно выразить, однако отрицание тогда относится не к союзу, а к так называемому указательному слову. Ср. *P не при условии, если Q; P не тогда, когда Q* и т. п. (Заметим, что т. н. контрастное отрицание ведет себя так же: ср. *P не тогда, когда Q, а когда R*.) Правда, отрицание *не* свободно присоединяется к т. н. расчлененным союзам; ср. *P не потому, что Q; P не благодаря тому, что Q*. Однако расчлененный союз можно трактовать как сочетание собственно союза (*что*) с указательным словом (*потому* или *благодаря*), а указательное слово нормально присоединяет к себе отрицание; ср. *P не тогда, когда Q; P не для того, чтобы Q* и т. п. (Иная трактовка отрицания при расчлененном союзе *потому, что* предложена Л. Н. Иорданской [1992; 2007]). Союзы — это, по-видимому, единственный большой класс неотрицаемых предикатов.

Много неотрицаемых предикатов среди частиц. Например, не присоединяет отрицания частица *даже*, контрастируя в этом отношении с *только*: нормально *не только*, однако невозможно **не даже*. Так, нормальны высказывания: *Ваня пил только коньяк* и *Ваня пил не только коньяк*; нормально также: *Ваня пил даже чистый спирт*, однако невозможно: **Ваня пил не даже чистый спирт*.

Остальные большие классы предикатов — глаголы, прилагательные, наречия, предлоги — в норме сочетаются с отрицанием *не*. Ср. *хотеть — не хотеть,*

жаль — не жаль, очень — не очень, (Это сделано) не для Васи, (Книга лежит) не на столе и т. п. Для многих прилагательных и наречий отрицание настолько нормально, что предикат «срастается» с *не* в новое слово; ср. *невеселый, нехорошо* и т. п. Такое «срастание» возможно и в случае глагола. Так, хорошо известно, что сочетание *не любить* может обозначать не только отсутствие любви, но и особое чувство, близкое неприязни (причем в обоих случаях сочетание глагола с *не* пишется в два слова). Казалось бы, в перечисленных частях речи не должно быть неотрицаемых предикатов. Однако, оказалось, что это не так: ср. наречие *впору*.

Некоторые неотрицаемые предикаты (частицы *вот* и *вон*; предикаты со значением сообщения, знания, истинности в составе вводной конструкции) описаны Ю. Д. Апресяном в работе [1978/1995], где показано, что неотрицаемость предиката обусловлена спецификой его значения: в семантической структуре такого предиката имеется компонент, препятствующий сочетанию предиката с отрицанием.

В каждом конкретном случае эта специфика семантической структуры предиката может проявляться по-своему. Иными словами, каждый неотрицаемый предикат должен описываться отдельно. Сосредоточимся на наречии *впору*.

2. Наречие «впору» в современном языке

Слово *впору* является уходящим: теперь оно встречается в гораздо более ограниченном круге контекстов, чем во второй половине 19 века, и имеет меньше значений. В частности, ушла из языка лексема *впору* ‘вовремя, когда нужно’, фиксируемая еще в [Словарь Ушакова]; ср. совершенно устаревшие примеры *Пришел впору; Приезд не впору мой* (А. С. Грибоедов).

В современном языке для слова *впору* характерны три следующих типа контекстов (мы опираемся на описание слова *впору*, предложенное Т. В. Крыловой для Активного словаря русского языка, разрабатываемого под рук. Ю. Д. Апресяна.):

(1) *Куртка ему впору; Сапоги пришлись впору.*

(2) *Ему жениться впору (а он балбесничает); Она полагала, что если доктор узнает ее настоящие годы, то не станет ее лечить и скажет, что ей впору умирать, а не лечиться* [А. П. Чехов. Мужики].

(3) *Такая тоска — впору запить.*

Вслед за Т. В. Крыловой будем считать, что в каждом из этих контекстов представлена отдельная лексема наречия *впору*. Это решение оправдано технически: каждый из этих контекстов удобно описывать отдельно, т. е. усматривать в каждом из них свою «единицу описания» наречия.

Начнем с контекстов (1), в которых представлена лексема *впору 1*.

2.1. «Туфли оказались ей впору»

В современном языке лексема *впору 1* употребляется только в предикативной позиции, однако это ограничение появилось относительно недавно. Ср. следующие примеры — в первом *впору* выступает в функции атрибута (несогласованного определения) существительного, а во втором — в функции обстоятельства:

(4) *У Ватажко нашлись высокие охотничьи сапоги впору Жоржу* [С. М. Степняк-Кравчинский. Андрей Кожухов].

(5) *— Крой новый кафтан, да к старому почаще прикидывай, а то, пожалуй, не впору сошьешь.* [П. И. Мельников-Печерский. На горах. Книга вторая].

В современном языке встречается употребление *впору* в функции атрибута, но подобные примеры выглядят устаревшими. Ср.

(6) *Электричество горит, трамваи ходят, вода из кранов льется, печки топятся. У меня на ногах — башмаки впору* [Л. К. Чуковская. Прочерк].

Сужение синтаксических функций лексемы *впору 1* свидетельствует о том, что данная лексема уходит из языка.

Лексема *впору 1* еще в конце 19 века вполне сочеталась с отрицанием. Ср.

(7) *Ветхое ситцевое платье всегда было ей [горничной] не впору и сильно стесняло могучие юные формы* [Д. Н. Мамин-Сибиряк. Приваловские миллионы].

(8) *— Мой фрак тебе не впору… — Впору; вот не впору! — перебил Тарантьев* [И. А. Гончаров. Обломов].

(9) *Хозяйский сюртук был не совсем впору и сильно тянул его руки назад* [А. Ф. Писемский. Сергей Петрович Хозаров и Мари Ступицына].

В современном языке подобное употребление тоже встречается, однако воспринимается нами как стилизованное. Ср.

(10) *Уходя с базара, он купил ей еще и хорошенькие татарские туфли, бархатные, украшенные золотой канителью, и в Державинском саду, испугавшись, что они ей не впору, предложил примерить* [В. А. Каверин. Перед зеркалом].

(11) *Ноги у него великанские; какие сапоги ни купит себе в Выборге или в Петербурге — все не впору: малы* [Л. К. Чуковская. Памяти детства: Мой отец — Корней Чуковский].

В идиолекте автора сочетание *не впору* возможно лишь в очень экспрессивном высказывании, когда говорящий возражает адресату, как бы передразнивая

его реплику: *Да что ты говоришь «впору, впору пиджак», не впору он мне!* Как указал анонимный рецензент, в современном русском языке употребление *не впору* (причем не экспрессивное) встречается не так уж редко, в частности, в блогах. В связи с этим возникает отдельная задача описания употребления лексемы *впору 1* в речи разных носителей языка. Описание узуса и нормы, однако, выходит далеко за рамки предлагаемой работы. Наша цель гораздо скромнее: поскольку существуют носители языка, для которых сочетание *не впору* ненормально, то следует найти причины такого запрета. Они носят семантический характер, и мы попытаемся это обосновать.

Лексема *впору 1*, представленная в контекстах (1), в первом приближении толкуется так:

(12) а. *X впору Y-у* ‘Одежда, обувь или головной убор X подходит по размеру человеку Y’.

Или так:

(12) б. *X впору Y-у* ‘Одежда, обувь или головной убор X соответствует по размеру человеку Y’.

Это толкование недостаточно объясняет факты языка. В частности, наречие *впору 1* не сочетается с показателями высокой степени. Невозможно: **совсем впору, *совершенно впору* и даже **очень впору*.

Эта сочетаемость лексемы *впору 1* может быть обусловлена каким-то оценочным компонентом ее значения. Ср. толкование этой лексемы в [МАС; Словарь Ушакова]: ‘как по мерке, как раз’, которое безусловно содержит оценку. Заметим, что предикаты *как по мерке* и *как раз* плохо сочетаются с теми же показателями степени. Ср. ??*совсем как по мерке*, ??*совсем как раз*, ??*совершенно как раз*, ?? *совершенно как по мерке, *очень как по мерке*, **очень как раз*. Сомнительно (в не экспрессивной речи) сочетание этих предикатов с отрицанием, ср. ??*Пиджак не как по мерке*, ??*Сапоги не как раз*. Ср. нормальные высказывания: *Пиджак — как по мерке! Сапоги как раз*.

На наш взгляд, лексема *впору 1* содержит оценочный компонент ‘в точности’ или, если воспользоваться менее идиоматичными выражениями, компонент ‘точно’ или ‘абсолютно’. (Выбор того или иного из этих компонентов не влияет на ход дальнейшего рассуждения.) Уточненное толкование этой лексемы таково:

(13) а. *X впору Y-у* ‘Одежда, обувь или головной убор X в точности подходит по размеру человеку Y’.

Или так:

(13) б. ‘Одежда, обувь или головной убор X в точности соответствует по размеру человеку Y’.

Компонент 'в точности' указывает на высшую степень Р и потому не допускает никаких иных показателей степени при Р. Невозможно: **очень в точности, *совсем в точности, *совершенно в точности*. Покажем теперь, что этот же компонент мешает сочетаться лексеме *впору 1* с отрицанием.

Представим себе возможный смысл 'не в точности подходит' или 'не в точности соответствует'. Что здесь отрицается? Или, более точно, какова сфера действия отрицания 'не'? Логически здесь возможны два варианта.

А) Отрицается целиком смысл 'не в точности подходит'. Более формально, сфера действия отрицания такова: 'не (в точности подходит)', или 'не (в точности соответствует)'. Иными словами, отрицается и точность соответствия, и само соответствие. Но такая организация смысла противоречит постулату коммуникативности Грайса «не говори лишнего». Действительно, если отрицается соответствие, то заведомо отрицается и точное соответствие. Значит, достаточно отрицать соответствие, т. е. сказать *не подходит*. С этой точки зрения сочетание **не впору* не употребляется, т. к. в его семантической структуре содержится лишний компонент.

Б) Отрицается только компонент 'в точности'. Тогда сфера действия отрицания такова: 'подходит не в точности', или 'соответствует не в точности'. В языке имеется специальное слово для выражения смысла 'не в точности' и аналогичных смыслов 'не абсолютно', 'не совершенно'. Это слово *почти*. Для того чтобы подвергнуть отрицанию компонент 'в точности' лексемы *впору 1*, его нужно «извлечь» из значения лексемы, и полученный смысл 'не в точности' выразить словом *почти*. Иными словами, отрицание не воздействует на компонент 'в точности' значения лексемы *впору*, почему и невозможно сочетание **не впору* (напомним, что мы описываем словоупотребление ряда носителей современного русского языка). Вместо этого сочетания употребляется сочетание *почти впору*. Ср.

(14) *Маскхалат мне почти впору, правда, брюки несколько коротки* [Владимир Богомолов. Иван].

Предложенный анализ можно подтвердить языковым материалом.

Выражение *почти Р* задает множество ситуаций, очень близких *Р* [Баранов, Плунгян, Рахилина 1993; Григорьева 2004а, 2004б]. Выражение *в точности Р*, а также наречия *абсолютно, совершенно, точно* и т. п., напротив, исключают подобные ситуации. Все данные предикаты плохо сочетаются с отрицанием 'не'.

Действительно, нормально: *абсолютно идеальное лицо*; *совершенно гладкая поверхность; характер точно такой же, как у отца*. Однако в нейтральном контексте странно: *не абсолютно идеальное лицо*; *не совершенно гладкая поверхность; характер не точно такой же, как у отца*. Подобные выражения нормальны в полемическом контексте, как возражение на реплику собеседника, когда говорящий ее как бы повторяет. Ср. — *У Евы абсолютно идеальное лицо. — Нет, у нее не абсолютно идеальное лицо*. Ответная реплика может

значить: (а) что лицо вообще неидеальное, ср. *Ну нет, у нее не абсолютно идеальное лицо, посмотри, какой нос!*; или (б) что оно близко идеальному, но не абсолютно идеальное. Однако смысл (б) в нейтральном контексте выражается словом *почти*, ср. *У нее почти идеальное лицо*. Аналогичные примеры: — *Это совершенно гладкая поверхность. — И не совершенно гладкая — местами даже шершавая* vs *Это почти гладкая поверхность; — Характер у него точно такой же, как у отца. — Вот и не точно такой же, совсем другой!* vs *Характер у него почти такой же, как у отца*. Аналогичная пара: *совершенно пустой зал — почти пустой зал*. В целом, выражение *Почти P* задает множество ситуаций очень близких P; а выражение *Абсолютно <совершенно> P* , напротив, отсекает множество этих ситуаций.

Точно так же плохо сочетается с отрицанием предикат *почти*. Нормально: *почти абсолютно идеальное лицо; почти совершенно гладкая поверхность; характер почти точно такой же, как у отца*. Однако лишь в полемическом контексте возможно: *не почти абсолютно идеальное лицо; не почти совершенно гладкая поверхность; характер не почти точно такой же, как у отца*. Действительно, в выражении *не почти абсолютно идеальное лицо* формально отрицается близость к идеалу, но что же имеется в виду? В соответствии с постулатами Грайса, имеется в виду: (а) либо то, что лицо вообще далеко от идеального, — и тогда данный смысл выражается словами *неидеальное лицо*; (б) либо то, что лицо не близко идеалу, а полностью ему соответствует, — но данный смысл выражается словами *абсолютно идеальное лицо*.

Вернемся к лексеме *впору 1*. В современном русском языке, в узусе ряда его носителей, лексема *впору 1* содержит оценочный компонент, указывающий на точность соответствия размера предмета одежды и т. п. размеру субъекта. В языке 19 века (а также в языке некоторых современных носителей языка) данная лексема *впору* не содержала этого компонента, почему и возможно было сочетание *не впору*. По-видимому, были нормальны и выражения *очень впору, совершенно впору, совсем впору*.

Заметим, что общее высказывание типа *Обувь должна быть впору* вполне подвергается отрицанию, ср. *Обувь не должна быть впору — лучше покупать туфли на размер больше*. Но отрицание ‘не’ относится здесь все-таки к предикату ‘должен быть’, а не предикату ‘впору’.

2.2. «Ей замуж впору»

Перейдем к контекстам типа

(2а) *Ему впору жениться (а он балбесничает)*.

(2б) *Она полагала, что если доктор узнает ее настоящие годы, то не станет ее лечить и скажет, что ей впору умирать, а не лечиться* [А. П. Чехов. Мужики].

В этих контекстах представлена лексема *впору 2 P*, которая указывает на то, что ситуация P больше всего подходит субъекту или подходит субъекту больше, чем другая, возможно подразумевая ситуация. При этом ситуация P не имеет места в реальности, а имеют место те ситуации, которым она противопоставляется. Так, в (2а) утверждается, что субъекту, который балбесничает, подходит жениться, — но это желаемая ситуация, которая не имеет места. В (2б) утверждается, что старухе подходит умирать и не подходит лечиться; при этом старуха не умирает, а старается лечиться. В высказывании *Кате замуж впору* утверждается, что Кате подходит выйти замуж, однако ситуация вступления Кати в брак не имеет места, и эта желаемая ситуация противопоставляется другим неназванным, но подразумеваемым ситуациям, в которых субъект участвует.

Данная лексема *впору* выражает еще и неодобрительное отношение к субъекту. Благодаря этому компоненту значения вряд ли возможны нейтральные высказывания с данной лексемой от первого лица. Плохо ??*Мне впору замуж выходить, а я тут с вашими детьмя нянчиться должна*.

Итак, в первом приближении лексему *впору 2* можно истолковать так:

(15) *Человеку A1 впору P, а не Q* 'По мнению говорящего, из множества возможных ситуаций субъекту A1 больше всего подошла бы ситуация P; P не имеет места; говорящий противопоставляет ситуацию P ситуации Q, которая имеет место; говорящий неодобрительно относится к A1'.

Данная лексема довольно близка по смыслу лексеме *пора*, представленной в контекстах типа *Ему пора жениться*. Однако лексема *пора* вполне подвергается отрицанию, ср. *Ему не пора жениться*. Что касается лексемы *впору*, то в ее значение оценочный компонент 'больше всего' — он и мешает взаимодействию данной лексемы с отрицанием.

Действительно, обсуждаемая семантическая структура могла бы иметь следующий вид:

(16) 'НЕ (подходить больше всего)'.

В этой структуре отрицаются сразу два предиката: 'подходить' и 'больше всего'. Рассмотрим возможность этой структуры с точки зрения постулатов Грайса.

Если отрицается соответствие, то заведомо отрицается и его максимальная степень. Значит, достаточно отрицать соответствие, т. е. сказать *не подходит*. С этой точки зрения сочетание **не впору* не употребляется, т. к. в его семантической структуре содержится лишний компонент.

Отрицание в (16) может относиться только к компоненту 'больше всего', и тогда (16) может значить 'подходит, но не больше всего' или 'подходит меньше всего'. Как видим, смысл 'больше всего' вполне сочетается с отрицанием, если он выражен отдельным словом или выражением. В нашем случае, однако, этот смысл «впаян» внутрь значения лексемы *впору 2*. Частица *не* с ним непосредственно не взаимодействует. В подобных случаях требуется отдельно выразить смысл 'подходит' и отдельно — смысл 'меньше всего'.

Есть и еще одна причина, по которой невозможно высказывание типа *Кому-л. не впору Р*. Возьмем его потенциальный смысл: ‘По мнению говорящего, ситуация Р, не имеющая места, меньше всего подходит субъекту; говорящий противопоставляет Р реальной ситуации Q’. Неясен коммуникативный замысел говорящего. Действительно, произнося высказывание *Человеку А1 впору Р*, говорящий противопоставляет желаемое положение дел реальному, что вполне естественно; ср. *Тебе нужно делать то-то и то-то, а ты этого не делаешь*. Непонятно, однако, зачем противопоставлять реальности нежелаемое положение дел, наименее всего подходящее субъекту.

Как видим, лексемы *впору 1* и *впору 2* устроены аналогично — обе они указывают на максимальную степень соответствия, ср. компонент ‘в точности’ в значении лексемы *впору 1* и компонент ‘больше всего’ в значении *впору 2*. Оба оценочных компонента мешают свободному взаимодействию данных лексем с отрицанием.

2.3. «Такая тоска — впору утопиться»

Примеры:

(17) *У нас жарища, впору ходить в плавках и сидеть по горло в воде* [Юлий Даниэль. Письма из заключения].

(18) — *Удивляюсь я на вас, Христофор Бонифатьевич: тут впору в сосульку обратиться, а вы еще фокусами развлекаетесь* [Андрей Некрасов. Приключения капитана Врунгеля].

Контексты данного типа можно представить в виде: *Q, впору Р*. Иными словами, лексема *впору 3* предполагает две ситуации. Первая из них (Q) выражена в предтексте, ср. *У нас жарища*. (Данный предтекст может быть достаточно отдаленным, ср. (18)). Вторая ситуция (Р) вводится лексемой *впору 3* и выражается инфинитивом. Связь между ситуациями Р и Q такова: Q так сильно влияет на субъекта, так что он готов совершить действие Р. При этом субъект Р обычно совпадает с говорящим. Высказывание типа *Ему там страшно скучно, впору запить* предполагает эмпатию к субъекту или передачу его речи. Более точно:

(19) *Q, впору Р* ‘Ситуация Q влияет на субъекта-говорящего так сильно, что он готов совершить Р; по его мнению, Р — это единственное, что он может сделать при наличии Q; при обычном положении дел ситуация Р для него невозможна’.

Теперь представим себе отрицание (19): *Q, не впору Р*. Подвергается отрицанию первый компонент толокования (19), ср.

(20) ‘Ситуация Q не влияет на субъекта-говорящего так сильно, что он готов совершить Р’.

Компонент (20), однако, вступает в противоречие с компонентом модальной рамки: 'по мнению субъекта, P — это единственное, что он может сделать при наличии Q'. Такова причина, по которой данная лексема *впору* не сочетается с отрицанием. Целый ряд предикатов, которые не сочетаются с отрицанием потому, что создается логическое противоречие компонента толкования с модальной рамкой, описан в работе [Апресян 1978/1995].

3. Заключение

Мы попытались рассмотреть три употребительные в современном русском языке лексемы наречия *впору*. В силу специфики своего значения ни одна из них не сочетается (по крайней мере, вполне свободно) с отрицанием *не*. Однако конкретная семантическая причина, препятствующая взаимодействию лексемы с отрицанием *не*, может быть в каждом случае своя. В случае лексемы *впору 3* (ср. *Такая тоска — впору утопиться*) отрицание порождает логическое противоречие между модальной рамкой лексемы и ее отрицаемым компонентом, а это влечет языковую аномалию высказывания. Эта причина языковой аномалии, в частности аномальности сочетания некоторых предикатов с отрицанием, описана Ю. Д. Апресяном [1978/1995]. Плохая сочетаемость с отрицанием лексем *впору 1* (ср. *Эти сапоги мне впору*) и *впору 2* (ср. *Ей впору генеральным директором быть*) объясняется через постулаты Грайса. Коммуникативные постулаты Грайса прежде всего описывают правила реального употребления высказываний для реализации коммуникативного замысла говорящего. Однако будучи приложимы к словосочетанию, они могут объяснить причину языковой аномалии, в частности — причину плохой сочетаемости предиката с отрицанием.

Остановимся на уходящей из языка лексеме *впору 1*.

Уход какой-либо единицы из языка — это постепенный процесс. Сначала данная единица, например лексема, перестает употребляться вполне свободно и закрепляется за каким-то кругом контекстов. Это влечет некоторую перестройку ее значения: лексема как бы вбирает в себя некоторые компоненты контекста (возможно — достаточно широкого). Благодаря этому ее значение обогащается новыми компонентами. Ситуация парадоксальна: уходя из языка, лексема становится более сложной по семантике, более конкретной и, как следствие, более лингвоспецифичной. Этот процесс прямо прпотивоположен грамматикализации, когда лексема начинает употребляться во все более широком круге контекстов, так что ее значение становится беднее и абстрактней. Отсюда, однако, не следует, что наиболее употребительные лексемы наименее лингвоспецифичны.

Данная работа является первым этапом описания слова *впору*. В современном языке встречаются и другие контексты с этим словом, которые раньше, по-видимому, были гораздо более употребительны. Предстоит задача описать данные контексты и на основании корпусных данных представить процесс ухода из языка данного слова. Другая задача состоит в том, чтобы на более широком

материале продемонстрировать, что действие постулатов Грайса распространяется и на семантическую структуру слововсочетания, причем нарушение этих постулатов в случае словосочетания влечет за собой языковую аномалию.

Литература

1. *Апресян Ю. Д.* Языковая аномалия и логическое противоречие // Апресян Ю. Д. Избранные труды. Т. II. М., 1995.
2. *Баранов А. Н., Плунгян В. А., Рахилина Е. В.* Путеводитель по дискурсивным словам русского языка. М., 1993.
3. *Григорьева С. А.* Словарная статья «СОВЕРШЕННО 1, АБСОЛЮТНО, СОВСЕМ 1» // Новый объяснительный словарь русского языка / Под общим руководством Ю. Д. Апресяна. 2-е изд. М. — Вена, 2004.
4. *Григорьева С. А.* Словарная статья «ПОЧТИ, БЕЗ МАЛОГО, ЧУТЬ НЕ, ЕДВА НЕ» // Новый объяснительный словарь русского языка / Под общим руководством Ю. Д. Апресяна. 2-е изд. М. — Вена, 2004.
5. *Иорданская Л. Н.* Перформативные глаголы и риторические союзы // Wiener Slawistischer Almanach. Sonderband 33. Wien, 1992.

НЕЗАВЕРШЕННОСТЬ ПРЕДЛОЖЕНИЯ VS. НЕЗАВЕРШЕННОСТЬ ТЕКСТА: АКЦЕНТЫ И АКЦЕНТОНОСИТЕЛИ

Янко Т. Е. (tanya_yanko@list.ru)
Институт языкознания РАН, Москва, Россия

Ключевые слова: устная речь, связный текст, просодия, незавершенность предложения, незавершенность текста, тема, рема, дискурс

SENTENCE INCOMPLETENESS VS. DISCOURSE INCOMPLETENESS: PITCH ACCENTS AND ACCENT PLACEMENT[1]

Yanko T. E. (tanya_yanko@list.ru)
Institute for Linguistics, Russian Academy of Sciences, Moscow, Russia

The prosodic cues for discourse incompleteness may be either identical with the prosodic means expressing the topic or independent of marking the communicative constituents of a sentence: the topic or the focus. The autonomous prosodic marking of discourse incompleteness becomes possible in the context of tails. A tail is a fragment of a sentence placed after the accent-bearer of the focus. (Thus in the sentence *Malo ja smyslju v muzhskoj krasote* 'Little I know about men's attractiveness' with *malo* 'little' as the accent-bearer of the focus the fragment *ja smyslju v muzhskoj krasote* is the tail). A tail may be either deaccented or it may be used to carry the rise of discourse incompleteness. Generating a tail is conditioned by activation of entities within a sentence, contrast, emphasis, and verification expressed either by lexemes or by prosody, or both. In Russian, a tail can also result from a specific word order transformation with the focus accent-bearer being shifted to the left in front of the finite verb. The sentence-final verb, therefore, transforms into the tail to be specifically used as the bearer of discourse incompleteness pitch accent. (Thus in the sentence *Ja pidzhak snjal*... literally: 'I my coat took off...' with *pidhzak* 'coat' as the accent-bearer of the focus the sentence-final verb *snjal* 'took off' is the tail). Sentences with tails are able, therefore, to display a full set of communicative meanings including topics, focus and discourse incompleteness expressed by separate accent-bearers carrying the respective pitch accents.

Key words: prosody, pitch accents, incompleteness, discourse, word order, topic, focus, oral speech

[1] This work has been supported by the Program of fundamental research "Language and literature in the context of social dynamics", 2012–2014, project "The computer model and electronic data base ‚Meaning-Spoken Speech'" and the Ministry of science and education of Russia, project 8009.

The prosodic markers of an uncompleted sentence and an uncompleted piece of narration are often identical. Sentence incompleteness and discourse continuation can be both expressed by a rise of the fundamental frequency placed on a word selected from segmental material of a single sentence[2]. In this case, there is no difference between a topic of a sentence whose focus is placed in the subsequent context and a sentence which is a component of an uncompleted narration: the applicable pitch accents and the due accent-bearers are identical. For instance, in the sentence *My priehali v Moskvu toljko v 1990 godu* 'We came to Moscow only in 1990' the segment *my priehali v Moskvu* 'we came to Moscow' with the rise on *Moskvu* is the topic, while in the story about a trip to Russia *My priehali v Moskvu. Potom avtobus otvëz nas v Bekasovo. Vecherom byl banket* 'We came to Moscow. Then a bus took us to Bekasovo. The party was in the evening.' *My priehali v Moskvu* is a non-final component of a connected text[3]. Topics, therefore, can have a syntactic structure of a sentence. Hence, whether a unit has a sentential structure or it is, for instance, a noun phrase cannot be decisive for distinguishing pieces of narration from topics.

A question then arises as to whether there could be cues for discourse continuation distinct from sentence continuation markers. The aim of the present paper is to provide a description of a set of prosodic patterns that are characteristic solely of discourse incompleteness.

The data for analysis are taken from oral corpora "Night dream stories" (see Kibrik, Podlesskaja (2009), http://spokencorpora.ru) "Stories about gifts and skiing" (see Podlesskaja (2012)), and the corpus of records based on TV and radio interviews, eyewitnesses accounts and examples of professional actor readings prepared by the author.

The prosody is explicated here in terms of the fundamental frequency changes as it is proposed in D. Bolinger (1958) and E. A. Bryzgunova (see Russkaja grammatika 1982: 97–101) rather than in terms of target levels as in J. B. Pierrehumbert (1980).

The instrumental study has been carried on by using the computer system of oral speech analysis Speech Analyzer.

1. The rise of frequency as a generalized marker of incompleteness. The strategy of serial topics

Major or minor rises[4] of frequency within a sentence can mark either the topic (as opposed to the focus of a sentence) or a link between a non-final syntactic component and other components within a syntactic construction. For instance, in the sentence *Prishla vesna* (literally: 'Came the spring') the rise on the sentence-initial word *prishla* 'came' marks the onset of a syntactic construction, while in *Vesna prishla*

---

[2] The principles of accent placement in communicative constituents of various lexical and syntactic structures are listed in Yanko (2008: 38–60).

[3] Details of identical expression of sentence and discourse incompleteness will be explicated by example (1) below.

[4] I do not focus here on various types of continuation rises in spoken Russian. The complete inventory is described in detail in Yanko (2008: 128–163).

tol'ko v mae (literally: 'The spring came only in May' the rise on *vesna* 'spring' is the marker of the topic. At the same time a rise can also express discourse incompleteness, i.e. the idea that the current step of narration is not final. To consider this point in details we need an example. The words carrying relevant pitch accents in examples below are underlined, the tonic syllables of the accent-bearers are capitalized.

(1) *Potom ja podoshla v* *druguju kOmnatu,* *vot u menja* *vybito steklO,*
Later I entered into another room, here at me broken window,

no zApaxa gaza *ne oshchushchAju.*
but smell gas. GEN not feel.1SG[5]

'Later I entered another room, I have got a broken window glass here, but as for the smell of gas, I do not feel any'

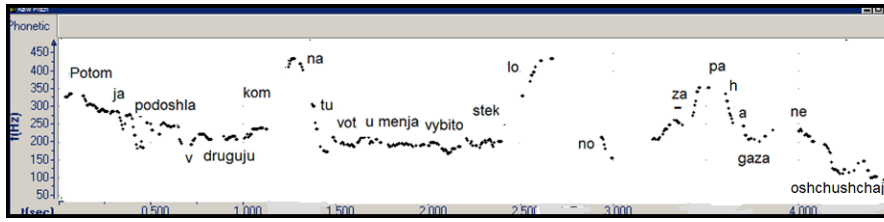


Fig. 1. Frequency tracing of sentence (1)

In example (1) from a gas explosion report the segments1.1 *Potom ja podoshla v druguju kOmnatu* 'Later I entered another room' with the accent-bearer *kOmnatu* 'room', 1.2 *vot u menja vybito steklO* 'Here I have a broken window' with the accent-bearer *steklO* 'window glass', and 1.3 *zApaxa gaza* 'smell of gas' with the accent-bearer *zApaxa* 'smell' have identical rises on the stressed syllables of the accent-bearers which are followed by frequency falls on the post-tonics (if any). These segments (1.1)–(1.3) can, therefore, be viewed as serial topics referring to the same focus *ne oshchjushchAju* 'I do not feel'. Indeed, in (1) there is no principal distinction between the prosody of the initial segments1.1 or 1.2 which really forward the narration ahead and the final topic *zapaxa gaza*.

However, while the strategy of "serial topics" is one of the most common mechanisms of discourse linkage in many languages, it cannot help deciding whether a prosodic constituent is a topic or it serves as a discourse constituent of an uncompleted piece of discourse. In this case, the prosody cannot distinguish between a communicative unit of a sentence — namely its topic — and a valid discourse unit.

[5] Details of identical expression of sentence and discourse incompleteness will be explicated by example (1) below.

2. Autonomous marking of topics and discourse links

The strategy of serial topics is the most common but not the only mechanism of expressing discourse incompleteness. A variety of strategies employing separate cues for sentential and discourse incompleteness is used. These strategies are based on specific accent-placement. Consider example (2) from the corpus "Night dream stories".

(2)	*A*	*kOshka,*	*ona*	*obidelas'*	*i*	*sprjAtalas'*	*ot*	*nashej*	*sobAchki…*
	And	cat	she	resented	and	hid	from	our	doggy
	'And	the cat,	it	got offended	and	hid	from	our	doggy…'

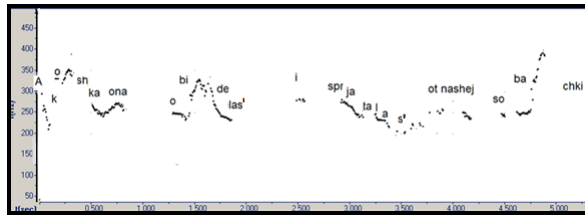


Fig. 2. Frequency tracing of sentence (2)

In example (2) the word *sprjAtalas'* 'hid' carries a fall of frequency designating the focus, while the rise on the sentence-final word *sobAchki* 'doggy.GEN' is a separate marker of discourse incompleteness. It shows that the narration is to be continued. The non-final position of the focus accent-bearer the word *sprjAatalas'* 'hid' in sentence (2) is accounted for by the fact that the argument *ot nashej sobachki* 'from our doggy' refers to the activated[6] knowledge of the hearer and is not, therefore, embodied in the focus: the dog appeared on the scene at the preceding stages of narration.

Sentence (2), therefore, has a full set of communicative meanings that shape a sentence as a speech act and at the same time as a component of a connected discourse. These are: the meaning of a topic (expressed by the rise on the tonic syllable of the accent-bearer of the topic *kOshka* 'the cat'), the meaning of the focus (designated by the fall on the accent-bearer of the focus *sprjAtalas'* 'hid'), and the discourse incompleteness (the rise on the accent-bearer of incompleteness *sobAchki* 'doggy.GEN'). The marker of discourse incompleteness is placed here after the marker of the focus. This order becomes feasible because the non-final placement of the focus accent-bearer leaves a sufficient segmental material free of any pitch-accents relevant for the topic-focus structure of the sentence.

E. Vallduvi and E. Engdahl (1996) employ the term "tail" for a fragment of a sentence placed after the accent-bearer of the focus. Consider their example:

---

[6] The notion of activation is used here on the terminology of Dryer (1996).

(3) *You shouldn't have brought chocolates to the president. He HATES chocolates.*

In example (3), *hates* is the focus, while the second occurrence of the word *chocolates* is the tail. Here, the tail is a fragment of a topic because it is borrowed from the first sentence and refers to the activated knowledge, yet it is placed sentence-finally because of the basic word order SVO characteristic of English. In (3), the tail *chocolates* does not carry any relevant pitch accents (as it is expected from a topic placed after the focus).

In example (4) from the reading of "Drama na ohote" ('The shooting party' by A.P. Chekhov), the tail is not a topic, as in (3), but a fragment of the focus placed after the accent-bearer of the focus:

(4)	*Ego*	*bol'shoe*	*muskulistoe*	*litso*	*ostalos'*	*navsegdA*	*v*	*moej*	*pamjati*
	'His	big	sinewy	face	left	forever	in	my	memory'

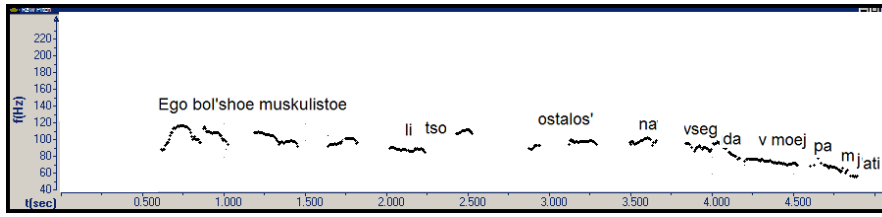


Fig. 3. Frequency tracing of sentence (4)

In sentence (4), the accent-bearer of the focus is the word *navsegdA* 'forever', while the fragment *v moej pamjati* 'in my memory' is placed sentence-finally and does not carry any relevant pitch accents. *V moej pamjati* here is the tail. The presence of lexically conditioned focus — the quantifier *navsegda* 'forever' — results in the non-final position of the focus accent-bearer and, as a consequence, in forming a tail; about the focus function of the Russian quantifiers see Bulygina, Shmelev (1997: 200–207). The prepositional phrase *v mojej pamjati* is the sentence-final fragment of the focus.

In examples (3) and (4) the tails are deaccented, while in (2) the tail *ot nashej sobachki* 'from our doggy' is used for marking the discourse link. So, a tail may consist either of a topic or of a fragment of the focus (remained after the accent-bearer of the focus). It may be either deaccented, as in (3) or (4), or it may carry the rise of discourse incompleteness as in sentence (2). Fig. 4 below exemplifies the difference between interpretations of an example from "Drama na ohote" read by two professional speakers: Aleksandr Balakirev (the upper panel) and Petr Korshunkov (the lower panel). In Balakirev's reading the tail is deaccented: it does not carry any relevant pitch accents. The idea that the fragment of Chekhov's text has not come to its logical end is only designated here by a frequency value of the final boundary tone which is slightly higher than the baseline of the speaker's voice. Whereas Korshunkov explicitly shows that the text is to be continued by a prominent rise of frequency on the sentence-final word'.

(5) *MAlo* *ja* *smyslju* *v* *muzhskoj* *krasotE*
Little I understand in men's attractiveness
'I know so little about men's attractiveness'

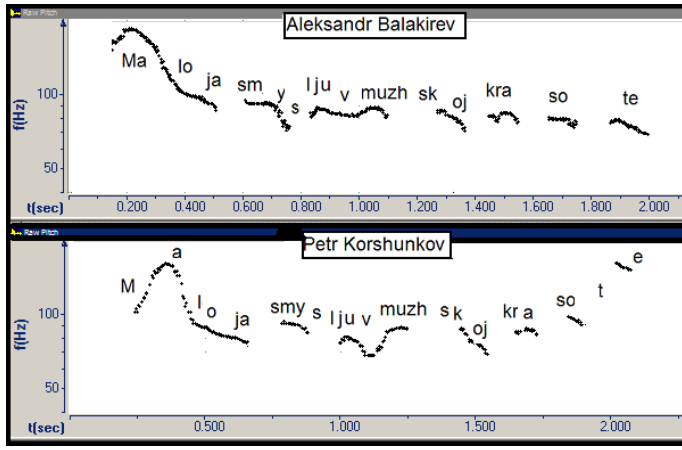


Fig. 4. Contrastive frequency tracings of sentence (5) read by two different readers

In both readings, the sentence begins with a high fall on the focal word *malo* 'little'. *Malo* is the accent-bearer of the focus in sentence (5).The fragment *ja smyslju v muzhskoj krasote* 'I know about men's attractiveness' is, therefore, the tail. Balakirev does not use the tail to designate incompleteness, while Korshunkov prefers to demonstrate explicitely that the current event of narration is not logically final. The focus structure of sentence (5) strongly depends on the semantics of the "focal word" *malo* 'little'; about the communicative functions of the Russian words of low quantity see Bulygina, Shmelev (1997: 205), about French *peu* see Ducrot (1973).

The reading strategy with fewer rises of discourse incompleteness — either placed in the tails of sentences or appearing as topics — is a characteristic parameter of professional readers' performance whose academic reading suggests that the listener makes a mental pause after each step of narration by using a "full stop" in order to reflect on the text. While in spontaneous speech or in its artistic imitating, the speaker — being afraid of losing the listener's attention — demonstrates by using "comma" strategies at every non-final step of narration that the discourse is not over. In this respect, in Balakirev's reading — as it is demonstrated by sentence (5) — and by other examples of his reading of "Drama na ohote", a measured style of reading with regular pauses and full stops dominates as if the speaker does not care much about whether the hearer is listening to the story or not.

A striking example of expressing a discourse link in the tail is sentence (6) from the corpus "Stories about gifts and skiing".

(6) *PokatAlsja* *On* *ne* *Ochen'* *udAchno…*
Skied he not very successfully
'His skiing was not a success'

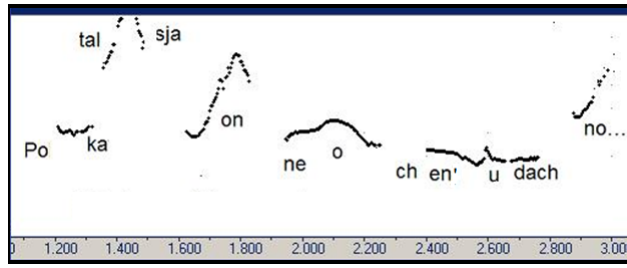


Fig. 5. Frequency tracing of sentence (6)

Sentence (6) is very short and has a syntactic structure of a simple sentence, nevertheless the two topics, the focus, and the discourse incompleteness have separate prosodic markers here. The verb *pokatAlsja* 'skied' (which carries a prominent rise on the stressed syllable) is the initial topic, the pronoun *on* 'he' is the second topic, the adverbial *ne Ochen'* 'not very' is the accent-bearer of the focus. It carries a specific accent of an emphatic focus. The emphatic prosody correlates here with the semantics of the word *ochen'* 'very'; about the prosody of emphasis see Yanko (2008: 83). The sentence-final *udAchno* 'successfully' is, thus, placed after the accent-bearer of the focus. It is, therefore, a tail, and it carries a specific type of a rise placed on the post-tonic syllable of the accent-bearer, while the tonic syllable carries a low level tone.

A discourse link placed in the tail can, in its turn, have its own tail. Consider example (7) from "Night dreams stories".

(7) *<Nas vseh* *razognali,* *skazali>* *nel'zjA* *gribY* *rvat'.*[7]
us.ACC all.ACC drove away told.they forbidden mushrooms.ACC to pluck
'We all were driven away; they told us that picking mushrooms was forbidden'

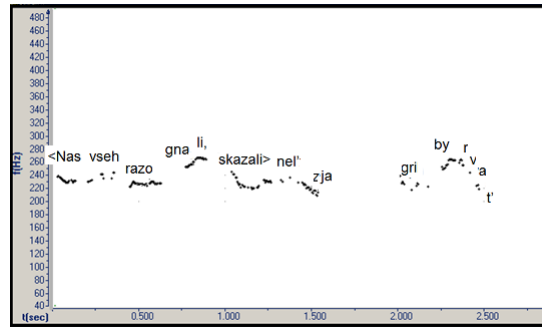


Fig. 6. Frequency tracing of sentence (7)

[7] The angle brackets designate the required context.

Here, the fall of frequency on the tonic syllable of the word *nel'zjA* 'forbidden' marks the focus of the sentence. The infinitive phrase *gribY rvat'* 'to pick mushrooms' is, therefore, a tail. As a tail it is employed for designating a discourse link: the frequency tracing displays the rise on the accent-bearer of the discourse link the word *griby* 'mushrooms'[8]. The sentence final infinitive *rvat'* 'to pluck, to pick' is, therefore, an atonic tail within the enclosing tail *griby rvat'* 'mushrooms to pick'.

3. A sentence-final finite verb as a marker of discourse linkage

In examples above the sources of the tails were either the sentence-final position of the activated entities (as in example (2)), or the non-final position of the focal words *navsegda* 'forever' (example (4)), *malo* 'little, few' (example (5)), and *nel'zja* 'forbidden' (example (7)), or the presence of the emphatic word *ochen'* 'very' (example (6)). There could be found other sources of tails, i.e. the conditions in which the focal pitch accent occupies the non-final position. For more comprehensive list of these contexts see Yanko (in press). The common parameters of the sentences viewed above are: 1) the sentences have sufficient segmental material of the tails to designate discourse linkage irrespective of designating the topic and the focus; 2) the finite verb which has arguments preserves its basic non-final position in a sentence. In this section, the cues for discourse incompleteness are analyzed which require the argument shift and, therefore, change the basic Russian word order.

The Russian spontaneous speech developed a specific strategy of separate marking the discourse linkage by a rise of frequency placed on the sentence-final finite verb. The sentence-final verb in this case serves as a tail to carry a discourse link. Its final position is attained by shifting the argument to the left in front of the finite verb. The basic word order SVO is, therefore, substituted for SOV (or OSV). This change is accounted for by the fact that a verb with two (or more) non-activated arguments generally has one of its arguments as the accent-bearer of the focus. The verb itself, therefore, does not carry any pitch accents relevant for designating the focus of a sentence. As a consequence, the verb is a component of a sentence the most free (statistically) of fulfilling the function of the focus accent-bearer. Being shifted to the left the focus accent-bearer paves the way for the verb to serve as a tail. Consider example (8) from "Night dreams stories" where the sentence-final verb is used for discourse linkage.

(8) *Ja iz kOmnaty vyhozhU, <kogda vhozhu ona uzhe napolovinu pustaja>*
I from room leave.PRES1SG when enter.PRES1SG it already half-empty
'I leave the room, when I come back it (a bottle of wine) is already half-empty'

[8] I do not focus here on various types of continuation rises in spoken Russian. The complete inventory is described in detail in Yanko (2008: 128-163).

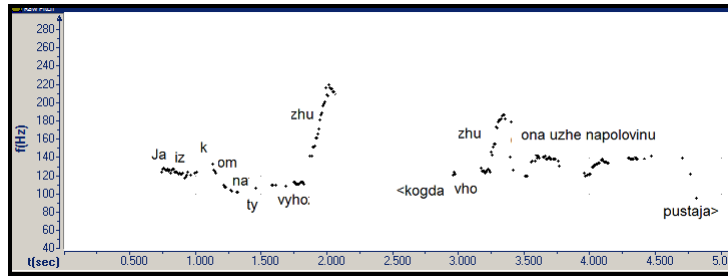


Fig. 7. Frequency tracing of sentence (8)

In sentence (8) the focus accent-bearer *iz komnaty* 'from the room' carries a fall, while the sentence-final verb *vyhozhU* 'I am leaving' carries a prominent rise of discourse incompleteness. The word order in sentence (8) is, therefore, SOV. The basic word order SVO for the syntactic structure of sentence (8) is displayed by sentence (8.1):

(8.1) *Ja* *vyhozhu* *iz* *kOmnaty*
I leave from room
'I am leaving the room'

Designating incompleteness by a rise on the sentence-final finite verb requires, therefore, a word order transformation. In spontaneous speech this strategy is highly frequent. Consider one more example (9) from the same corpus.

(9) *Kogda* *obrAtno* *uzhe* *bezhAli…*
When back already ran
'When we were already running back…'

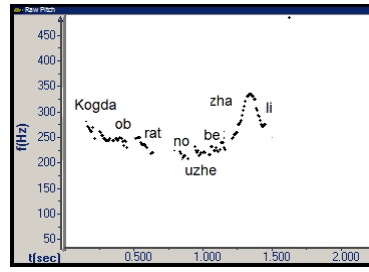


Fig. 8. Frequency tracing of sentence (9)

In sentence (9), the focus accent-bearer the word *obrAtno* 'back' carries a fall on the tonic syllable followed by a subsequent fall on the post-tonic syllable, while the sentence-final verb *bezhAli* 'were running' carries a prominent rise on the tonic syllable followed by a fall on the post-tonic. The basic word order for sentence (9) is represented by sentence (9.1):

(9.1) *Kogda* *uzhe* *bezhali* *obrAtno* …
when already ran back
'When we were already running back…'

The sentence-final position of a finite verb which carries a specific rise of frequency to designate the discourse linkage can only be interpreted as a result of a transformation because a verb with arguments even in Russian (with its unlimited scrambling) generally is not placed sentence-finally. I assume that this transformation is a specific mechanism of discourse linkage elaborated by the Russian spontaneous speech. In more formal styles of communication it is not employed.

***

The prosodic cues for discourse incompleteness may be either identical with the prosodic means expressing the topic or independent of marking the communicative constituents of a sentence: the topic or the focus. The autonomous prosodic marking of discourse linkage becomes possible in the context of tails. Generating a tail is conditioned by the basic topic-focus structure of a sentence, activation of entities within a sentence, contrast, emphasis, and verification expressed either by lexemes or by prosody, or both. In Russian, a tail can also result from a specific word order transformation with the focus of a sentence being shifted to the left in front of the finite verb. The sentence-final verb, therefore, transforms into the tail to be specifically used as a bearer of discourse incompleteness pitch accent.

References

1. *Bolinger D.* (1958) A Theory of Pitch Accent in English. Word. Vol. 14.
2. *Bulygina T., Shmelev A.* (1997) Jazykovaja kontseptualizatsija mira (na materiale russkoj grammatiki) [On the linguistic conceptualization of the world]. School "Yazyki russkoj kul'tury", Moscow.
3. *Dryer M. S.* (1996) Focus, pragmatic pressupposition, and activated propositions. Journal of pragmatics 26. Pp. 475–523.
4. *Ducrot O.* (1973) French *peu* and un *peu*. A semantic study. Generative Grammar in Europe. Dordrecht.
5. *Kibrik A., Podlesskaja V.* (2009) Rasskazy o snovidenijah: Korpusnoe issledovanie ustnogo russkogo diskursa [Nightdream stories: A corpus-based study of spoken Russian discourse]. Jazyki slavjanskih kul'tur, Moscow.
6. *Pierrehumbert J. B.* (1980) The Phonology and Phonetics of English Intonation. PhD thesis. MIT. Amherst.
7. *Podlesskaya V.* (2012) Russian complement clauses in prosodically annotated spoken corpora. The Fifth international conference on cognitive science. Abstracts. Vol. 2. Kaliningrad. P. 783–785.
8. *Russkaja grammatika* [Russian grammar] (1982). Vol.1. Nauka. Русская грамматика Т. 1, М., Наука.
9. *Vallduví E., Engdahl E.* (1996) The linguistic realization of information packaging. Linguistics, 34. Pp. 459–519.
10. *Yanko T.* (2008) Intonatsionnye strategii russkoj rechi [Intonational strategies of the Russian speech]. Jazyki slavjanskih kul'tur, Moscow.
11. *Yanko T.* (in press) Linejno-aktsentnye struktury rem [Linear-accentual structures of a focus]. Aktual'nye voprosy teoreticheskoj i prikladnoj fonetiki [Current issues in theoretical and applied phonetics]. Moscow State university publishing houses. Moscow

COMPARISON OF OPEN INFORMATION EXTRACTION FOR ENGLISH AND SPANISH

Zhila A. (alisa_zh@mail.ru), **Gelbukh A.** (www.gelbukh.com)

Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico

Open Information Extraction (IE) is the task of extracting relational tuples representing facts from text, with no prior specification of relation, no pre-specified vocabulary, or a manually tagged training corpus. Part-of-speech based systems are shown to be competitive with parsing-based systems on this task and work faster for large-scale corpora. Nevertheless, implementation of such a system requires language-specific information. So far, all work has been done for English. We present a relation extraction algorithm for Open IE in Spanish, based on POS-tagged input and semantic constraints. We provide a description of its implementation in an Open IE system for Spanish ExtrHech. We compare its performance with Open IE systems for English, including a comparison on a parallel English-Spanish dataset, and show that the performance is comparable with the state-of-the-art systems, while the system is more robust to noisy input. We give a comparative analysis of errors in extractions for both languages.

Keywords: open information extraction, fact extraction, cross-lingual, Spanish, English

1. Introduction

Traditional Information Extraction (IE) is focused on detection of precise, pre-specified type of information that would satisfy requests narrowed to a certain domain or area of activity. For example, an IE system could be trained for extraction of information of some certain classes, e.g. *ACQUIRE*(*argument*$_1$, *argument*$_2$, …, *argument*$_n$) with a fixed number of arguments *n*. Normally, an IE system would learn an extractor from a large tagged corpus for a specific relation marked up by human annotators or in a semi-supervised manner [8, 10, 14]. Although this approach might be efficient for a certain target relation or classes (e.g., *people* or *cities* as in [10]), it requires very expensive resources for training and, more importantly, this approach does not scale to large corpora such as the Web, where the number of possible relations is very large or where the target relations cannot be specified beforehand.

Open information extraction (Open IE) was introduced by Banko et al. [2] in 2007 as a new extraction paradigm that facilitates domain independent discovery of relations in text and can be readily scaled to a large and versatile corpus such as the Web. An Open IE system extracts *all* possible relations and assertions without requiring any prior specification of relations, manually tagged training corpora, example seeds tailored for the target relations, or any other relation-specific input. This

guarantees scalability, and the system can satisfy unanticipated user needs. Open IE is necessary when the number of relations is large and the relations are not pre-specified [3]. Consequently, it can serve purposes distinct from the traditional IE: fact extraction at sentence level (e.g. <*Mozart*, *was born in*, *Salzburg*>), new perspective on search as question answering (e.g. "*Where was Mozart born?*") in an unrestricted form [6], or assessment of the quality of text documents at the Web scale [7].

Independency from relation pre-specification is achieved through the implementation of a compact set of relation-independent lexico-syntactic patterns that allow identification of *arbitrary* relations [2]. However, the patterns are language dependent. All previous work in this field has been done for English [4, 6, 12, 15]; no language-related issues not specific for English have been addressed.

We present an Open IE for Spanish *ExtrHech*, compare its performance with that of a similar Open IE system for English ReVerb [6] on a parallel dataset, and perform analysis of errors for both languages.

The paper is organized as follows. Related work is reviewed in Section 2. Section 3 presents the Open IE approach for Spanish and describes the *ExtrHech* system. Section 4 describes the experimental results for two datasets in Spanish and a parallel English-Spanish dataset. In section 5, the analysis of errors is presented. Section 6 draws the conclusions and outlines future work.

2. Related Work

Open IE is the task of extracting arbitrary relations with their corresponding arguments from text without pre-specification of relations or manually tagged training corpora. The first step of any Open IE system is extraction of relations from a sentence. For example, in a sentence "*The policeman saw a boy who was crossing the street*", two assertions can be identified: <*the policeman*, *saw*, *a boy*> and <*a boy*, *was crossing*, *the street*>. A large corpus of text such as the Web is highly redundant, and many assertions are expressed repeatedly in different forms. After being encountered many times in various sources, an assertion has a significantly higher probability to be true.

The basic idea is that most sentences contain highly reliable syntactic clues to their structure [2]. There are three major approaches to relation identification in Open IE.

1. *Self-supervised learning* involves three steps: automatic labeling of relations using heuristics and distant supervision; learning of a relation phrase extractor; and extraction, for which a candidate pair of arguments is detected and then a relation extractor is applied to detect a relation between these arguments. Examples of such systems are TextRunner [2], WOE^{pos}, and WOE^{parse} [15]. One of its shortcomings is that potential arguments are detected before the relation is defined and cannot be backtracked. Therefore, a noun that actually belongs to a relation phrase can be marked as an argument. For example, in the relation "*to make a deal with*", *deal* can be incorrectly recognized as an argument. Consequently, the output of such systems contains many incoherent or uninformative extractions.
2. *Context analysis*, implemented in OLLIE system [12]. This approach overcomes various limitations of the other approaches. First, it extracts not only relations

expressed via verb phrases, but also relations mediated by adjectives, nouns, etc. Second, it is not limited to binary relations and can detect more than two arguments of a relation. Yet deeper context analysis requires syntactic parsing, which is time- and resource-consuming and makes real-time processing at Web scale impractical. Syntactic parsing analysis using heuristic rules is also implemented in the Open IE for Spanish FES [1].

3. *Syntactic and lexical constraints* implemented in the form of rules, as in ReVerb [6]. In contrast to the first approach, it initially detects a verb phrase, and then searches for its possible arguments, which reduces incoherent and uninformative extractions. It also has more light-weight implementation and faster execution time than context-analysis systems because it is based on part-of-speech analysis.

These approaches have been evaluated only for English. However, their relation extraction algorithms are language dependent: they use either part-of-speech or syntactic dependency information. It is not known how language affects implementation and output of Open IE.

We present a relation extraction algorithm for Open IE in Spanish, following the third approach. Since the datasets used for evaluation of the Open IE systems in [2, 6, 12] (300 to 500 sentences) are not available, we have created two comparable datasets for evaluation of our system.

3. ExtrHech, an Open IE system for Spanish

ExtrHech, an Open IE system for Spanish, is a POS-tag based system using syntactic and lexical constraints; see Figure 1.

Fig. 1. Processing pipeline of ExtrHech

The system takes a POS-tagged text as an input. For POS-tagging we used Freeling-2.2 [13] that uses EAGLES POS-tagset for Spanish.

ExtrHech performs sentence-by-sentence processing. First, it looks for a verb phrase that is limited to be either a single verb or a verb immediately followed by dependent words till a preposition (*nació en*) or a preposition followed immediately by an infinitive (*sirven para acentuar*). The expression for a verb phrase is:

$$\text{VREL} \rightarrow (\text{V W*P})|(\text{V}),$$

where V stands for a single verb possibly preceded by a reflexive pronoun (*se caracterizaron*), or a participle (*relacionadas*); V W*P stands for a verb with dependent words, where W can be a noun, an adjective, an adverb, a pronoun, or an article, and P stands for a preposition possibly immediately followed by an infinitive. Here * stands for zero or more occurrences, | stands for choice of a variant; ? (see below) stands for zero or one occurrence.

Next, the system looks to the left of the verb phrase for a noun phrase that could be a first argument in a relation. Then it searches to the right from the verb phrase for a second argument. The following expression describes noun phrases in ExtrHech:

$$\text{NP} \rightarrow \text{N (PREP N)?},$$

where N stands for a noun optionally preceded by an article (*los gobernantes*), an adjective or ordinal number (*los primeros homínidos*), a number (*3.5 millones*), or their combination, optionally followed by a single adjective (*el epíteto heroico*), a single participle (*las fuentes consultadas*), or both (*los documentos escritos antiguos*). PREP stands for a single preposition. In our system a noun phrase can be either a single noun with optional modifiers or a noun with optional modifiers followed by a dependent prepositional phrase, consisting of a preposition and another noun with its optional modifiers (*la historia de la civilización romana*).

If a noun is followed by a participle clause terminating with another noun, then the participle phrase is resolved into a separate relational tuple. For an example,

(1) *Los egipcios se caracterizaron por sus creencias relacionadas con la muerte.*
"The Egyptians were characterized by their beliefs related with death."

gives two relational tuples:

(2) <Arg1 = *Los egipcios;* Rel = *se caracterizaron por;* Arg2 = *sus creencias*>
(3) <Arg1 = *sus creencias;* Rel = *relacionadas con;* Arg2 = *la muerte*>,

with (2) corresponding to the main verb of sentence (1) and (3) corresponding to the participle clause.

ExtrHech also resolves coordinating conjunctions for verbal and noun phrases into independent relations or arguments correspondingly. Relative clauses are also resolved into independent assertions. Lexical constraints currently limit the length of relational phrases to prevent over-specifying of a relation. We use EAGLES POS-tag set and properly treat reflexive pronouns for verbal phrases. Currently we do not tackle anaphora, zero subject construction, and free word order. Still ExtrHech's precision is comparable with that of other Open IE systems (see Section 4.1).

4. Experimental Results

4.1. Experiments on Different Spanish Datasets

We analyzed ExtrHech's performance on two datasets.[1] The first one, FactSpCIC [1], contains 68 grammatically and orthographically correct and consistent

[1] Both datasets are available on www.gelbukh.com/resources/spanish-open-fact-extraction.

sentences manually selected from school textbooks. The second one contains 159 sentences randomly extracted from CommonCrawl 2012 corpus [9], which is a corpus of web crawl texts from over 5 billion web pages. It contains the sentences in their original form as they were crawled from the Internet. As evaluated by a human judge, 36 sentences (22% of the corpus) were either grammatically incorrect or incoherent.

Two human judges independently evaluated each extraction as correct or incorrect. For FactSpCIC dataset, they agreed on 89% of extractions (Cohen's kappa k = 0.52), which is considered to be moderate agreement [11]. For the raw Web text dataset of 159 sentences, they agreed on 70% of extractions (Cohen's kappa k = 0.40), which is considered the lower bound of moderate agreement. The number of correct extraction was calculated as an average for the two judges.

Precision of the system is the fraction of returned extractions that are correct. *Recall* is the fraction of correct extractions in the number of all possible correct extractions. To estimate the latter, we made a list of all extractions that the system is expected to return. Then, this set was extended by the extractions returned by the system that both annotators considered correct. This gives a lower bound estimation of all possible extractions that could be detected in the datasets, which gives the *upper bound* for recall; see Table 1.

Table 1. Performance of ExtrHech system on a grammatically correct and on a noisy datasets

Dataset	Precision	Recall
FactSpCIC (grammatically correct)	87%	70%
Raw Web text (noisy)	55%	49%

4.2. Experiments on Parallel Spanish and English Dataset

To analyze differences in performance between systems for Spanish and English, we formed a parallel Spanish-English dataset. The original Spanish FactSpCIC dataset was manually translated into English by a professional human translator. Then, the fact extractor for Spanish ExtrHech was run on the 68 sentences in Spanish, and ReVerb was run on the English translation.

The evaluation of extraction for Spanish is presented in Section 4.1. The output in English was also evaluated by two human judges. The judges agreed on 85% of extractions (Cohen's kappa $\kappa = 0.60$), similar to 89% agreement for Spanish, which is the upper bound for the moderate agreement range and is slightly higher than $\kappa = 0.52$ for Spanish.

Precision and recall for the extraction in English were calculated as described in Section 4.1.1; see Table 2, where the number of correct extractions is averaged by the two human judges.

Table 2. Comparison of Open IE systems for Spanish and English on a parallel dataset

System	Precision	Recall	correct extractions	found extractions	expected extractions
ExtrHech (Spanish)	87%	70%	99.5	115	137
ReVerb (Englist)	76%	50%	71	93	139

For the parallel dataset, precision and recall for the English ReVerb system are lower than those for the Spanish ExtrHech. Yet this might be due to overadjustment of ExtrHech to the dataset, which was also used during development of the system (note that no learning was involved). However, the total number of assertions extracted by ReVerb is lower than the amount of extractions by ExtrHech. Thus the Spanish extractor is more robust than the English one. Higher number of expected extractions for the English dataset (139) is due to the absence of the zero-subject phenomenon in English.

To show that ExtrHech performs at the level comparable with the state-of-the-art Open IE systems, we provide the comparative data on performance of the Open IE systems described in Section 2 based on the data provided in [1, 6, 12] and ExtrHech for different datasets; see Table 3.

Table 3. Comparative data for various Open IE systems

System	Approach	Dataset (# of sentences)	Precision	Recall	Running Time
ExtrHech (Spanish)	syntactic and lexical constr. over POS-tagged text	FactSpaCIC (68)	0.87	0.73	< 5 min
		raw Web text (159)	0.55	0.49	
ReVerb (English)	syntactic and lexical constr. over POS-tagged text	FactSpaCIC (68), translated	0.76	0.50	< 5 min
		Yahoo (500)	0.87 0.60	at 0.20 at 0.50	
Text-Runner (English)	self-learning on POS-tagged text	Yahoo (500)	< 0.64	at >0	< 5 min
WOE^{parse} (English)	self-learning on parsed text	Yahoo (500)	0.87	at 0.15	hours
OLLIE (English)	context analysis of parsed text	news, Wikipedia, biology textbooks (300)	0.66–0.85	N/A (various yield levels from [11])	N/A, probably hours
FES (Spanish)	heuristic rules on parsed text	FactSpaCIC (68)	0.87	0.91	hours

Table 3 shows that for the dataset of raw Web sentences, ExtrHech's 0.55 precision and 0.49 recall are slightly lower than those of ReVerb system for the Yahoo

dataset (0.60 and 0.50 correspondingly). However, the Yahoo dataset is not available, so we do not know whether it is raw Web text including incorrect and incoherent sentences common for texts on the Internet that hinder fact extraction.

ExtrHech speed is at the same level as that of other POS-tag based systems. It is much faster than syntactic parsing based systems, which perform significantly slower although with better precision.

5. Error Analysis

To analyze errors in assertion extraction for ReVerb and ExtrHech for the parallel English-Spanish dataset FactSpaCIC, first, we compared the distributions of the types of errors found in extractions. The type classification was modified from [6] to clearly distinguish between error types and their reasons. Table 4 shows the fractions of each type of errors in the total number of extractions for both systems.

Table 4. Distribution of the types of errors in all extractions

System and total number of extractions	Incorrect relation phrase	Incorrect arguments	Correct relation phrase, incorrect arguments	Incorrect argument order
ExtrHech (Spanish), 115	0.09	0.22	0.16	0.04
ReVerb (English), 93	0.12	0.26	0.13	–

Very similar distributions can be seen for the first 3 types of errors for both languages. Incorrect argument order was not observed for English because of highly dominant direct word order.

Causes of errors in assertion extractions from the parallel FactSpaCIC dataset are shown in Table 5.

Table 5. Causes of errors in assertion extraction from parallel FactSpaCIC, percent of all errors

	ExtrHech	ReVerb
N-ary relation	24%	41%
Underspecified noun phrase	10%	9%
Incorrect POS-tagging	10%	5%
Incorrect coordinative conjunction	43%	14%
Incorrect relative clause	19%	9%
Non-contiguous relation	5%	–
Over-specified relation phrase	5%	–
Inverse word order	14%	–
Infinitive	–	9%
Underspecified relation phrase	–	5%
Over-specified noun phrase	–	5%
No extraction	–	23%

One of the main issues for both languages is N-ary relations, i.e. relations requiring more than two arguments (e.g. "*The boy gave a book to the girl*"). Other issues frequent for both languages are incorrect relative clause resolution and incorrect coordinating conjunction resolution; however, both are more typical for ExtrHech system. Relative clauses can be more common and complicated for Spanish language because relative pronouns readily take prepositions (e.g. *en el cual* vs. less common *in which*), although this needs linguistic proof. Resolution of coordinating conjunction is implemented differently in each system. The English-language system would sometimes either leave out all but the first of coordinated elements or consider all coordinated elements as one argument. Consequently, they were either not counted as extracted assertions at all or considered as one correct extraction.

Interestingly, for neither of the systems incorrect POS-tagging is among top causes of errors, due to the high precision of the modern POS-taggers.

Several issues are encountered only for one language. Non-contiguous relation phrases, although present in English too, are more common in Spanish since they can be caused by free word order. Over-specification vs. under-specification of relational phrases is causes by differences in system implementation. Another observation is that ReVerb does not attempt detecting facts in 5 sentences from the dataset. In this experiment, ExtrHech showed more robust behavior.

6. Conclusions and Future Work

We have presented the Open IE system for Spanish language, ExtrHech, based on syntactic and lexical constraints. It takes a POS-tagged text as an input and outputs a list of extracted binary relations per sentence.

It ExtrHech performs at the precision and recall levels comparable with the state-of-the-art systems for English based on similar approach, i.e. syntactic and lexical constraints and POS-tagging. 87% precision and 70% recall were obtained for the dataset with grammatically correct sentences, and 55% precision and 49% recall were observed on the raw Web text dataset, which included incorrect or incoherent sentences. Although the recall of ExtrHech is lower than that of the syntactic parsing based systems, the precision is at the same level, and the speed is much higher.

We also performed the analysis of errors in extractions made by ReVerb and ExtrHech system from the parallel English-Spanish dataset of 68 grammatically correct sentences. It shows that the major error causes are common for both languages. Interestingly, incorrect POS-tagging is not among the major issues for extraction errors. There are sets of issues that are typical either for one language. Some of them are related to the language properties, others are caused by systems' implementation differences. However, ExtrHech was more robust on the dataset used in the experiment.

Future work includes detailed analysis on how POS-tagger accuracy affects POS-tag based Open IE. We also plan to conduct a comparative experiment for an English-Spanish parallel or comparable dataset containing incoherent or incorrect sentences to better understand the robustness in different languages. Additionally, we will

continue improving ExtrHech's handling of the inverse word order, relative clauses, and coordinating conjunctions.

Acknowledgements. The work was partially supported by European Union via 7th FWP (*Web Information Quality Evaluation Initiative*, WIQ-EI, project 269180), Government of Mexico via CONACYT (50206-H), IPN (SIP 20131702), and Mexico City Government (ICYT PICCO10-120).

References

1. *Aguilar Galicia H.* (2012), Extracción automática de información semántica basada en estructuras sintácticas, MSc thesis, IPN, Mexico.
2. *Banko M., Cafarella M. J., Soderland S., Broadhead M., and Etzioni O.* (2007), Open information extraction from the Web, *Proc. IJCAI 2007*, pp. 2670–2676.
3. *Banko, M., & Etzioni, O.* (2008). The Tradeoffs between Open and Traditional Relation Extraction. *Proc. ACL 2008*.
4. *Etzioni O., Banko M., Soderland S, and Weld D. S.* (2008), Open information extraction from the web, *Commun. ACM* 51(12):68–74.
5. *Etzioni, O.* (2011). Search needs a shake-up. *Nature*, 476(7358), pp. 25–26.
6. *Fader A., Soderland S., and Etzioni O.* (2011), Identifying relations for open information extraction, *Proc. EMNLP 2011*, pp. 1535–1545.
7. *Horn C., Zhila A., Gelbukh A., Kern, R., and Lex E.* (2013), Using Factual Density to Measure Informativeness of Web Documents, *Proc. NoDaLiDA 2013*, in print.
8. *Kim J., Moldovan, D.* (1993), Acquisition of semantic patterns for information extraction from corpora, *Proc. of 9th IEEE AIA*, pp. 171–176.
9. *Kirkpatrick M.* (2011), New 5 Billion Page Web Index with Page Rank Now Available for Free from Common Crawl Foundation, readwrite.com/2011/11/07/common_crawl_foundation_ announces_5_billion_page_w.
10. *Kozareva, Z., Hovy, E., & Rey, M.* (2010). Not All Seeds Are Equal: Measuring the Quality of Text Mining Seeds. *Proc. HLT 2010*, pp. 618–626.
11. *Landis J. R., Koch G. G.* (1977), The Measurement of Observer Agreement for Categorical Data, *Biometrics* 33(1):159–174.
12. *Mausam, Schmitz M., Bart R., Soderlund S., and Etzioni O.* (2012), Open Language Learning for Information Extraction, *Proc. EMNLP 2012*.
13. *Padró L., Collado M., Reese S., Lloberes M., and Castellón I.* (2010), FreeLing 2.1: Five Years of Open-Source Language Processing Tools, *Proc. LREC 2010*.
14. *Soderland S.* (1999), Learning Information Extraction Rules for Semi-Structured and Free Text, *Machine Learning* 34(1–3):233–272.
15. *Wu F., Weld D. S.* (2010), Open information extraction using Wikipedia, *Proc. ACL 2010*, pp. 118–127.

БЕЗЛИЧНЫЕ КОНСТРУКЦИИ С ПЕРЕХОДНЫМ ГЛАГОЛОМ В СЛАВЯНСКИХ И ГЕРМАНСКИХ ЯЗЫКАХ: НУЛЕВЫЕ ПОДЛЕЖАЩИЕ И СЕМАНТИЧЕСКИЕ РОЛИ

Циммерлинг А. В. (fagraey64@hotmail.com)

Институт современных лингвистических исследований МГГУ имени М. А. Шолохова

Ключевые слова: аргументная структура предиката, безличные конструкции, семантические роли, нулевые подлежащие, переходность, типология

TRANSITIVE IMPERSONALS IN SLAVIC AND GERMANIC: ZERO SUBJECTS AND THEMATIC RELATIONS[1]

Zimmerling A. V. (fagraey64@hotmail.com)

Institute for Modern Linguistic Research, SMSUH, Moscow, Russia

The paper argues that transitive impersonals in Russian, Ukrainian and Icelandic can be accounted for in terms of Mel'čuk's zero lexemes reanalyzed here as pronouns in the nominative case acting as agreement controllers. An alternative analysis resorting to Burzio's Generalization stipulates defective *v*P for different classes of verbs licensing transitive impersonals but fails to make correct predictions. The distribution of impersonals in Russian and Ukrainian does not depend on the distinction of unaccusative vs unergative vs psych predicates. Most Russian verbs labeled 'psych' in the previous generative research are either semantic causatives or agentive verbs with an external argument and valency grid <Agent, Patient>.

Keywords: argument structure, event structure, impersonal construction, thematic relations, zero subjects, transitivity, typology

[1] The paper is written with financial support of the Russian Foundation for the Humanities, project RFH 11-04-00282 'Typology of morphosyntactic parameters'. I am grateful to the anonymous reviewer for the valuable comments.

1. Introduction: Burzio's Generalization and transitive impersonals

Transitive impersonals of the type Russ. *ulicu*$_{\text{ACC.SG.F.}}$ *zasypalo*$_{\text{PRT.3SG.N.}}$ *peskom*$_{\text{INSTR.SG}}$ 'The street was shuttered with sand', Icel. *Bátinn*$_{\text{ACC.SG.M.DET}}$ *rak*$_{\text{PRT.3SG}}$ *að landi*$_{\text{DAT.SG}}$ 'The boat drifted ashore' which are attested in a number of languages with accusative alignment[2] challenge a controversial statement known as Burzio's generalization (BG). In its original form BG claims that only verbs that can assign (structural) accusative to some object, can assign an external theta-role (Agent) to its subject [Burzio 1986: 178]. BG makes two wrong predictions that: a) verbs without an Agent subject cannot assign accusative, b) any verb with an agent subject can assign accusative. Obvious counterexamples to both predictions exist, cf. Russ. *ego*$_{\text{ACC.SG.M}}$ *eto*$_{\text{NOM.SG}}$ *ochenj tjagotit*$_{\text{PRS.3SG}}$ where *tjagotit'* 'be a burden to someone' is a psych verb — according to standard assumptions, psych verbs lack an Agent argument and have a valency grid <Expericiencer, Stimulus>[3]. Further counterexamples to BG are impersonal passives from transitive verbs as Ukr. *bulo vidhyleno* in (1) since passive participles in the Minimalist program do not assign case[4].

(1) Ukr. ***Statt'u*** $_{\text{ACC.SG.F.}}$ <u>*bulo*</u> $_{\text{PRT.3SG.N.}}$ <u>*vidhyleno*</u> $_{\text{PRT.3SG.N.}}$. 'The paper has been declined'

A revised form of BG tries to predict Nominative case marking on the object. Nominative objects (internal arguments of verbs from different classes) are attested in Old Russian, Modern and Old Icelandic, North Russian dialects.

(2) O.Russ. *Ontan-e*$_{\text{NOM.SG.M}}$ <u>*prislal-e*</u>$_{\text{PRT.3SG.M}}$, *Ovdokim-u*$_{\text{DAT.SG}}$ [$_{\text{CoP}}$ *dva klesča*$_{\text{ACC.PL}}$ *da* ***sčuk-a***$_{\text{NOM.SG}}$].
'Ontane has sent two breams and **a pike** to Ovdokime'.

(3) Icel. Jón-i$_{\text{DAT.SG}}$ <u>líka</u>$_{\text{PRS.3SG}}$ / <u>líka-r</u> $_{\text{PRS.3PL}}$ [$_{\text{DP}}$ **þess-ar**$_{\text{NOM.PL.F}}$ **stúlk-ur**$_{\text{NOM.PL.F}}$].
'John likes these girls.'

Nominative case marking on the object is also attested in embedded clauses where the subject may preserve idiosyncratic marking with Dative case, cf. (4a). In (4b) the subject of the embedded IP gets Accusative case from the matrix verb *telja* 'to think'.[5]

---

[2] Mostly — in languages with a standard nominative-accusative sentence pattern, without ergative case-marking on the subject argument, cf. Hindi.

[3] In terms of 'theta-roles' and 'internal/external arguments' this statement means that psych verbs lack an external subject argument and have two internal arguments.

[4] This is a framework-internal issue since in a different framework one can stipulate that assignment of Case (at least in some languages) does not depend on the Voice value — Active vs passive.

[5] Both (4a) and (4b) can be analyzed as instances with Exceptional case marking (ECM), while the older term 'subject-to-object raising' fits only (4b), where the subject of the embedded clause, DP *Jón* gets accusative case as predicted by the valency grid of the matrix verb *telja*.

(4) a. Icel. *Sigg-a*$_{\text{NOM.SG}}$ *tal-d-i*$_{\text{PRT.3SG}}$ [$_{\text{IP}}$ ***Jón-i***$_{\text{DAT.SG}}$ *líka*$_{\text{INF}}$ [$_{\text{DP}}$ **þess-ar**$_{\text{NOM.PL.F}}$ **stúlk-ur**$_{\text{NOM.PL.F}}$].
'Sigga thought that John liked these girls.'
b. Icel. *Sigg-a*$_{\text{NOM.SG}}$ *tal-d-i*$_{\text{PRT.3SG}}$ [$_{\text{IP}}$ ***Jón***$_{\text{ACC.SG}}$ *líka*$_{\text{INF}}$ [$_{\text{DP}}$ **þess-ar**$_{\text{NOM.PL.F}}$ **stúlk-ur**$_{\text{NOM.PL.F}}$].
'The same'.

A revised form of BG stipulates that an object only gets Nominative case when there is no Nominative subject. It is falsifiable too, as shown by Woolford (2003) who gives up the idea of rigid conditions linking argument structure with case marking and explains the competition of structural Acc and Nom by preferences of less marked case forms[6]. Woolford's OT-driven description of Nom/Acc case marking of internal arguments in Icelandic and Faroese [Woolford 2003: 307–319] partly overlaps with Zimmerling's (2002: 755–775) analysis of the same data in terms of parametric typology.

1.1. Phrase-structural accounts of BG and defective *v*P

Although there is general consensus that BG is a simplistic observation that even in the framework of the Minimalist Program is considered an epiphenomenon, cf. Reuland (2000), there is a bulk of recent attempts to save BG in its original form. These attempts restate BG in phrase-structural terms and are based on Chomsky's (1995) idea of little *v* as a phrasal category responsible both for the projection of an external argument and structural accusative assignment. This gives a chance to account for cross-linguistic variation since one can add projections for different types of impersonal constructions *ad libitum*. A sketch incorporating earlier proposals is given in [Lavine 2012: 5] and reproduced below as (i); the upper shell of *v*P is identified as Voice Phrase while the lowest shell immediately above big V[7] is tagged 'v-TelicP/QuantP' and treated as a projection headed by some aspectual head [+Telic/Quant], cf. [Svenonius 2002].

(i) [$_{\text{voiceP}}$ External Argument [v-Voice v-Fate [$_{\text{v-CauseP}}$ v-Cause [$_{\text{v }\varphi\text{/AgrOP}}$ [$_{\text{v-TELIC/QUANTP}}$ v-TELIC/QUANT [$_{\text{VP}}$ Acc]]]]]].

Some authors claim that transitive impersonals are compatible with BG since little *v* containing a verb assigning Accusative and the theta-role of Theme (Patient) to its object also projects a silent argument which on some reasons is not spelled-out. Sigurðsson (2011) adds a projection called FATE for Icelandic verbs like *reka* 'drive' in sentences like (5) where they denote elemental processes. He claims that FATE is a special uncontrolled process feature blocking or turning off the usual Voice

---

[6] Woolford postulates a markedness scale 'nominative is a less marked case than accusative, accusative is a less marked case than dative' and derives her OT-constraints *accusative and *dative from it.

[7] The category hosting an object DP in the accusative case.

feature that otherwise introduces AGENT. In plain words, we are told that if an Icelandic sentence is about uncontrolled events the Agent argument is not projected since the *v*P in a 'Fate' context is defective but a transitive verb still assigns Accusative case.

(5) Icel. ***Bát-inn***$_{\text{ACC.SG.M.DET}}$ *rak*$_{\text{PRT.3SG}}$ *að landi*$_{\text{DAT.SG}}$
'The boat drifted ashore.'

It is unclear what is specific for Icelandic compared to Russian examples like (6) for which a zero subject argument has been postulated in [Mel'čuk 1995] and [Zimmerling 2009].

(6) Russ. ***Lodk-u***$_{\text{ACC.SG.F}}$ *prigna-l-o*$_{\text{PRT.3SG.N}}$ *obratno k beregu*$_{\text{DAT.SG}}$.
'The boat drifted back ashore'.

The absence of a Nominative subject in (5) and (6) due to a presumably defective *v* or the presence of a zero subject responsible for controlling φ-features of the verb forms *rak*$_{\text{PRT.3SG}}$ or *prignal-o*$_{\text{PRT.3SG.N}}$ are not observable things. Indeed, neither Russian nor Icelandic require that every sentence has a Nominative DP so the idea that transitive verbs like Icel. *reka* og Russ. *prignat'* are defective in 'fate' contexts has poor motivation except for the wish to save BG. The Zero-subject-theory has more motivation since it explains agreement features. If one after Mel'čuk (1995) assumes that φ-features of the verb form in 3Sg in (5) and (6) are controlled by a zero lexeme, then it is natural to assume that the subject in (5) and (6) is a zero pronoun $\varnothing^{3Sg}$ in the nominative case, 3Sg.N., cf. Zimmerling (2007). Along the same lines, the plural form of the Russian verb *prignat'* prompts that its controller is a zero pronoun $\varnothing^{3pl}$ in the Nominative case, 3Pl, cf. (7).

(7) Russ. *Lodk-u*$_{\text{ACC.SG.F}}$ *prignal-i*$_{\text{PRT.3PL}}$ *obratno k beregu*$_{\text{DAT.SG}}$.
'One drove the boat back ashore'.

In Modern Russian, zero subjects of the 3rd p. are distributed complimentary in situations denoting processes controlled by a human Agent ($\varnothing^{3Pl}$) and processes not involving any human Agent ($\varnothing^{3Sg}$). The participant triggering uncontrolled processes is called Elements in [Mel'čuk 1995] and Causer in [Lavine 2012]. I define it as non-animated Agent since all Russian and Icelandic transitive impersonals have a thematic argument (Patient).

(ii) Transitive impersonals in Russian, Icelandic and Ukrainian have event structure with an overt Patient argument expressed by an accusative DP and a covert argument with the value 'non-animated, non human Agent with a generic reference'.

2. Zero subjects and *φ*-features

Unless a syntactic theory stipulates that case is only assigned to spelled-out elements (Ns/NPs/DPs) or that zero subjects of finite clauses cannot not have role semantics, one must postulate nominative case to all zero subjects of transitive impersonals

since overt subjects of the same transitive verbs, cf. *reka* and *prignat'* both of which mean 'to drive' are invariably marked with Nominative. There is no hint that their lexical semantics is changed when they shift from overt subject to zero. Pereltsvaig (2000), Svenonius (2002) and Richardson (2007: 102–107) argue that transitive impersonals arise due to some modification of grammar e.g. only if some aspectual feature like +TELIC/QUANT is realized i.e. in the presence of some telic marker or in a quantity reading. Unfortunately, such a feature was not found, as Lavine (2012) points out, and the assignment of Accusative in Slavic impersonals is independent from their aspectual characteristics — Perfective vs Imperfective Aspect, ± telicizing prefix, ± quantity reading[8]. I conclude that 'Fate accusative' and 'Telic-Quant' shells of *v*P are fake notions postulated ad hoc for Germanic and Slavic impersonals respectively, in order to save an even more dubious requirement, BG. Modern Russian zero subjects $\varnothing^{3Pl}$ and $\varnothing^{3Sg}$ have following φ- and role-and-reference features.

(iii) Russ. $\varnothing^{3Sg}$: Zero pronoun, Nominative case, 3rd person singular, neuter, non−Human, non-animated generic Agent triggering a non-controlled process.

(iv) (iv) Russ. $\varnothing^{3Pl}$: Zero pronoun, Nominative case, 3rd person plural, generic animated[9] Agent triggering a controlled process.

Modern Icelandic does not have zero subjects associated with 3Pl while generic human subject is expressed by an overt indefinite pronoun *maðr* 'one' in Nom.Sg. The 3Sg form is linked both with generic non−Human Agents, cf. (5) above, and with generic human Agents. The latter is possible in two types of passives — impersonal passives from verbs taking dative and genitive objects[10], cf. *hvelfa* 'to turn down' in (8) and impersonal passives from transitive and ditransitive verbs, cf. *skamma* 'to scold' in (9). The construction in (8) is standard, while (9), so called 'new passive' occurs in sub-standard Icelandic only.

(8) Icel. *Bátu-num*$_{\text{DAT.PL.DET}}$ *var*$_{\text{PRT.3SG}}$ *hvolf-t*$_{\text{PARTII.SG.N}}$ ***viljandi***$_{\text{PARTI}}$.
'The boats have been turned down <by some people> **on purpose**'.

(9) Colloq. Icel. *Var*$_{\text{PRT.3SG}}$ *skamma-ð*$_{\text{PARTII.SG.N}}$ *þig*$_{\text{2ACC.SG}}$?
'Were you scolded?'

---

[8] Notably, transitive impersonals occur in Russian in the imperfective aspect, also in generic and habitual contexts, cf. *Pri takom vetre* ***ulicu***$_{\text{ACC.SG.F}}$ *zameta-e-t*$_{\text{PRS.3SG}}$ *snegom*$_{\text{INSTR.SG}}$ *za chas* 'With such a wind, the street gets covered with snow in a hour'.

[9] The requirement {+Human} Agent is too strong for $\varnothing^{3Pl}$ given the possibility of such sentences as Russ. $\varnothing^{3Pl}$ *pokusali*$_{\text{PRT.3PL}}$ ***men'a***$_{\text{1ACC.SG}}$ *sil'no* ' <they, i.e. some living beings, probably — insects> bit me terribly'.

[10] Icelandic has a large class of verbs taking dative and genitive objects. In most cases verbs from these classes do not license standard passives with an agreeing participle.

Given that (5), (8) and (9) exemplify one and the same type of zero subjects, the specification of Icelandic zero subjects is following:

(v) Icel. $\varnothing^{3Sg}$: Zero pronoun, nominative case, 3rd person singular, neuter[11] generic Agent.

The φ-feature '3Sg.N' may be too strong for Icelandic and other Germanic languages since the participle II form used in the perfect tenses and passives like (8) and (9) is morphologically neuter but can be interpreted as a non-agreeing form in syntax. Anyway, this is not a problem for our analysis: if we deny agreement features of the neuter form of participle II in (8) and (9), then this φ-feature of the zero subjects should be recast simply as '3Sg'. Ukrainian shares with Russian both types of zero subjects $\varnothing^{3Pl}$ and $\varnothing^{3Sg}$ distributed complementary in active sentences, and adds one more type — impersonal transitive passive. The pattern (9) with a passive voice and a generic human Agent remains marginal in Icelandic but is grammaticalized in Ukrainian, cf. (10b). Since Ukrainian also retains generic human Agents associated with 3Pl, this may lead to contextual synonymy of active and passive structures without an overt subject, cf. (10a)

(10) Ukr.

a. ***Tyremnyj***$_{\text{ACC.SG.M}}$ ***termin***$_{\text{ACC.SG.M}}$ *Berlusconi* $\varnothing^{3Pl}$ *skoroty-l-y*$_{\text{RT.3PL}}$ *do odnogo roku*.
'The prison sentence of Berlusconi was abridged up to one year.'

b. ***Tyremnyj***$_{\text{ACC.SG.M}}$ ***termin***$_{\text{ACC.SG.M}}$ *Berlusconi* (*bulo*$_{\text{PRT.3SG.N}}$) *skoroche-n-o*$_{\text{PARTII.SG.N}}$.
'The prison sentence of Berlusconi has been abridged'.

Mel'čuk's approach to transitive impersonals is similar to the phrase-structural account of Lavine & Freidin (2002) who stipulate for them a φ-complete *v* and a probe-head relation (abstract object agreement). Russian, Ukrainian and Icelandic show rich agreement morphology which prompts that the inflectional features of an impersonal verb are controlled by some syntactic category. A radical form of Lavine & Freidin's idea is that only languages with a φ-complete *v* can be accounted for in terms of zero subject categories serving as agreement triggers. It is probable, since no zero subjects have been found in languages with poor verbal agreement. Later, Lavine (2012) revised his account since it failed to predict the ungrammaticality formed by the basic, so called monadic unaccusatives as Russ. *zamerznut'* 'to be frozen (over)', *lopnut'* 'to burst', *vylinjat'* 'to molt' which do not assign accusative while so called dyadic unaccusatives asserting 'a causative sub-event', as Russ. *zamorozit'* 'to freeze smth', *zamesti snegom* 'to cover smth with snow' still can.

[11] As in Modern Russian and Ukrainian, the neuter form is overtly marked in the perfect tenses which is historically due to the fact that Germanic and Slavic participle II has nominal morphology. The Slavic verbal ending Nom/Acc.Sg.N. *-o* as an impersonal marker (cf. Russ. *svetal-o, ego stošnil-o, lodku prignal-o k beregu*) is a late borrowing of a nominal ending into verbal morphology.

(11) Russ.
a. **Rek-u*$_{\text{ACC.SG.F}}$ *zamerz-l-o*$_{\text{PRT.3SG.N}}$
b. *Rek-a*$_{\text{NOM.SG.F}}$ *zamerz-l-a*$_{\text{PRT.3SG.F}}$
'the river froze up'.

(12) Russ.
a. **Utk-u*$_{\text{ACC.SG.F}}$ *polin'a-l-o*$_{\text{PRT.3SG.N}}$
b. *Utk-a*$_{\text{NOM.SG.F}}$ *polin'a-l-a*$_{\text{PRT.3SG.N}}$
'the duck molted'.

(13) Russ. *Vesj*$_{\text{ACC.SG.M}}$ *gorod*$_{\text{ACC.SG.M}}$ *zamorozi-l-o*$_{\text{PRT.3SG.N}}$.
'The whole city was frozen over'.

(14) Russ. *Stolbiki*$_{\text{ACC.Pl.M}}$, *zame-l-o*$_{\text{PRT.3SG.N}}$ *snegom*$_{\text{INSTR.Sg}}$.
'The stakes got covered by snow'.

3. Semantic roles and verb classes

The distribution of (11)–(14) is easily explained without recourse to syntax since *zamerznut'* 'to be/get frozen' or 'to be/get frozen up/over', *polin'at'* 'to molt', 'to shed hair' are Statives[12] but not Activities or Actions. Statives do not occur in transitive impersonals in Russian, since they do not project an Agent event role, as required by (iii) above, while verbs denoting processes and projecting an Agent event role can occur in transitive impersonals, although this is not a sufficient condition[13]. It is bizarre that the class of unaccusatives hosts both Statives like *zamerznut'* 'to be/get frozen' and transitives/causatives like *zamesti* 'to cover smth with snow' and *zamorozit'* 'to freeze smth', since *zamorozit' 'X freezes Y'* is just a semantic causative to *zamerznut'* 'X makes Y *zamerznut' '*. The origin of unaccusative theory, cf. Perlmutter (1978) was that intransitives split into verbs with an Agent-like argument (unergatives) and verbs with a Patient-like argument (unaccusatives). Initially, 'unaccusative' was a cover term for inactive intransitives, their only argument being Patient-like but lacking the canonic marking of Patient with the Accusative case, hence the ill-formedness of (11a) and (12a). The next claim was that unaccusatives make up a semantic class in UG, their sole argument being straightforwardly identified as Patient. A further stipulation was that unaccusatives get uniform syntactic diagnostics across languages, such as distribution of BE- and HAVE- auxiliaries in perfect tenses in Dutch or Danish, possibility of transitive impersonals or distributive *po-* constructions in Russian etc. Both claims are controversial, cf. Plungian (2011: 117–121). Even if uniform diagnostics of verb classes exists, it does not prove that there is a general meaning behind them. The final claim was that the notion of grammatical subject has different value for transitives,

[12] I.e. verbs denoting static situations or transitions from one state to another.

[13] The sufficient conditions are that a) a Russian verb does not select for +Animate subjects only, b) the state of affairs can be described as resulting from a non-controlled activity.

unergatives and unaccusatives: since these allegedly are semantic classes, their subjects originate in different positions in UG, as prescribed by a universal hierarchy of thematic roles and show different movement patterns (subject raising)[14]. If we first stipulate that Icel. *reka*, Russ. *prignat'*, *zamorozit'*, Ukr. *skorotyty* are unaccusatives we do not need to project a subject position for them since we already know that such verbs produce a defective *v*.

In Russian and Ukrainian transitive impersonals are not licensed by a single semantic feature. The necessary condition is that a verb is not a Stative and can take an Agent subject. The sufficient conditions for Russian are that A) the verb does not select for +Animate subjects only, B) the resulting event can be interpreted as an outcome of some non-controlled activity. The condition A) is illustrated by the transitive *proexat' ostanovku*$_{\text{Acc}}$ 'to miss one's stop' that has an Agent subject but invariably selects {+Animate; +Referential} subjects. Such verbs do not license transitive impersonals with $\varnothing^{\text{3Sg}}$. Suppose that a train has been set in motion due to mechanical failure[15] and drives past a stop. Even then, it is still impossible to use (15) in standard Russian.

(15) Russ. **etu*$_{\text{ACC.Sg.F}}$ *ostanovku*$_{\text{ACC.SG.F}}$ *proexa-l-o*$_{\text{PRT.3SG.N}}$.
Intended: 'The vehicle missed a stop as a result of an uncontrolled motion.'

The condition B) is illustrated by the pair of transitive verbs *kol'nut'* 'to sting' and *ukusit'* 'to bite', 'to sting'. Both can denote a situation like 'A mosquito stang/bit a man'. But *ukusit'* invariably selects {+ Animate; + Referential} subjects, while *kol'nut'* does not show this condition: accordingly, **ego ukusilo* would mean that X has been bit by a non-referential subject, while *ego kol'nulo* entails that X felt consequences of a sting or was able to detect it. Therefore, (16a) is grammatical, while (16b) is not.

(16) Russ.
a. ***Ego***$_{\text{3SG.ACC.M}}$ *kol'nu-l-o*$_{\text{PRT.3SG.N}}$ *v ščeku*$_{\text{ACC.PREP}}$ *Komar*$_{\text{NOM.SG.M}}$?[16]
'He felt a sting in the cheek. A mosquito?'
b. ****Ego***$_{\text{3SG.ACC.M}}$ *ukusi-l-o*$_{\text{PRT.3SG.N}}$ *v ščeku*$_{\text{ACC.PREP}}$ *Komar*$_{\text{NOM.SG.M}}$?

3.1. Causatives and psych verbs

Lavine (2012: 10) argues that Russian and Ukrainian psych verbs do not license transitive impersonals. Transitive psych verbs have a grid <Experiencer, Stimulus>. The absence of an Agent argument could account for the ungrammaticality of (17a)

[14] Cf. claims that postverbal subjects in SVO/SOV languages and narrative …SV → VS or locative inversion are primarily or exclusively characteristic of unaccusative subjects [Babyonyshev 1996: 137–144].

[15] Events of this type have been attested, cf. http://lenta.ru/articles/2013/01/30/train/

[16] From the viewpoint of Russian grammar *komar* 'mosquito' behaves as an {+Animate} subject. It takes the standard endings of the animated declension.

where Causer/Stimulus is expressed by a DP *igruškoj* 'by a toy' in the instrumental case but a similar example (17b) where Causer/Stimulus argument is expressed by a DP *vspyškami molnii* 'by flashes of lightning' in the same instrumental case is acceptable. The ill-formedness of (17a) is due to the lexical filling, not to general semantic characteristics of the verbs like Russ. *napugat'* 'frighten smb'.

(17) Russ.
a. **Mal'čika*$_{\text{ACC.SG.M}}$ *napuga-l-o*$_{\text{PRT.3SG.N}}$ *igruškoj*$_{\text{INSTR.SG.F}}$.
Intended: 'The boy was frightened by a toy'.
b. *?Mal'čika*$_{\text{ACC.SG.M}}$ *napuga-l-o*$_{\text{PRT.3SG.N}}$ *vspyškami* $_{\text{INSTR.PL}}$ *molnii*$_{\text{GEN.SG.F}}$
'The boy was frightened by flashes of lightning'.

Russ. *napugat'* and its Ukrainian counterpart *nalyakaty* both have active uses with a Agent subject, cf. (18a) and semi-active uses with a Stimulus subject, cf. (18b–c). (18a) denotes a controlled process the result of which is triggered by the subject's intentional activity. (18b) denotes a process controlled by the subject but its effect on another participant is not directly related to the subject's intentional activity. (18c) denotes an uncontrolled process triggered by a non-Human Causer. The label 'psych verb' is applicable to (18c) and, with some reservations, to (18b), but not to (18a) where *napugat'* behaves as standard causative verb linked to an intransitive middle verb *napugat'sa* 'to be frightened'[17], cf. (18d). The middle verb *napugat's'a* has a reflexive marker *-s'a/s'*: its subject is marked with Nominative too but does not get the roles of either Agent or Patient.

(18) Russ.
a. *Direktor*$_{\text{NOM.SG.M}}$ ***umyšlenno*** *napuga-l*$_{\text{PRT.3SG.M.}}$ *sekretaršu*$_{\text{ACC.SG.F}}$.
'The director intentionally frightened the lady receptionist.'
b. ***Prixod***$_{\text{NOM.SG.M}}$ ***direktora***$_{\text{GEN.SG}}$ *sil'no* *napuga-l*$_{\text{PRT.3SG.M}}$ *sekretaršu*$_{\text{ACC.SG.F}}$
'The arrival of the director frightened the lady receptionist terribly
(not necessarily because the director wished to).'
c. ***Vspyški***$_{\text{NOM.PL}}$ ***molnii***$_{\text{GEN.SG.F}}$ *napuga-l-i*$_{\text{PRT.3PL}}$ *mal'čika*$_{\text{ACC.SG.M}}$
'The flashes of lightning frightened the boy.'
d. ***Sekretaršа***$_{\text{NOM.SG.F}}$ *sil'no* *napuga-l-a-s'*$_{\text{PRT.REFL.3SG.F}}$ *iz-za prixoda*$_{\text{GEN.SG}}$ direktora$_{\text{GEN.SG}}$
'The lady receptionist was terribly frightened because of the director's arrival.'

In (18a–c) the active argument is marked with Nominative and the other participant is marked with Accusative. This allows to describe *napugat'* as a standard

[17] Lavine (2012: 7) argues that dyadic unaccusatives specify a causative sub-event. I would restate this by claiming that dyadic verbs conforming to a formula 'X causes Y do $V_{intrans}$', like *zamorozit'* = 'X makes Y *zamërznut*', *napugat'* ='X makes Y *napugat'sa*' are not unaccusatives but causatives with an Agent argument in the subject position. Morphological causatives from intransitives are typical for Russian (and Ukrainian and Icelandic as well). Morphological causatives from transitive verbs are rare in Russian, cf. *poit'* 'to give smb to drink smth', 'to water cattle' and *pit'* 'to drink'. A similar pair is attested in Icelandic, cf. *drekkja* which is a causative to a transitive verb *drekka* 'to drink'.

semantic causative conforming to a formula 'X causes Y to make Z' and including a component 'to be frightened' normally expressed in Russian by a middle (stative) verb *napugat'sya* having a reflexive marker.

(vi) Russ. '*X napugal Y-a*' = 'X caused Y to make Z', 'Z = *napugat'sa*'.

Given that Russ. *napugat'* both denotes processes controlled by a referential human Agent, cf. (18a) and uncontrolled processes not involving human Agents, cf. (18c), it is puzzling that it blocks a transitive impersonal in (17a). I offer a multifactor explanation: a) an impersonal form of a causative verb is blocked or hampered in Russian, if there is a middle form i.e. a verbal form with a reflexive marker and an inactive meaning, cf. *napugat'sa* derived from the same stem[18] b) the sub-event associated with the second overt participant expressed by an Instrumental DP can be interpreted as part of a major event caused by a non-Human Agent triggering a non-controlled process, taking effect over the first overt participant expressed by an Accusative DP. The contrast of (17a) and (17b) can be explained in this way:

(17a) a. **Mal'čika*$_{\text{ACC.SG.M}}$ <u>*napuga-l-o*</u>$_{\text{PRT.3SG.N}}$ ***igruškoj***$_{\text{INSTR.SG.F}}$.

(17a') 'The sub-event associated with the second participant expressed by an Instrumental DP *igruškoj* cannot be interpreted as part of the effect of an uncontrolled process triggered by a covert argument and taking over the first participant expressed by an Accusative DP *mal'čika*'.

(17b) ?*Mal'čika*$_{\text{ACC.SG.M}}$ <u>*napuga-l-o*</u>$_{\text{PRT.3SG.N}}$ ***vspyškami***$_{\text{INSTR.PL}}$ ***molnii***$_{\text{GEN.SG.F}}$

(17b') 'The sub-event associated with the second participant expressed by an Instrumental DP *vspyškami molnii* can be interpreted as part of the effect of an uncontrolled process triggered by a covert argument and taking over the first participant expressed by an Accusative DP *mal'čika*'.

In short, an event like 'A toy frightened a boy' cannot be interpreted in Russian as contributing to an event 'A boy was frightened by an uncontrolled process', while an event 'Flashes of lightning frightened a boy' marginally can. This has nothing to do with either the conjecture that *napugat'* is an unaccusative or to the conjecture that it is a psych verb.

Our next claim is that the label 'psych verb' does not correspond to any semantic class. The background idea was that these verbs denote states of mind that typically lack a Nominative subject or, at least, an external argument with the role of Agent. If one turns to Russian verbs denoting uncontrolled reactions of a human subject, one can find some 10-20 transitive verbs selecting a {+Human} argument in the Accusative

---

[18] This condition is though neither necessary nor sufficient in Russian. Cf. *ego*$_{\text{3Sg.M}}$ <u>*udari-l-o*</u>$_{\text{Prt.3Sg.N}}$ (*tokom*$_{\text{Instr.Sg.}}$, *kuskom*$_{\text{Instr.Sg}}$ *armatury*) 'X has been hit by a discharge of current/by a rod fragment) and *On*$_{\text{3Nom.Sg.M}}$ <u>*udari-l-s'a*</u>$_{\text{Pret.Refl.3Sg.M}}$ 'X bumped (against something)'.

case and licensing transitive impersonals[19]. None of these denotes a specific mental state — they rather describe uncontrolled reactions, including pathogen or symptomatic states (typically, bouts of illness and remission). Cf. Russ. *Men'a*$_{1SG.ACC}$ *tošnit*$_{PRS.3SG}$, *znobit*$_{PRS.3SG}$, *lixoradit*$_{PRS.3SG}$, *mutit*$_{PRS.3SG}$, *rvet*$_{PRS.3Sg}$, *pučit*$_{PRS.3SG}$, *raspiraet*$_{PRS.3SG}$ *ot gazov/lyubopytstva* which are possible in an actual-durative context, and *men'a*$_{1SG.ACC}$ *razneslo*$_{PRT.3SG.N}$, *razvezlo*$_{PRT.3SG.N}$, *skryučilo*$_{PRT.3SG.N}$, *prixvatilo*$_{PRT.3SG.N}$, *otpustilo*$_{PRT.3SG.N}$, *proneslo*$_{PRT.3SG.N}$, *proskvozilo*$_{PRT.3SG.N}$, *probralo*$_{PRT.3SG.N}$, *razobralo*$_{PRT.3SG.N}$ which are mostly used in the past tense in a perfective context. Some of them, as *tošnit', znobit', lixoradit', pučit',* do not take overt nominative subjects in Russian and are true impersonal verbs but this fact, contrary to Babby (2002) does not prove that they do not project zero subject $\varnothing^{3Sg}$ specified as {−Human} non-referential Agent of an uncontrolled process. I claim that the Accusative argument of all Russian verbs selecting an overt {+Human} object is Patient (Theme), not Stimulus, and they select an overt or covert Agent argument in the Nominative case.

(vii) So called psych verbs licensing transitive impersonals in Russian are transitive agentive verbs, typically with a grid <{−Human Agent}, {+Human Patient}>.

The absence of an overt nominative subject by *tošnit'* in (18) is an idiosyncratic feature, while the ability of the verb *rvat'* in (19) to take an overt nominative subject is a default option. Both (18) and (19) signal the same meaning 'X felt sick and vomited (due to the impact of an outer uncontrolled process)'. The main difference is that *tošnit'* is a transitive agentive verb with a narrow meaning that can only denote a class of situations 'Y makes X feel sick' and invariably selects a {−Human, −Animate} Agent, while *rvat'* is a transitive agentive verb with a broad meaning 'to tear', 'to rend', 'to pull out' which can denote a broader class of situations, both with a {+Human} and {−Human} Agent.

(18) Russ. *Ego*$_{3SG.ACC.M}$ $\varnothing^{3Sg}$ *stošni-l-o*$_{PRT.3SG.N}$.
'He nauseated', 'he vomited'.

(19) Russ. *Ego*$_{3SG.ACC.M}$ $\varnothing^{3Sg}$ *vyrva-l-o*$_{PRT.3SG.N}$.
'He vomited', 'he threw up'.

It is essentially redundant to postulate additional types of zero subjects for Russian impersonals with a {+Human} argument in Dative and Accusative case as Mel'čuk's initial analysis seems to hint (1995, 188) or to treat transitive

[19] The exact number is unclear since it is difficult to separate uses subcategorizing for a {+Human} argument in the Accusative and uses subcategorizing for a {−Human} argument in the same case if both of them license transitive impersonals. Cf. Russ. *Mashinu*$_{Acc.Sg.F}$ {−Human} *pripodn'-a-l-o*$_{Prt.3Sg.N}$ *i pones-l-o*$_{Prt.3Sg.N}$ *vetrom*$_{Instr.Sg.M}$ 'The car got lifted and carried away by the wind' and Russ. *Ego*$_{Acc.Sg.F}${+Human} *pone-s-l-o*$_{Prt.3Sg.N}$ 'He started talking / expressing his emotions unrestrained', both of which seem to instantiate one and the same underlying meaning of the agentive verb *ponesti*, lit. 'to start to carry smth'. A similar picture is with Russ. *perekosit'* 'to warp' or 'to twist', *skosobočit'* 'to make smth get lop-sided'.

impersonals from the unaccusative and psych groups differently. The φ-features and role-and-references properties of the Russian zero subject pronoun $\varnothing^{3Sg}$ {−Human, −Animate Agent of an uncontrolled process} apparently do not depend on either the fact whether the Patient (Theme) argument is specified as {+Human} or {−Human} or on the fact whether a sentence is about non-controlled physiological reactions or about other non-controlled processes. The last point can be demonstrated on impersonal uses of the transitive verbs *pronesti* and *vyrvat'*. In (20a–b) the event structure is identical.

(20) Russ.
  a. ***Pacienta***$_{\text{ACC.SG}}$ $\varnothing^{3Sg}$ *vyrva-l-o*$_{\text{PRT.3SG.N}}$ *i* $\varnothing^{3Sg}$ *prones-l-o*$_{\text{PRT.3SG.N}}$.
  'The patient vomited and his bowels moved (due to the impact of an outer uncontrolled process).'
  b. *Uragannym*$_{\text{INSTR.SG.M}}$ *vetrom*$_{\text{INSTR.SG.M}}$ ***pacienta***$_{\text{ACC.SG.M}}$ $\varnothing^{3Sg}$ *vyrva-l-o*$_{\text{PRT.3SG.N}}$ *iz gamaka*$_{\text{PREP.GEN}}$ i $\varnothing^{3Sg}$ *proneslo*$_{\text{PRT.3SG.N}}$ *des'at' metrov po vozduxu*.
  'The patient has been pulled out from a hammock by a hurricane (due to the impact of an outer uncontrolled process) and got carried away ten meters through the air.'

The proposed generalized account arguably extends to Russian ditransitive impersonals i.e. constructions with <$\varnothing^{3Sg}$, an overt Patient argument in the Accusative case, specified as {−Human} and an overt Experiencer/External Possessor argument in the Dative case specified as {+Human}>, cf. (21) and (22).

(21) Russ. ***Emu***$_{\text{3SG.Dat.M}}$ {+Human} ***nogu***$_{\text{ACC.SG.F}}$ {−Human} $\varnothing^{3Sg}$ *sve-l-o*$_{\text{PRT.3SG.N}}$.
'He got a cramp in his leg'.

(22) Russ. ***Emu***$_{\text{3SG.DAT.M}}$ {+Human} ***pam'at'***$_{\text{ACC.SG.F}}$ {−Human} $\varnothing^{3Sg}$ *otšib-l-o*$_{\text{PRT.3SG.N}}$.
'He had a lapse of memory.'

A formal analysis of (21) and (22) depends on the treatment of the Dative argument as a subject-like element[20], or as an indirect object. For the reasons of space I assume that $\varnothing^{3Sg}$ can be recognized as subject of (21) and (22).

4. Zero subjects in Ukrainian transitive impersonals

The final section of the paper briefly discusses transitive impersonals in Ukrainian. Here an equivalent of (17b) with *nalyakaty* 'to frighten' and $\varnothing^{3Sg}$ {−Human} is ill-formed, cf. (23). A passive with *nalyakaty* 'to frighten' and $\varnothing^{3Sg}$ {+Human} is, as expected, ill-formed too since this peculiar combination of arguments and verb forms would mean that a frightening effect of the flashes of lightning results from some controlled process triggered by a {+Human} Agent.

[20] Cf. Zimmerling (2009, 2012) for the analysis of two other Dative structures in Russian.

(23) Ukr.*__Xlopčyka__$_{\text{ACC.SG.M}}$ nalyaka-l-o$_{\text{PRT.3SG.N}}$ ∅$^{\text{3Sg}}$ {−Human} *spoloxamy*$_{\text{INSTR.PL}}$ *blyskavky*$_{\text{GEN.SG.F}}$.
Intended: 'The boy was frightened by flashes of lightning.'

(24) Ukr. *__Xlopčyka__$_{\text{ACC.SG.M}}$ *bu-l-o*$_{\text{PRT.3SG.N}}$ nalyaka-n-o$_{\text{PARTII.SG.N}}$ ∅$^{\text{3Sg}}$ {+Human} *spoloxamy* $_{\text{INSTR.PL}}$ *blyskavky*$_{\text{GEN.SG.F}}$.

An overly similar verb *zalyakati* 'to bully', 'to frighten' licenses impersonal passive.

(25) Ukr. __ïx__$_{\text{3.ACC.PL}}$ ∅$^{\text{3Sg}}$ {+Human} zalyaka-n-o$_{\text{PARTII.SG.N}}$ i ∅$^{\text{3Sg}}$ {+Human} zmuše-n-o$_{\text{PARTII.SG.N}}$ *movčaty*$_{\text{INF}}$.
'They were bullied and forced to keep silent.'

The contrast of (24) vs (25) may be explained by a filter on middle verb formation, proposed above for Russian pairs Causative : Morphological decausative like *napugat'* : *napugat'sa*. Indeed, there is a decausative *nalyakatys'a,* but not **zalyakatys'a* (|| Russ. **zapugat'sa*). *Zalyakaty* 'to bully' only selects {+Human} Agentive subjects while *nalyakaty* also takes {Human} subjects in the active voice[21]. Consequently, an elimination of a referential {+Human} subject leads to a well-formed passive structure with ∅$^{\text{3Sg}}$ {+Human}. Amazingly, *zalyakaty* also licenses active transitive impersonal construction, cf. (26).

(26) Ukr. *Zgadajte*$_{\text{IMP.2PL}}$, *jak* __Varku__$_{\text{ACC.SG.F}}$ *peklom*$_{\text{INSTR.SG.N}}$ zalyaka-l-o$_{\text{PRT.3SG.N}}$.
'Remember, how Barbara was frightened by hell/by stories about hell.'

The well-formedness of (26) indicates that ∅$^{\text{3Sg}}$ in Ukrainian active sentences is not associated with the value {−Human}. The context of (26) is unclear — either the woman was frightened by Hell as an imagined reality — {−Human Agent} or by stories about Hell told by some people — {+Human Agent}. I prefer to analyze the meaning of (26) as vague, not two-way ambiguous. Ukrainian passive construction with ∅$^{\text{3Sg}}$ and Ukrainian active construction with ∅$^{\text{3Pl}}$ are both unambiguous. Their zero subjects have different φ-features but the same role semantics {+Human Agent}, so the two constructions compete, cf. (27) vs (28).

(27) Ukr. __Oficeriv__$_{\text{ACC.PL}}$ ∅$^{\text{3Sg}}$ zalyaka-n-o$_{\text{PARTII.3SG.N}}$ ∅$^{\text{3Sg}}$ zaturka-n-o$_{\text{PART.3SG.N}}$, ∅$^{\text{3Sg}}$ zaklьova-n-o$_{\text{PART.3SG.N}}$, *usi*$_{\text{NOM.PL}}$ robl'at' $_{\text{PRS.3PL}}$ *use i vodnočas ne* robl'at' $_{\text{PRS.3PL}}$ *ničogo*.
'The officers are bullied, scared and cowed, all of them do everything and at the same time do nothing'.

(28) Ukr. __Oficeriv__$_{\text{ACC.PL}}$ ∅$^{\text{3Pl}}$zalyaka-l-i$_{\text{PRT.3PL}}$ ∅$^{\text{3Pl}}$zaturka-l-i$_{\text{PRT.3PL}}$, ∅$^{\text{3Sg}}$ zaklьova-l-i$_{\text{PRT.3SG.N}}$.
'The officers are bullied, scared and cowed'.

---

[21] A sentence like **Dark forests bullied the boy* is impossible in English, while a sentence like *Dark forests frightened the boy* is OK. The same holds for Ukr. *zalyakaty* and *nalyakaty*, respectively.

The Ukrainian participle ending 3Sg.N.-*o* used in impersonal passives like (1), (10b) and (27) is morphologically different from the agreeing participle ending 3Sg.N.-*e*. This parameter has a typological dimension[22]: overt and covert controllers of Ukrainian subject agreement seem to have different properties. However, since Ukrainian impersonal passives are copular structures with a slot for an overt copula *bu-l-o* 3Sg.N. in the past tense, one can give a uniform description of Ukrainian, Russian and, probably, Icelandic passives with participle II and a zero subject.

5. Preliminary conclusions

1. Transitive impersonals in Russian, Ukrainian, Icelandic and typologically similar languages can be explained in terms of zero subject pronouns controlling φ-features of the verb and showing role-and-reference properties of non-referential Agents.
2. Burzio's generalization (BG) does not predict the distribution of transitive impersonals. Phrase-structural accounts of BG add problems rather than solve them by stipulating fake categories as 'Accusative-of-fate-P', *'Accusative-of-nausea-P' etc. Licensing of transitive impersonals or, in other terms, merging of zero subjects, is conditioned by grammar principles, not in the lexicon.
3. Unaccusative verbs are at best a syntactic group, not a semantic class. So called psych verbs are a loosely related group of verbs selecting a {+Human} argument. Many verbs analyzed as belonging to the 'psych' group in Russian actually are agentive verbs with an external argument and valency grid <Agent, Patient>.
4. BG, the unaccusative and psych hypotheses do not make accurate predictions and have little value for computational linguistics. The relevant parameters can be predicted by implementing tags for thematic roles (Agent, Patient, etc), subcategorization options {±Refererential}, {±Animate}, {±Human}, {±controlled process}, derivational verb types — Stative, Causative, Decausative etc.

[22] An exact parallel is known from Modern Swedish.

References

1. *Babyonyshev, M.* (1996). Structural connections in Syntax and Processing: Studies in Russian and Japanese. MIT.
2. *Babby, L.* (2002) Subjectlessness, External Subcategorization, and the Projection Principle. Journal of Slavic Linguistics. Vol. 10.
3. *Burzio, L.* (1986). Italian Syntax. Dordrecht: Reidel.
4. *Chomsky, N.* (1995). The Minimalist Program. Cambridge, MA: MIT Press.
5. *Lavine, J.* (2012). Varieties of v and the Structure of 'Anti-Burzio' Predicates. // Typology of morphosyntactic parameters, Moscow 14–16 November, 2012.
6. *Lavine, J. & R. Freidin* (2002). The subject of defective Tense in Slavic. // Journal of Slavic linguistics 18: 1, 101–130.
7. *Mel'čuk, I.* (1995). Syntactic, or Lexical Zero in Natural Language. // The Russian Language in the Meaning-Text Perspective. Wiener Slawistischer Almanach. Sonderband 39. Moskau; Wien, 1995.
8. *Pereltsvaig, A.* (2000). On accusative adverbials in Russian and Finnish. // Adverbs and adjunction, eds. A.Alexiadou and P.Svenonius. 155–176. Linguistics in Potsdam, 6.
9. *Plungian, V.* (2011). Vvedenie v grammatičeskuju semantiku. Grammatičeskie značenija I grammatičeskie sistemy yazykov mira. Moscow: RGGU, 2011.
10. *Reuland, E.,* ed. (2000). Argument and case: Explaining Burzio's Generalization. Amsterdam: John Benjamins.
11. *Richardson, K.* (2007). Case and Aspect in Slavic. Oxford: OUP.
12. *Sigurðsson, H.* (2011). On the new passive. Syntax 14: 148–178.
13. *Svenonius, P.* (2002). Case in uninterpretable aspect. Proceedings of Perspectives in Aspect, University of Utrecht.
14. *Woolford, E.* (2003). Burzio's Generalization, Markedness and Locality Constraints on Nominative Objects. Ellen Brandtner and Heike Zinsmeister (eds.). New Perspectives in Case Theory. 299–327.
15. *Zimmerling, A.* (2002). Tipologičeskij sintaksis skandinavskix yazykov. Moscow: Yazyki slavyanskoj kul'tury. 2002. 896 p.
16. *Zimmerling, A.* (2007). Zero Lexemes and Derived Sentence Patterns. Wiener Slawistischer Almanach. Sonderband 69. Wien.
17. *Zimmerling, A.* (2009) Dative Subjects and Semi-Expletive pronouns. G. Zybatow, U. Junghanns, D. Lenertová, P. Biskup (eds.). Studies in Formal Slavic Phonology, Syntax, Semantics and Information Structure. Frankfurt am Main; Berlin; Bern; Bruxelles; New York; Oxford; Wien, 2009.
18. *Zimmerling, A.* (2012). Nekanoničeskie podležasčii v russkom yazyke. Ot značenija k forme, ot formy k značeniju: Sbornik statej v čest' 80-letija Aleksandra Vladimiroviča Bondarko. Moscow: Yazyki slavyankix kul'tur. P. 568–590.

Abstracts

THE IMPACT OF SYNTACTIC STRUCTURE ON VERB-NOUN COLLOCATION EXTRACTION

Akinina Y. S. (jakinina@hse.ru), **Kuznetsov I. O.** (iokuznetsov@hse.ru), The Center for Semantic Technologies, Higher School of Economics, Moscow, Russia,
Toldova S. Y. (toldova@yandex.ru), Lomonosov Moscow State University, The Center for Semantic Technologies, Higher School of Economics, Moscow, Russia

Automatic verb-noun collocation extraction is an important natural language processing task. The results obtained in this area of research can be used in a variety of applications including language modeling, thesaurus building, semantic role labeling, and machine translation. Our paper describes an experiment aimed at comparing the verb-noun collocation lists extracted from a large corpus using a raw word order-based and a syntax-based approach. The hypothesis was that the latter method would result in less noisy and more exhaustive collocation sets. The experiment has shown that the collocation sets obtained using the two methods have a surprisingly low degree of correspondence. Moreover, the collocate lists extracted by means of the window-based method are often more complete than the ones obtained by means of the syntax-based algorithm, despite its ability to filter out adjacent collocates and reach the distant ones. In order to interpret these differences, we provide a qualitative analysis of some common mismatch cases.

ALEXANDR EVGENJEVICH KIBRIK: FROM STRUCTURALISM TO NEW IDEAS

Alpatov V. M. (v-alpatov@yandex.ru), The Institute of Linguistics RAS, Moscow, Russia

The activities of A.E. Kibrik reflected the process of movement from the structural linguistic to the functional one; this process is characteristic for the modern linguistics but Kibrik formulated the main principles of a new paradigm very precisely, especially in his article "Linguistic postulates" (1983–1992). He pointed to the narrowness of the structural linguistics studying only the structure of the language. He wrote that it is necessary to study linguistic phenomena with the mental activity of speaking persons, called to reveal linguistic processes as a matter of fact, emphasized the central role of semantics in the language. Now linguistics develops in the direction that was determined by A. E. Kibrik thirty years ago.

CONDITIONAL RANDOM FIELD MODELS FOR THE PROCESSING OF RUSSIAN

Antonova A. Ju. (a-antonova@list.ru), HSE, Moscow, Russia,
Soloviev A. N. (a.solovyev@i-teco.ru), CJSC I-Teco, Moscow, Russia

The paper aims to illustrate the applicability of conditional random field (CRF) models to Russian texts. Introduced in 2001, CRF method has been successfully exploited and proved its efficiency for a variety of NLP tasks. Its main advantage over HMM is the possibility to model the dependencies and interdependencies in sequential data. Yet this approach has not been widely used for Russian. Since CRF operates with language-independent features, its initial adaptation for Russian can be minimalistic. We show how CRF models produce state-of-the-art quality for several basic NLP tasks, including named entity recognition, part-of-speech tagging and object-oriented sentiment analysis. We exploited CRF-Suite tool to train and evaluate our models. We used a corpus of news texts for NER and POS-tagging tasks and a subcorpus from Russian Twitter for SA. The results of the evaluation were compared to other existing methods for each of the three tasks.

SEMANTICS OF EMOTION CAUSATIVES: THE STATUS OF THE CAUSATIVE COMPONENT

Apresjan V. Yu. (vapresyan@hse.ru), National Research University Higher School of Economics, Vinogradov Russian Language Institute, Moscow, Russia

The paper considers semantic structure of emotion causatives and their interaction with negation, namely, its narrow or wide scope. Emotion causatives are defined as a group of causatives with their specific semantic properties that distinguish them from other groups of causatives. One of those properties concerns their relation with corresponding decausatives, which, unlike causatives, do not license wide scope of negation. There are several factors that enable negation to have scope over the causative element in emotion causatives — their imperfective aspect, generic referential status of the causative NP phrase, agentivity and conativity of the causative. Non-agentive causatives never license the negation of the causative component. Agentive conative causatives license the negation of the causative component more frequently and easily than agentive non-conative causatives, prompting the assumption that in their semantic structures the causative component has different statuses (assertion in the former, presupposition in the latter). It also has different forms for conatives and non-conatives. Conativity vs. non-conativity of emotion causatives is related to the emotion type, with conative synthetic causatives being limited to basic emotions. The greatest degree of conativity and, hence, the assertive status of the causative component characterizes three emotion causatives — *zlit'* 'to make mad'*, veselit'* 'to cheer up'*, and pugat'* 'to frighten'.

CORRECTING COLLOCATION ERRORS IN LEARNERS' WRITING BASED ON PROBABILITY OF SYNTACTIC LINKS

Azimov A. E. (mitradir@gmail.com), Moscow State Lomonosov University, Moscow, Russia,
Bolshakova E. I. (eibolshakova@gmail.com), National Research University Higher School of Economics, Moscow State Lomonosov University, Moscow, Russia

The paper describes a novel method for automatic collocation error correction in NL texts written by language learners or translated from another NL with the aid of machine translators. We assume that the main cause of collocation errors is the strategy of word-by-word translation used by authors of the texts or by machine translators, so the errors essentially depend on the source language. While processing a sentence from the text, the method considers as potential correcting variants all its paraphrases that have the same syntactic structure and are built by replacing all words of the sentence by their substitutes. Substitutes are automatically generated using word translation equivalents taken from a translation dictionary. To detect an error in the sentence, we propose a relevance degree function computed from the probability of the word's syntactic links and applied to the sentence and its paraphrases. If the function value for the sentence is lower than for some of its paraphrases, our method signals an error, then it is corrected by an appropriate sentence paraphrase. The method was evaluated by correcting collocation errors in English texts written by Russian speakers. Stanford Parser and an English text collection were used to gather statistics and compute the probability of English word syntactic links. Within certain limitation, the experiments gave promising results: our method detected about 80% of collocation errors (with words of various POS) and 87% of proposed correcting paraphrases contained a proper correction.

SEMANTICS OF THREAT IN FORENSIC LINGUISTICS

Baranov A. N. (baranov_anatoly@hotmail.com), IRL RAS, Moscow, Russia

The paper considers the semantics and pragmatics of threat as a speech act. In lexical semantics, the concept of a threat is often explained as a unified (single) notion. It is shown that speech acts of threat in Russian are divided into two types: threat-penalty and threat-warning. The latter type of threat — threat-warning — has a specific variety — threat-compulsion. Threat-penalty is a kind of a threat situation in which something bad occurred and speaker informs the hearer (who is responsible for this) that he will be punished. Threat-warning presupposes that no bad thing has occurred yet and the speaker shows the hearer that he should not do this bad thing. The realization of threat-compulsion assumes that the speaker tries to force the hearer to do something under threat of penalty. Distinguishing the three kinds of threat is important for forensic linguistics. In cases of extremism, murder, bribe, exaction and others articles of law detection of body of the crime presupposes an analysis of criminal intention, which is reflected

apart from everything else in kinds of threat. Implicit ways of threatening are the most complicated to analyze in forensic linguistics. The analysis of implicit threat presupposes that all parts of semantic representation of this speech act (variables with terms and constants) should be identified in the text. The paper focuses on the case of implicit threat. The specific feature of the case analyzed consists in the implicit expression of penalty.

CORPUS AS LANGUAGE: FROM SCALABILITY TO REGISTER VARIATION

Belikov V. (vibelikov@gmail.com) , RSUH, Moscow, Russia,
Kopylov N. (Nikolay_Ko@abbyy.com), RSUH, ABBYY, Moscow, Russia,
Piperski A. (apiperski@gmail.com), RSUH, Moscow, Russia,
Selegey V. (Vladimir_S@abbyy.com), RSUH, ABBYY, Moscow, Russia,
Sharoff S. (s.sharoff@leeds.ac.uk), RSUH, Moscow, Russia; University of Leeds, UK

The main research question of any corpus investigation, either while experimenting with the Internet or working with the RNC or any other corpus, should be the question of the object of investigation: do we study a particular corpus, search engine or the language "overall"? Unfortunately, researchers usually accept as self-evident the assumption of "scalability" of the results obtained with a specific corpus study to the whole body of language. The article examines the criteria to justify the possibility to scale specific data and proposes an approach to assessing the limits of discovered facts, as adopted in the framework of an ongoing project to create the General Internet Corpus of Russian (GICR). One of the basic ideas of this project is that scaling the results is a very limited operation. For the majority of linguistic and lexicographical problems, corpus analysis should be carried out within a well-defined genre and sociolinguistic parameters.

COMPUTATIONAL TREATMENT OF SUPPORT VERB CONSTRUCTIONS IN ITALIAN AND IN RUSSIAN

Benigni V. (valentina.benigni@uniroma3.it), Università "Roma Tre", Roma, Italia,
Cotta Ramusino P. (paola.cottaramusino@unimi.it), Università degli Studi, Milano, Italia

We aim at comparing some corpora-based computational resources that enable us to analyse the collocational profiles of the SVCs in both languages. The resources include SketchEngine, which works for both languages, Lexit for Italian and NKRJA for Russian. The case study focuses on the Italian verb *mettere* followed by a prepositional phrase with the prepositions *in* and *a*, and the corresponding Russian verb *stavit'/postavit'* followed by a prepositional phrase with the prepositions *v* and *na*. We discuss the options offered by the tools at the syntax-semantic interface. A closer comparison of the three tools shows that they provide different data. We have explored some aspects of the semantic tagging of Lexit and NKRJA and propose an integration of the two tools. It seems that further development of semantic tagging could be helpful in creating Italian-Russian lexicographic resources.

RESEARCH OF LEXICAL APPROACH AND MACHINE CROWDSOURCING MORPHOLOGICAL ANNOTATION

Bocharov V. V. (bocharov@opencorpora.org), **Alexeeva S. V.** (alexeeva@opencorpora.org), **Granovsky D. V.** (granovsky@opencorpora.org), **Protopopova E. V.** (protoev@gmail.com), **Stepanova M. E.** (mariarusia@gmail.com), **Surikov A. V.** (ksurent@opencorpora.org), OpenCorpora.org

Manually annotated corpora are very important and very expensive resources: the annotation process requires a lot of time and skills. In OpenCorpora project we are trying to involve into annotation works native speakers with no special linguistic knowledge. In this paper we describe the way we organize our processes in order to maintain high quality of annotation and report on our preliminary results.

DESCRIPTION OF THE RUSSIAN EXTERNAL POSSESSOR CONSRUCTION IN A NATURAL LANGUAGE PROCESSING SYSTEM

Bogdanov A. V. (bidon@inbox.ru), **Leontyev A. P.** (taonick@yandex.ru) ABBYY, Moscow, Russia

The paper shows how Russian external possessor constructions are treated in the ABBYY Compreno® system. The specific tasks of our system require that sentences with external possessor constructions be considered as synonymous with those with internal possessors. Accordingly,

the semantic structure is generated in such a way that the possessor, whether external or not, and the possessum form a single constituent. This is not the case with the syntactic structure because there is much evidence that the external possessor is not syntactically dependent on its possessum. The semantic and syntactic structures of external possessor constructions are not isomorphic so we have to apply a syntax-semantic interface to derive one from the other. We show that two different kinds of interface must be used. For constructions with strong lexical restrictions we use a special normalization module while leaving the syntactic description relatively simple. In contrast, constructions with fewer lexical restrictions require a more sophisticated syntactic description where movements are postulated.

THOSE WHO SEEK, WILL THEY FIND? (SEARCH FUNCTION OF VERBAL HESITATIONS IN RUSSIAN SPONTANEOUS SPEECH)

Bogdanova-Beglarian N. V. (nvbogdanova_2005@mail.ru), Saint-Petersburg State University, Philological Faculty, Saint-Petersburg, Russia

The article is dedicated to verbal hesitations used in Russian spontaneous speech when a speaker is trying to find a better way of expressing his idea. The search process always mates hesitation, sometimes self-correction, and sometimes stays incomplete. Our conclusions are based on the reflections on the Russian Speech Corpus (balanced annotated textothec and the One Speech Day block).

A COMPUTER DICTIONARY OF RUSSIAN PARONYMS BASED ON A FORMAL CRITERION OF PARONYMY

Bolshakova E. I. (eibolshakova@gmail.com), Moscow State Lomonosov University, National Research University Higher School of Economics, Moscow, Russia,
Bolshakov I. A. (iabolshakov@gmail.com), Independent researcher, Moscow, Russia

We note that Western European lexicography has neither precise definition of paronymy nor dictionaries of paronyms. However, such dictionaries can help us correct malapropisms like *massive evacuation* or *sensitive shoes*. Although three comprehensive dictionaries of Russian paronyms have been published in the recent decades, it remains unclear what additional features of similarity of two words of the same root and the same POS are needed to consider the words paronymous. Based on the collected statistics of affix proximity of paronyms in the largest printed dictionary of Russian paronyms, we propose a formal criterion of paronymy. Two words of the same root and the same POS are considered formally paronymous if their affix differences (separately for suffices and prefixes) are limited to particular values. Affix difference equals the minimal number of editing operations on affixes (deletion, insertion or substitution) that transform an affix chain of one word into that of the other. Aiming to develop a computer dictionary of formal paronyms, we first compiled a computer dictionary of 23,000 Russian words divided into 2,400 same-root, same-POS groups. All words were split into morphs: prefixes, the root, suffixes, and the ending. Then affix distances between word pairs from the groups were automatically computed, and all formally paronymous pairs were selected. These pairs constitute the resulting computer dictionary of paronyms, which contains 21,800 word entries with their 190,000 paronyms, larger than all known dictionaries of paronyms.

MODELING PECULIAR CONDITIONS OF UNDERSTANDING UTTERANCES (THE CASE OF IRONY)

Borisova E. G. (efcomconf@list.ru), Moscow Teachers Training University, Moscow, Russia,
Pirogova Yu. K. (adv-pirogova@yandex.ru), National Research University Higher School of Economics, Moscow, Russia

The article deals with modeling the understanding of natural language texts in special cases that differ from the trivial 'normal' condition 'what is said is what is meant' (literal understanding). This includes hints, metaphors etc. The article is focused on irony, which seems to be a paradox: 'what is meant' is different from 'what is said'. By thorough analysis of examples of irony both in the literature and in common usage (including texts of media and the Internet) we classify the cases of irony. The sense components of utterances which are to be understood in the

opposite way have been identified. They are not only parts of the dictum but of the modal frame as well. The pragmatic analysis showed the intentions of the Speaker using irony, including the cases when the object of mockery is the Speaker himself, or the Hearer. Correlation of irony vs. mockery and irony vs. quotations is investigated. The results should be used by designing the model of natural text understanding.

DICTIONARY-BASED AMBIGUITY RESOLUTION IN RUSSIAN NAMED ENTITIES RECOGNITION. A CASE STUDY

Brykina M. M. (m.brykina@gmail.com), ZAO Eventos, Moscow, Russia; Lomonosov Moscow State University, Moscow, Russia, **Faynveyts A. V.** (fainalex@lyandex.com), Freie Universität Berlin, Berlin, Germany; ZAO Eventos, Moscow, Russia, **Toldova S. Yu.** (toldova@yandex.ru), Lomonosov Moscow State University, Moscow, Russia; The Center for Semantic Technologies, Higher School of Economics, Moscow, Russia

The Information Extraction task and the task of Named Entities recognition (NER) in unstructured texts in particular, are essential for modern Mass Media systems. The paper presents a case study of NER system for Russian. The system was built and tested on the Russian news texts. The method of ambiguity resolution under discussion is based on dictionaries and heuristic rules. The dictionary-oriented approach is motivated by the set of strict initial requirements. First, the target set of Named Entities should be extracted with very high precision; second, the system should be easily adapted to a new domain by non-specialists; and third, these updates should result in the same high precision. We focus on the architecture of the dictionaries and on the properties that the dictionaries should have for each class of Named Entities in order to resolve ambiguous situations. The five classes under consideration are Person, Location, Organization, Product and Named Event. The properties and structure of synonyms and context words, expressions and entities necessary for disambiguation are discussed.Key words: Named Entities Recognition, Named Entities ambiguity, Named Entities disambiguation, rule-based approach.

COMPUTATIONAL REFINING OF A RUSSIAN-LANGUAGE TAXONOMY USING WIKIPEDIA

Chernyak E. L. (echernyak@hse.ru), **Mirkin B. G.** (bmirkin@hse.ru), Department of Applied Mathematics and Informatics, National Research University Higher School of Economics, Moscow, Russia

A two-step approach to devising a hierarchical taxonomy of a domain is outlined. As the first step, a coarse “high-rank” taxonomy frame is built manually using the materials of the government and other representative sites. As the second step, the frame is refined topic-by-topic using the Russian Wikipedia category tree and articles filtered of “noise”. A topic-to-text similarity score, based on annotated suffix trees, is used throughout. The method consists of three main stages: 1) clearing Wikipedia data of noise, such as irrelevant articles and categories; 2) refining the taxonomy frame with the remaining relevant Wikipedia categories and articles; 3) extracting key words and phrases from Wikipedia articles. Also, a set of so-called descriptors is assigned to every leaf; these are phrases explaining aspects of the leaf topic. In contrast to many existing taxonomies, our resulting taxonomy is balanced so that all the branches are of similar depths and similar numbers of leaves. The method is illustrated by its application to a mathematics domain, “Probability theory and mathematical statistics”.

A CORPUS OF RUSSIAN AS L2: THE CASE OF DAGHESTAN

Daniel M. A. (misha.daniel@gmail.com), **Dobrushina N. R.** (nina.dobrushina@gmail.com) NRU HSE / MSU, Moscow, Russia

In Daghestan, the number of Russian speakers has been dramatically increasing over the last few decades. Russian has assumed the functional niche previously vacant in this extremely multilingual setting, becoming the first ever lingua franca of the region as a whole. Russian is acquired in a situation of strong interaction with local languages and shows contact properties on various linguistic levels: phonetics, morphology, syntax and lexicon. Its regional variant is also

visibly developing as a self-identification device. The aim of this paper to discuss some (socio) linguistic properties of this idiom, attribute them either to interference or to imperfect learning, and to argue for building a corpus of Daghestanian Russian.

EXPRESSIVE COMPONENT STRUCTURING OF THE RUSSIAN THESAURUS RUSSNET

Degteva A. V. (degteva.anna@gmail.com), **Azarova I. V.** (ivazarova@gmail.com), Saint-Petersburg State University, St. Petersburg, Russia

The paper deals with the structure of expressive attributive word meanings implemented in the wordnet-type thesaurus for Russian (RussNet). The adjectives involved express the appraisal of objects and situations denoted by nouns, the assessment depending on the intrinsic qualities of objects or rendering the subjective attitude of the speaker. The research was based on a 21 million word corpus of modern texts. The sentiment meaning in RussNet is structured according to three parameters: Polarity, Domain, and Objectivity. "Polarity", the intrinsic parameter of the class, describes a positive or negative sentiment value and its measure. "Domain" represents one of the three most commonly expressed standpoints: pragmatic, moral, and aesthetic, as well as actualization of lexical functions Ver/AntiVer, Pos/AntiPos, and Bon/AntiBon defining semantic interaction with hierarchical groups of nominal meanings (semantic trees and subtrees of the RussNet thesaurus). "Objectivity" describes the assessment source as either being personal or custom, usual or individual for the object described. The parameters listed above are organized into a rather intricate scheme but in practical work its structure can be simplified, Yet, detailed analysis can help structuring fuzzy sentiment expressions and detecting versatile evaluative content.

DEVELOPMENT OF LEXICAL BASIS FOR THE UNIVERSAL DICTIONARY OF UNL CONCEPTS

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

The paper describes the current state of development of the lexical basis of an open and free lexical-semantic resource — the Universal Dictionary of UNL Concepts (UNLDC). The resource serves as a lexicon of an artificial intermediary language UNL (Universal Networking Language). It links the elementary units of UNL — concepts with lexicons of natural languages and various external lexical and semantic resources, including Wordnet and SUMO ontology. The dictionary's main goal is to support automated semantic analysis, encoding the meaning of the text as UNL semantic graphs and subsequent generation of text in different natural languages.

GERMAN-RUSSIAN IDIOMS ONLINE: ON A NEW CORPUS-BASED DICTIONARY

Dobrovol'skij D. O. (dm-dbrv@yandex.ru, dobrovolskij@gmail.com), Russian Academy of Sciences, Russian Language Institute, Moscow, Russia

The paper focuses on the structure and principles for constructing a new German-Russian phraseological dictionary based on corpus data. Fragments of this dictionary are available on the website of the German Language Institute in Mannheim: "Deutsch-russische Idiome online" http://wvonline.ids-mannheim.de/idiome_russ/index.htm. Relevant information is also made available via the Europhras homepage at http://www.europhras.org. In section 1, I formulate certain general principles of modern bilingual phraseology.. Section 2 discusses the state of the art of German-Russian phraseography and explains the need for a new German-Russian phraseological dictionary. In Section 3, key features of the new corpus-based dictionary are considered. The basic difference between the present dictionary and traditional ones is that all examples of idiom usage are taken from text corpora DeReKo and DWDS, and in individual cases from the German-language Internet. Parallel texts from the Russian National Corpus (RNC) are also used. The use of authentic examples based on text corpora is a new approach in bilingual lexicography. Traditional dictionaries were based on a limited body of randomly selected examples, and the use of the idioms was often not even exemplified. The advantages of using

corpora consist not only in more detailed and well thought-out illustrations of the expressions being described, but also in additional possibilities that the corpus provides for compiling the idiom-list and structuring entries.

DIALOGUE MODELLING IN PSYCHOLINGUISTICS: LITERAL AND ANALOGICAL PERSPECTIVES AS BASES FOR ADULTS' AND CHILDREN'S REFERENCES

Fedorova O. V. (olga.fedorova@msu.ru), **Delikishkina E. A.** (skaista_diena@mail.ru), **Slabodkina T. A.** (goodword@yandex.ru), **Tsipenko A. A.** (eire.morrigan@gmail.com), Lomonosov Moscow State University, Moscow, Russia

Dialogue is a fundamental part of language use. In search of systematic evidence how the dialogue mechanisms work we turn to the referential communication task originally devised by R. Krauss and specified by H. Clark. In our experiment, two students or children were seated at tables separated by an opaque screen, in front of each were 12 cards of so-called Tangram figures. For the Director the cards were already arranged in a target sequence, and for the Matcher the same figures lay in a random sequence. The Director's job was to get the Matcher to rearrange his or her figures to match the target ordering. They carried out the task in four trials. All conversations (36 adults' and 8 children's dialogues) were transcribed, including changes of speaker, back-channel responses, hesitations, and false starts. We consider a prediction proposed by H. Clark that people prefer analogical perspective, which focuses on the resemblances of the figures to natural objects, to literal perspective, which focuses on the literal features of the objects, i.e. their geometric parts. Our results confirm the hypothesis; we also describe some peculiarities of the child dialogue strategies.

PARSE THICKET REPRESENTATIONS OF TEXT PARAGRAPHS

Galitsky B. (boris.galitsky@ebay.com), **Ivovsky D.** (dilv_ru@yahoo.com), **Kuznetsov S.** (skuznetsov@hse.ru), **Strok F.** (fdr.strok@gmail.com), eBay Inc.; National Research University Higher School of Economics, Moscow, Russia

We develop a graph representation and learning technique for parse structures for sentences and paragraphs of text. We introduce parse thicket as a set of syntactic parse trees augmented by a number of arcs for inter-sentence word-word relations such as coreference and taxonomies. These arcs are also derived from other sources, including Rhetoric Structure and Speech Act theory. We introduce respective indexing rules that identify inter-sentence relations and join phrases connected by these relations in the search index. We propose an algorithm for computing parse thickets from parse trees. We develop a framework for automatic building and generalizing of parse thickets. The proposed approach is used for evaluation in the product search where search queries include multiple sentences. We draw the comparison for search relevance improvement by pair-wise sentence generalization and thicket-level generalization.

PAPER SUCH A PAPER. ON A REDUPLICATION PATTERN IN MODERN RUSSIAN

Gilyarova K. A. (hilaris@gmail.com), Russian State University for the Humanities, Moscow, Russia

The paper presents a semantic and pragmatic analysis of noun reduplication in colloquial Russian and the Internet language. We consider the repetition of a noun within the same prosodic unit separated by a particle "takoj" ('such') as in "statja takaja statja" ('paper such a paper'). Drawing on a corpus of examples gathered from Internet texts we categorize the semantics of this reduplication pattern into six types: (1) prototype and connotation, (2) non-fitting a stereotype, (3) condescension and irony, (4) expression of emotions, (5) discourse topic and scene-setting topic (6) object nomination and ellipsis. Compared to the model "such X-X", the model "X such X" more often points to the negative attitude. We also consider the syntactic structure of the given reduplication pattern.

LINGUISTIC ANALYSIS OF SOCIAL MEDIA

Grefenstette G. (Gregory.Grefenstette@3ds.com), 3DS Exalead, Paris, France

One can look upon the Web as a large corpus that can teach us about language use, and also about the real world. In order to determine what is new or interesting we need to know what the norm for language use is. This involves creating a language model that corresponds to what is found on the web. Since the web is so big, it is impossible to download it all and count appearances of words and phrases, so one must use the technique of probing: generating things to be tested and submitting them to a search engine to find their frequency of occurrence. It has been shown using Google to gather statistics is perilous since Google does not provide exact counts but rather estimates the number of pages containing an expression. These counts can be very far from the reality of what is really in Google's index. Using another search engine, such as Exalead, is one solution, but then the problem of index coverage comes into play. Google has declared having seen 1 trillion unique URLs (in 2008) but estimates of the size of Google's index are about 50 billion pages, so some hidden choice has been made of what is in the index and what is not. This means that frequency based language models derived from search engines are only approximate. Nonetheless, it is possible to make rough, relative judgments of how often one linguistic phenomenon appears with respect to another, and using probing can provide some information of the relative frequency of these phenomena. Over a long period, it is possible to generate and test a great number of possibilities, some examples of the usefulness of this technique are finding what words commonly occur with other words, what colors are often associated with nouns, what are the most common translation of multiword expressions, what are the most likely transliteration of English terminology and names into Japanese, for example.

The Web is not a uniform corpus, far from it. There are many different language registers even within one language: there are professionally edited well written articles, there are more colloquial blog posts, there are hastily written error-filled comments, all which generate different language models. One recent exploitation of user-generated content on the web has been the mining of opinions concerning some subject, or company, or product. Affect analysis is now a thriving market and a true commercial success for natural language processing. Many other areas of text mining remain to be explored. For example, the particular language used to tag photos in social media sites (such as Panoramio or Flickr) and reveal many things about the user (especially in conjunction with GPS and time data). This language is different from that found in the general web, or on Wikipedia. We can use it to find out the interesting things to visit in a city, we can predict where a tourist can go, we can even guess whether a user is a woman or a man, from their tagging behavior. Mining this information can lead to additional applications that exploit this new knowledge.

GESTURAL PROFILES OF RUSSIAN PREFIXES

Grishina E. A. (rudi2007@yandex.ru), Institute of Russian Language RAS, Moscow, Russia

The study analyzes the main types of gestures, which accompany the Russian verbs with/without prefixes. The gestures are described from the topological point of view: any hand/head movement is placed along the Cartesian coordinates and the statistical correspondence between prefixes and topological characteristics of gestures is detected. The paper presents the gestural profiles (the set of gestural attributes) of 16 Russian prefixes. The study makes use of the data of the Multimodal Russian Corpus (MURCO).

CHITAT' NE CHITAL, NO…: ON A RUSSIAN CONSTRUCTION WITH REPEATED LEXICAL ELEMENTS

Iomdin L. L. (iomdin@gmail.com), Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

The paper discusses several types of Russian microsyntactic units — non-standard syntactic constructions and syntactic idioms with repeated verbal elements. The primary construction discussed is of the type *chitat' ne chital (no sdelal chto-to menee sil'noe)* ≈ 'one did not really read it (but one did something less strong)'. In this construction, two copies of the same verb in different inflectional forms (one in the infinitive and the other in finite form) are present, the latter preceded by the negative particle. Since lexical instantiation of the verbal positions is virtually

free, the only restriction being imposed on their lexical coincidence, the construction should be treated as lexically unbound and, hence, as a non-standard microsyntactic construction. There are two more constructions that appear to be lexically and syntactically close to the primary one: the so-called emphatic tautological infinitive construction of the type *s'jest'-to on s"est* ≈ 'he will definitely eat it' and a syntactic idiom with lexically bound repeated verbal elements of the type *Ja tebja znat' ne znaju* ≈ 'I don't know you and have no wish to do anything with you'. We focus on the semantics of these three units and ways to discriminate them in human and automatic natural language processing tasks.

MAG VEL MOT: LANGUAGE INNOVATIONS IN EVERYDAY LIFE TERMINOLOGY

Iomdin B. L. (iomdin@ruslang.ru), **Lopukhina A. A.** (nastya-merk@yandex.ru), V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences, **Panina M. F.** (mar-fed@yandex.ru), **Nosyrev G. V.** (grigorij-nosyrev@yandex.ru), Yandex, **Matissen-Rozhkova V. I.** (heinin@mail.ru), **Vill M. V.** (vill.margarita@yandex.ru), **Zajdel'man L. Ja.** (gde.vyhod@gmail.com), **Vinokurov F. G** (fedor-win@ya.ru), Russian State University for the Humanities, **Vybornova A. N.** (anna@179.ru), Higher School of Economics, Moscow, Russia

The paper continues research into words denoting everyday life objects in the Russian language. This research is conducted for developing a new encyclopedic thesaurus of Russian everyday life terminology. Working on this project brings up linguistic material which leads to discovering new trends and phenomena not covered by the existing dictionaries. We discuss derivation models which gain polularity: clipped forms (*komp* < *komp'juter* 'computer', *nout* < *noutbuk* 'notebook computer', *vel* < *velosiped* 'bicycle', *mot* < *motocikl* 'motorbike'), competing masculine and feminine contracted nouns derived from adjectival noun phrases (*mobil'nik* (m.) / *mobilka* (f.) < *mobil'nyj telefon* (m.) 'mobile phone', *zarjadnik* (m.) / *zarjadka* (f.) < *zarjadnoe ustrojstvo* (n.) 'AC charger'), hybrid compounds (*plat'e-sviter* 'sweater dress', *jubka-brjuki* 'skirt pants', *shapko-sharf* 'scarf hat', *vilkolozhka* 'spork, foon'). These words vary in spelling and syntactic behaviour. We describe a newly formed series of words denoted multifunctional objects: *mfushka* < *MFU* < *mnogofunkcional'noe ustrojstvo* 'MFD, multifunction device', *mul'titul* 'multitool', *centr* 'unit, set'. Explaining the need to compose frequency lists of word meanings rather than just words, we offer a technique for gathering such lists and provide a sample produced from our own data. We also analyze existing dictionaries and perform various experiments to study the changes in word meanings and their comparative importance for speakers. We believe that, apart from the practical usage for our lexicographic project, our results might prove interesting for research in the evolution of the Russian lexical system.

USING STATISTICAL METHODS FOR PROSODIC BOUNDARY DETECTION AND BREAK DURATION PREDICTION IN A RUSSIAN TTS SYSTEM

Khomitsevich O. G. (khomitsevich@speechpro.com), **Chistikov P. G.** (chistikov@speechpro.com), Speech Technology Center Ltd, St. Petersburg, Russia

The paper deals with statistical methods for predicting positions and durations of prosodic breaks in a Russian TTS system. We use CART and Random Forest classifiers to calculate probabilities for break placement and break durations, using grammatical feature tags, punctuation, word and syllable counts and other features to train the classifier. The classifiers are trained using a large high-quality speech database consisting of read speech. The experimental results for prosodic break prediction show an improvement compared to the rule-based algorithm currently integrated in the VitalVoice TTS system; the Random Forest classifier shows the best results, although the large size of the model makes it more difficult to use in a commercial TTS system. To make the system more flexible and deal with the remaining break placement errors, we propose combining probabilities and rules in a working TTS system, which is the direction of our future research. We observe good results in experiments with predicting pause durations. A statistical model of break duration prediction has been implemented in the TTS system in order to make synthesized speech more natural.

SEMANTIC ROLES AND CONSTRUCTION NET IN RUSSIAN FRAMEBANK

Kashkin E. V. (egorkashkin@rambler.ru), Lomonosov Moscow State University, Moscow, Russia,
Lyashevskaya O. N. (olesar@gmail.com), NRU Higher School of Economics, Moscow, Russia

The paper reports on a research project in progress which involves a dictionary of Russian lexical constructions and a corpus tagged with FrameNet-like annotation scheme. Russian FrameBank, originally conceived as an analogue of Berkeley FrameNet, takes into account some recent approaches adopted in Construction Grammar and Russian lexical semantics, as well as certain features of the Russian lexical system and grammar.

We focus on the semantic annotation of constructions in FrameBank. First, the article describes the inventory of semantic roles used in FrameBank which correlates with the semantic classification of verbs and other predicates. Semantic roles form a hierarchy: 88 roles are classified into six clusters (those of Agent, Patient, Experiencer, Instrument, Addressee, Circumstances), which are further subdivided into some smaller groups. The hierarchical organization makes the inventory of semantic roles more flexible for use in theoretical research and computational applications (such as automatic semantic role labeling). We also show that many examples are annotated in a more appropriate way by introducing syncretic semantic roles (e. g. Instrument-Place or Result-Manner).

Second, we touch upon an ongoing project on the systematization of semantic shifts in verbal lexemes (metaphor, metonymy, and rebranding, which is argued to be a special type of a semantic shift, see, for example, [Rakhilina et al. 2010a]) and the corresponding changes in argument structure constructions (including changes of a morpho-syntactic pattern, omission of a participant which belongs to a known class, etc.). The labels for the shifts are provided, along with examples of their realization. Lexical constructions are defined on constant (lexicalized) slots, mainly verbs and other predicates in a particular meaning. Frames are thus seen as the signifié side of constructional clusters formed by synonymous predicates, aspectual pairs, etc. Since it is not uncommon for polysemous lexemes that the formal façade of constructions is inherited from sense to sense, we claim that the frame nets cannot be routed without taking into account sense relations in polysemous predicates.

The final discussion deals with the relation between semantic classes of verbs, semantic roles, and lexical/semantic constraints on the classes of participants as provided by FrameBank data.

DISCOURSE TAXONOMY

Kibrik A. A. (aakibrik@gmail.com), Institute of Linguistics RAS; Lomonosov Moscow State University, Moscow, Russia

Among the central issues in the theory of discourse is discourse taxonomy, that is elucidation of the parameters classifying discourses into types. There are several such parameters, and they are often cofused. The main ones include mode, genre, and functional style. The distinction in mode concerns the medium: spoken or written. Genres are related to the typical communicative goals, acknowledged by discourse communities, and are characterized by standard schemata. Functional styles are identified in connection with the various domains of human existence. There are other discourse taxonomies as well, in particular, quite important is the distinction between types of presentation that characterize not whole discourses but their fragments, or passages. Each discourse taxonomy reflects upon grammatical, lexical and other local linguistic choices. Such choices are a resultant of all factors stemming in discourse taxonomies. Even though discourse taxonomies are in principle independent from each other, discourse types established on the basis of different parameters may have similar properties. For example, the written mode and the official functional style have similar reflexes in the linguistic structure.

MORE THAN ONE: RUSSIAN IDIOMS WITH *ODIN/EDIN* COMPONENT

Kiseleva K. L. (xenkis@mail.ru), **Voznesenskaja M. M.** (voznes-masha@yandex.com),
Kozerenko A. D. (akozerenko@mail.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper deals with a part of Russian phraseology:, the idioms containing the *odin/edin* ('one', 'single') lexical component, e.g. *vse kak odin, odin-edinstvennyj, vse do edinogo, odnoj levoj, ni odna zhivaja dusha, iz odnogo testa* etc. (English equivalents for: 'one and all', 'all alone', 'all down to the last one', 'with one hand tied behind one's back', 'not a one living soul', 'cut from the same cloth'). We observe that, first, the meaning of the idioms containing *odin/edin* depends on

the meaning of the word *odin* in this context (ex. in *smekh odin* and *v odin prisest* we have two different lexical meanings of *odin*). Second, we try to classify these idioms according to the inner form model that we see in each case. For example, *vse do odnogo* is based on the model labeled "exhaustion" while the similar idiom *vse kak odin* is based on another model, labeled "matching". Apart from suggesting several classes of idioms depending on their inner form model, we show that the presence of the component *odin* systematically brings two semantic effects to the meaning of the idioms: uniqueness, oneness, wholeness vs. insufficiency, poorness, lameness.

CLAUSES COMBINING WITH *TO CHTO* IN SPOKEN RUSSIAN

Korotaev N. A. (n_korotaev@hotmail.com), Russian State University for Humanities, Moscow, Russia

In spoken Russian discourse, complement clauses introduced by a combination of *to* (originally — a correlative pronoun in nominative or accusative case) and *chto* (complementizer) may exhibit specific features that are not possible in standard written speech. Based on the data from several spoken corpora, the present study claims that *to chto* is regularly used as a compound complementizer. In prosodic terms, *to chto* is often pronounced together with the subordinate clause, while *to*-pronoun usually adheres to the main predicate, a strong intonation boundary appearing between it and the *chto*-clause. In semantic terms, *to chto*-constructions may violate the condition of 'givenness' that presumably licenses the use of the correlative pronoun *to* in standard speech. In syntactic terms, *to chto* may be used with predicates that require a different case (genitive, instrumental) or a prepositional phrase. Also, coordination of *chto*-clauses and *to chto*-clauses are possible, and *to chto*-clauses appear in contexts with other correlative pronouns in the main clause (like *takoj*).

COMPENSATION OF COMMUNICATION STIMULI IN THE EMOTIONAL DIALOGUE

Kotov A. A. (kotov@harpia.ru), National Research Center "Kurchatov Institute", Moscow, Russia

An utterance is generated as an expression of an internal communication stimulus. As indicated in the theory of politeness, contradicting tendencies may interfere with the expression of an initial stimulus, in particular an initial face threatening act may be modified by the strategies of negative and positive politeness. Basing on the observations on a multimodal emotional corpus we argue that a certain number of expressive cues in a similar way compensate and modify an initial communication stimulus. (a) A speaker may compensate the changes in gaze direction through gestures, showing iconic gestures when looking aside, and closing gestures when looking at the addressee. We show that "looking aside" is usually combined with addressed gestures (demonstration, iconic gestures). (b) Smiles may also compensate the definitiveness of the main utterance. We show that smiles usually appear in the postposition to an utterance and reduce face threatening in the situations of failure or doubtful proposal — in these cases smiles do not express pleasure and are not connected to jokes.

HUMAN BODY AND ITS PARTS IN DIFFERENT LANGUAGES AND CULTURES (THE RESULTS OF THE SCIENTIFIC PROJECT)

Krejdlin G. E. (gekr@iitp.ru), **Pereverzeva S. I.** (P_Sveta@hotmail.com), Russian State University for the Humanities, Moscow, Russia

The paper presents the main results of a project aimed at constructing semiotic representations of human body and corporality in different natural languages (English, Arabic (the Egyptian dialect), Lithuanian, German and Hindi) and the corresponding body languages. The lexical system of a body language consists of gestures (in a broad sense of the word), i.e. gestures proper (manual gestures, gestures of legs, etc.), postures, meaningful glances, touches and some other semiotic classes of units. The primary directions of the project are (1) to describe somatic objects and their significant combinations; (2) to describe major classes of these objects, such as the human body itself, body parts, bones, biological liquids; (3) to examine the features of these objects and their values as well as those of their names; (4) to exhibit different kinds of gestures with somatic objects, among them those expressing human relationships. We also focus on some results in the field of applied nonverbal semiotics, i.e. (a) description of Russian symptomatic gestures performed by a patient in a conversation with a doctor. These gestures

may serve to characterize a patient's disease; (b) semantic analysis of Russian phraseological units with names of somatic objects; (c) exploration of meaning and functional characteristics of the so-called Bible somatisms — linguistic expressions in the Bible texts with names of somatic objects as well as of the gestures; (d) analysis of theatrical corporeal behavior.

ADVERBIAL EXPRESSIONS BASED ON VERBAL NOUNS

Kustova G. I. (galinak03@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper discusses a stage of abstract noun grammaticalization — namely, transformation into adverbial expressions, cf. *v ozhidanii* 'waiting', *pod okhranoj* 'under protection', *po priglasheniju* 'by invitation', *v blagodarnost'* 'in gratitude'. Two types of such adverbials are distinguished: 1) the agent of the adverbial is not expressed (*Passazhiry khodili po perronu* ***v ozhidanii*** *poezda* 'The passengers were strolling along the platform waiting for the train'); 2) the agent of the adverbial is necessarily expressed (*Prijekhal* ***po priglasheniju djadi*** 'came by invitation of his uncle'). In contrast to adverbials, nominalizations can express all arguments.

A TYPOLOGICALLY ORIENTED DATABASE OF QUALITATIVE FEATURES

Kyuseva M. V. (mkyuseva@gmail.com), **Reznikova T. I.** (tanja.reznikova@gmail.com), **Ryzhova D. A.** (daria.ryzhova@mail.ru), HSE, Moscow, Russian Federation

The article presents the Typological Database of Qualities, which aims at providing a new tool for research in lexical typology. The database contains information on the lexicalization of several semantic fields of adjectives in different languages (like 'sharp' — 'blunt', 'empty' — 'full', 'solid' — 'soft', 'thick' — 'thin', 'smooth' — 'rough', etc.). We discuss issues concerning database structure (in particular, the choice of information units that would make the meanings from different languages comparable to each other). Special attention is devoted to the representation of figurative meanings in the Database which allows to investigate the models of their derivation from the literal meanings. The developed database can be used for solving both theoretical and practical tasks. On the practical level, the Database may serve as a multilingual dictionary which accounts for fine-grained differences in meaning between individual words. On the theoretical side, the Database allows for various generalizations on cross-linguistic patterns of polysemy and semantic change.

PROPERTIES OF ZERO COPULA IN RUSSIAN IN COMPARISON WITH PROPERTIES OF NON-ZERO VERBS

Letuchiy A. B. (alexander.letuchiy@gmail.com), National Research University Higher School of Economics, Moscow, Russia

The article is focused on the properties of the zero copula used as a present tense form in Russian. The principal aim is to check whether the zero copula can be used in the same contexts as non-zero verbs or if it has particular features. I find out that there are contexts where the zero copula is allowed while non-zero verbs in the present tense are prohibited; conversely,, there are constructions which require a non-zero verb and prohibit the zero copulas. The former contexts include mainly biclausal constructions. The reason is that the zero copula lacks morphological tense and mood markers and does not apparently contradict any syntactic restrictions. The latter contexts, where the zero copula is prohibited belong to constructions with temporal meanings and constructions with predicatives. In the end I draw attention to the fact that constructions with the zero copula are not simply a reduced variant of some full structures, they have some particular rules of use which differ in some respects from those of non-zero verbs.

ON THE CAUSATIVE MEANING OF THE RUSSIAN CONJUNCTION *A TO*

Levontina I. B. (irina.levontina@mail.ru), Russian Language Institute (Vinogradov Institute), Russian Academy of Sciences, Moscow, Russia

The Russian conjunctions *a to* as well as *a ne to* '≈ or else' have repeatedly become objects of linguistic studies. First of all researchers were interested in semantic distinctions between these conjunctions and conditions of their interchangeability. Besides, much attention has been paid to the

structure of polysemy of these items, especially *a to*. Yet one of the interesting meanings of the conjunction *a to* seems not to have received an adequate description. It is the meaning which is usually described as causal: *Sxodi v bulochnuju, a to xleba net* 'Go to the baker's since we are out of bread';. *Pojdem domoj, a to zactra rano vstavat'* 'Let's go home because tomorrow we have to get up early tomorrow'; *Net li u tebja soli, a to u menja konchilas'* 'Do you have some salt, since mine is over?' Apparently, the idea of cause alone is absolutely insufficient. The paper addresses this causative meaning of *a to* contrasting it with other senses of the conjunction and other words of causation'.

REPORTED SPEECH IN SPOKEN DISCOURSE: INTONATION AS A MEANS OF INTEGRATION

Litvinenko A. O. (allal1978@gmail.com), M. V. Lomonosov Moscow State University, Moscow, Russia

I discuss typical intonation patterns in Russian reported speech constructions, based on the data from the Prosodically Annotated Corpus of Spoken Russian which consists of 4 experimental subcorpora of Russian spoken discourse (the current version of the corpus is available on the website http://spokencorpora.ru/). More than 400 occurrences of reported speech of different types (direct speech, indirect speech, semi-direct speech) have been analyzed. I have attempted to show that (i) intonation patterns in preceding framing clauses (falling tone in main phrasal accent, rising tone in main phrasal accent and absence of main phrasal accent) correspond to the type of the reported speech (direct, indirect and semi-direct, accordingly); (ii) however, this correspondence is more a tendency than a cause-and-effect relationship; (iii) there are some typical patterns in semi-direct speech that use 'mixed' intonation in order to keep the 'original' illocutionary meanings and to integrate the reported speech into the following context as much as possible: the ***list pattern*** and the ***head-tail-pattern***.

MULTILINGUAL COMPOUND SPLITTING COMBINING LANGUAGE DEPENDENT AND INDEPENDENT FEATURES

Loginova-Clouet E. A. (elizaveta.loginova@univ-nantes.fr),
Daille B. (beatrice.daille@univ-nantes.fr), Nantes University, Nantes, France

Compounding is a common phenomenon for many languages, especially those with rich morphology. Dealing with compounds is a challenge for NLP systems since compounds are not often included in the dictionaries and other lexical sources. We present a compound splitting method combining language independent features (similarity measure, corpus data) and language specific component transformation rules. Due to the usage of language independent features, the method can be applied to different languages. We report on our experiments in splitting of German and Russian compound words, giving positive results compared to matching of compound parts in a lexicon. To the best of our knowledge. elaborated compound splitting is a rare component of NLP systems for Russian, yet our experiments show that it could be beneficial to use a specialized vocabulary.

DATA VISUALIZATION FOR BUILDING THE CATALOGUE OF RUSSIAN LEXICAL CONSTRUCTIONS (BASED ON RNC)

Lyashevskaya O. N. (olesar@gmail.com), NRU Higher School of Economics, Moscow, Russia; Vinogradov Institute of Russian Language RAS, Moscow, Russia,
Mitrofanova O. A. (alkonost-om@yandex.ru), Saint-Petersburg State University, St. Petersburg Russia, **Panicheva P. V.** (ppolin86@gmail.com), EPAM Systems, Russia

Our research aims at automatic identification of constructions associated with particular lexical items and its subsequent use in building the catalogue of Russian lexical constructions. The study is based on the data extracted from the Russian National Corpus (RNC, http://ruscorpora.ru). The main accent is made on extensive use of morphological and lexico-semantic data drawn from the multi-level corpus annotation. Lexical constructions are regarded as the most frequent combinations of a target word and corpus tags which regularly occur within a certain left and/or right context and mark a given meaning of a target word. We focus on nominal constructions with target lexemes that refer to speech acts, emotions, and instruments. The toolkit that pro-

cesses corpus samples and learns up the constructions is described. We provide analysis for the structure and content of extracted constructions (e. g. r:ord der:num t:ord r:qual|*pervyj 'first'* + LJUBOV' 'love'; LJUBOV' 'love' + PR|*s 'from'* + ANUM m sg gen|*pervyj 'first'* + S f inan sg gen|*vzgljad 'sight' = love at first sight*). As regards their structure, constructions may be considered as n-grams (n is 2 to 5). The representation of constructions is bipartite as they may combine either morphological and lemma tags or lexical-semantic and lemma tags. We discuss the use of visualization module PATTERN.GRAPH that represents the inner structure of extracted constructions.

LEXICO-GRAMMATICAL FREQUENCY DICTIONARY: A PRELIMINARY DESIGN

Lyashevskaya O. N. (olesar@gmail.com), NRU Higher School of Economics, Moscow, Russia

A new electronic frequency dictionary shows the distribution of grammatical forms in the inflectional paradigm of Russian nouns, adjectives and verbs, i. e. the grammatical profile of individual lexemes and lexical groups. While the frequency hierarchy of grammatical categories (e. g. the frequency of part of speech classes or the average ratio of Nominative to Instrumental case forms) has long been the standard topic of research, the present project shifts the focus to the distribution of grammatical forms in particular lexical units. Of particular concern are words with certain biases in grammatical profile, e. g. verbs used mostly in Imperative, in past neutral or nouns used often in plural. The dictionary will be a source for many of the future research in the area of Russian grammar, paradigm structure, grammatical semantics, as well as variation of grammatical forms.

The resource is based on the data of the Russian National Corpus. The article addresses some general issues such as corpora use in compiling frequency resources and technology of corpus data processing. We suggest certain solutions related to the selection of data and the level of granularity of grammatical profile. Text creation time and language registers are discussed as parameters which may shape the grammatical profile fluctuations.

TOGETHER OR SEPARATELY? ON THE SEMANATIC CATEGORY OF TWONESS IN RUSSIAN

Mikaelian I. (irina-mikaelian@yandex.ru), The Pennsylvania State University, State College, PA, USA, **Zalizaniak Anna A.** (anna.zalizaniak@gmail.com), Institute of Linguistics, Russian Academy of Sciences; Institute of Informatics, Russian Academy of Sciences, Moscowm Russia

This paper attempts to refine our understanding of the grammatical and semantic features of the Russian collective numerals using data of corpora. The focus of our attention is the word *dvoe* considered in comparison with other quantity words comprising the meaning 'two' in their semantics, i. e. the numerals *dva* 'two' and *oba* 'both', as well as the noun *para* 'pair, couple'. The importance for the Russian language of the semantic category of "twoness" has been shown, and a new term *gemina tantum* has been introduced to designate the class of nouns that tend to be used in plural form and normally refer to two objects forming a pair or a couple, cf. *shoes, boots, eyes, parents, spouses.* Semantic analysis of the words *dvoe* and *oba* in the context of human nouns has shown that these words practically never interchange because, despite similar assertions, they carry different presuppositions and implications.

IS THE DEVIL IN THE DETAILS?... A MEASURE TO ASSESS THE MATCHING OF IDIOSTYLE ELEMENTS IN THE TEXTS OF ONE — OR IS IT TWO? — AUTHORS (AGEEV-SIRIN/NABOKOV-LEVI)

Mikheev M. Yu. (m-miheev@rambler.ru), NIVC MGU, Moscow, Russia

We analyze N. Struve's hypothesis that the author of the text of *The romance with cocaine* (published in 1936 under the pseudonym M. Ageev) was Vladimir Nabokov. We compare the idiostyle features of this text and all of Nabokov's texts, as well as what is available in the Russian National Corpus, published before the Ageev and Nabokov works and after them. The general conclusion is that Nabokov seems not to be involved in this text. This problem was stated by Nikita Struve, who rejected biographical arguments and required that "philological", literary or poetic arguments should be given. We consider all of these arguments.

A COREFERENTIALLY ANNOTATED CORPUS AND ANAPHORA RESOLUTION FOR CZECH

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz), **Mírovský J.** (mirovsky@ufal.mff.cuni.cz), **Novák M.** (mnovak@ufal.mff.cuni.cz), Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic

The paper presents an overview of a finished project focused on annotation of grammatical, pronominal and extended nominal coreference and bridging relations in the Prague Dependency Treebank (PDT 2.0). We give an overview of existing similar projects and their interests and compare them with our project. We describe the annotation scheme and the typology of coreferential and bridging relations and give the statistics of these types in the annotated corpus. Further we give the final results of the inter-annotator agreement with some explanations. We also briefly present the anaphora resolution experiments trained on the coreferentially annotated corpus and the future plans.

THE PROSPECTS OF APPLICATION OF SEMANTIC MARKUP TO THE NAMED ENTITY RECOGNITION PROBLEM

Nekhay I. V. (nekhayiv@gmail.com), Department of image recognition and text processing, DIHT MIPT, Moscow, Russia

The paper describes an attempt to construct a Named Entity classifier upon ABBYY Compreno Syntactic and Semantic Parser that was presented at the "Dialogue" conference in 2012. The classifier employs supervised learning technique, namely the Conditional Random Fields model, developed under heavy constraints on the available feature set: no external NE lists or non-local features are used. The system is evaluated on the NER field's "gold standard" evaluation corpus of CoNLL-2003 achieving F-scores of 91.61% on dev and 87.51% on test set. The classifier outperforms several other systems developed under the same constraints on features, but underperforms a single system that makes use of significantly richer local context. The gain of individual classifier features based on parser attributes is explored; it is demonstrated that Compreno's semantic hierarchy and surface (syntactic) slots provide classifier with the most valuable features used to locate and classify NEs. This reliance on parser results, however, leads to error propagation from parser to classifier, as shown in the section on error analysis. Final conclusions make an attempt to offer several topics for following research.

EGOCENTRICALS AND THEIR REGISTERS OF INTERPRETATION

Paducheva E. V. (elena.paducheva@yandex.ru), VINITI, Russian Academy of Sciences, Moscow, Russia

Linguistic entities (words, grammatical categories, syntactic constructions) are called EGOCENTRICALS, if their semantics presupposes the SPEAKER as one of the participants in the situation described, cf., for example, *sejčas*, as in *On sejčas doma* ['he's now at home', the speaker is the holder of the moment of speech], *edva li* 'unlikely', as in *On edva li pridet* ['he's unlikely to come', the speaker is the subject of doubt], subjunctive mood, as in *Byla by sejčas vesna!* ['if it were spring now!', the speaker is the subject of volition]. Only CANONICAL communicative situations can afford a sterling, i.e. full value, speaker — with the synchronous addressee, with the field of vision common to the speaker and the addressee, etc. In NON-CANONICAL communicative situations, such as NARRATIVE or HYPOTAXIS, when the speaker is not accessible as a performer of his/her presupposed role, and some substitute of the speaker comes into play, different egocentricals behave differently. Two types of egocentricals are discerned — SHIFTABLE (i.e. secondary) egocentricals, which can be used in all types of communicative situations, and HARD (i.e. primary) egocentricals, which stick to the canonical communicative situation, thus belonging to the so called MAIN CLAUSE PHENOMENA. One egocentrical is discussed in detail: the adverb *odnaždy* 'once upon a time'.

CONTEXT-INDEPENDENT AUTOCORRECTION OF QUERY SPELLING ERRORS

Panina M. F. (mar-fed@yandex-team.ru), **Baytin A. V.** (baytin@yandex-team.ru), **Galinskaya I. E.** (galinskaya@yandex-team.ru), Yandex, Moscow, Russian

While analyzing errors in the search queries, it is easy to notice that the most part of query spelling errors are trivial typos. Such errors usually do not depend on the surrounding words and their correction can be done in the automatic mode. In this work we tried to define a class of query spelling errors that can be corrected automatically. For the selected class we developed a classifier dividing corrections into reliable (suitable for automatic query spelling correction) and low-reliable (suitable only for the query spelling suggestion). As candidates for autocorrections we used query speller suggestions familiar to the users of search engines by "Did you mean..." function. For the classifier training we used typical lexical and statistical features. The experiments showed high performance of the word-level features and the ability to configure the classifier for a given level of accuracy. The application of the proposed method of trivial typo correction can significantly improve the quality of the query spelling errors correction.

BREEDS OF COOCCURRENCE: AN ATTEMPT AT CLASSIFICATION

Paperno D. A. (denis.paperno@unitn.it), Università degli studi di Trento, Trento, Italy, **Roytberg A. M.** (cvi@yandex.ru), **Khachko D. V.** (mordol@lpm.org.ru), IMPB RAS, Pushchino, Russia, **Roytberg M. A.** (mroytberg@lpm.org.ru), IMPB RAS, Pushchino, Russia and RSU HSE, Moscow, Russia

The paper proposes a substantial classification of collocates (pairs of words that tend to cooccur) along with heuristics that can help to attibute a word pair to a proper type automatically. The best studied type is frequent phrases, which includes idioms, lexicographic collocations, and syntactic selection. Pairs of this type are known to occur at a short distance and can be singled out by choosing a narrow window for collecting cooccurrence data. The next most salient type is topically related pairs. These can be identified by considering word frequencies in individual documents, as in the well-known distributional topic models. The third type is pairs that occur in repeated text fragments such as popular quotes of standard legal formulae. The characteristic feature of these is that the fragment contains several aligned words that are repeated in the same sequence. Such pairs are normally filtered out for most practical purposes, but filtering is usually applied only to exact repeats; we propose a method of capturing inexact repetition. Hypothetically one could also expect to find a forth type, collocate pairs linked by an intrinsic semantic relation or a long-distance syntactic relation; such a link would guarantee co-occurrence at a certain relatively restricted range of distances, a range narrower than in case of a purely topical connection, but not so narrow as in repeats. However we do not find many cases of this sort in the preliminary empirical study.

INCORPORATION IN VERB FORMS IN RUSSIAN

Pazelskaya A. G. (avis39@mail.ru), I-Teco JSC, Moscow, Russia

This paper investigates constraints on incorporation of nominal roots into compound verbs in Russian. This type of incorporation is generally impossible. The author examines several apparent exceptions from this generalization and proposes an explanation to the constraint itself as well as to the exceptions. A special attention is paid to the relation between (non-existing) compound verbs and compound nominals corresponding to the same nominal+verbal complex. Exceptions from the general constraint "no nominal roots within a compound verb" include deverbal adjectives which are formally equivalent to participles, verbs with reflexive and reciprocal "pronominal" components, verbs derived from compound nominals and compound verbs that have lost their semantic interpretability as complex verb. This interpretability is postulated to be the crucial feature correlating with the constraint on verbal compounds with nominal component, for the reason that this interpretability indicates the presence of two independent nodes (V and NP) in the structure of the compound. If such two node structure becomes a verb, the inner NP node receives case from higher structure levels and cannot incorporate into compound verb.

SEMANTIC FACTORS OF NOUN GOVERNMENT CHANGES IN MODERN RUSSIAN

Pestova A. R. (pestova2012@gmail.com), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The paper deals with the new variants of noun government in the modern Russian language. These variants are accounted for by certain semantic factors, such as development of meaning and semantic analogy. Due to the development of meaning the nouns *avarija* 'accident', *piruèt* 'pirouette' and *kontseptsija* 'conception' get new variants of government (*avarija s* + instrumental, *piruèty s* 'with'+ instrumental */ vokrug* 'around' + genitive, *kontseptsija po* 'on'+ dative). By semantic analogy the nouns *bum* 'boom', *fobija* 'phobia' and *vostrebovannost'* 'demand' adopt syntactic features of their synonyms. *Bum* 'boom' accepts a PP *na* 'on'+ accusative (by analogy with words *moda* 'fashion' or, *spros* 'demand'). *Fobija* 'phobia' governs either *pered* 'before'+ instrumental (by analogy with the noun *strakh* 'fear'), or *k* + dative (by analogy with words belonging to the semantic group 'attitude (positive or negative) toward smb, or smth', e.g. *neprijazn'* 'dislike', *uvazhenije* 'respect'). *Vostrebovannost'* 'demand' governs *v* 'in'+ prepositional case by analogy with the semantically similar word *potrebnost'* 'need'. Well-educated native speakers were asked to fill in questionnaires containing phrases with these variants. Their answers are presented.

VAGUE REFERENCE IN RUSSIAN: EVIDENCE FROM SPOKEN CORPORA

Podlesskaya Vera (podlesskaya@ocrus.ru), Russian State University for the Humanities, Moscow, Russia

The paper focuses on phenomena that fall under a broad category of what is called "loose uses" of language or "vague reference". These are lexical, grammatical and prosodic resources that allow the speaker to refer to objects and events for which the speaker fails to retrieve the exact name, or simply finds the exact name to be unnecessary or inappropriate. Based on first-hand corpus data of spoken Russian, the paper investigates expressions that are used in a language to temporarily substitute a delayed constituent, as well as those that do not imply any later substitution, but rather suggest an approximate nomination sufficient at the current moment of communication. These expressions can be used instead of their supposed exact correlate or together with it. The first option implies that an expression is used as a generic, or as a cover bleached nomination. The second option implies that the speaker doesn't take the full responsibility for the given actual nomination the expression is added to, since it is in some sense incomplete or not fully appropriate. The study of lexical resources of vague reference in spoken Russian is complemented by investigating also the associated syntactic and prosodic patterns.

A GRAMMAR DICTIONARY FOR AUTOMATIC ANALYSIS OF THE XVIII–XIXTH CENTURY TEXTS: FIRST RESULTS

Polyakov A. E. (pollex@mail.ru), Ushinsky State Scientific Pedagogical Library RAE, **Savchuk S. O.** (savsvetlana@mail.ru), **Sitchinava D. V.** (mitrius@gmail.com), Vinogradov Institute for the Russian Language RAS, Moscow, Russia

The paper presents the key principles of building a grammar dictionary and a morphological analyzer for XVIII–XIXth century Russian texts based on orthographical, morphological and lexical features exemplified by the Russian National Corpus (RNC). The analyzer should involve different modules applicable to different kinds of texts depending on their respective orthographical and grammatical phenomena. Several alternative ways of implementing orthographical and morphological rules are discussed (including pre-processing, online normalization etc.). Evaluation data of the first analysis results are presented.

UNSUPERVISED LEARNING OF PART-OF-SPEECH DISAMBIGUATION RULES

Protopopova E. V. (protoev@gmail.com), **Bocharov V. V.** (victor.bocharov@gmail.com), Saint Petersburg State University, Saint Petersburg, Russia

Morphological disambiguation is one of the key aims of part-of-speech tagging. The task is considered to be solved, though all the tools for disambiguation use a lot of manually created data. This paper describes an attempt to disambiguate Russian corpus without manually annotated data. The method used was proposed about twenty years ago but has not been applied to synthetic languages yet. The main idea of our approach is to derive disambiguation rules automatically from a corpus with ambiguous annotations using only a few statistical data. It can be done in a simple way by means of unsupervised learning. The results are quite high and can be compared to results of existing systems. We also tried to measure the size of the corpus necessary to produce a reasonable set of disambiguation rules and showed that it can be comparable in size with the corpora used to train statistical disambiguation models.

CONDUCTOR, PRESS THE BRAKES…

Rakhilina E. V. (rakhilina@gmail.com), National Research University Higher School of Economics, Institute for Russian Language / Russian Academy of Sciences, Moscow, Russia

The paper examines the lexical semantics and syntax of the form *postoj* (attenuated imperative of Russian verb *postojat'* 'stand (for a while)') describing it as one of the quasi-grammatical markers of continuous prohibitive, such as *prekrati, perestan', xvatit, budet, ostav', xoroš,* etc. All of them mark the illocution for interrupting the ongoing situation. *Postoj* differs from the other markers by its attenuative semantics, so that the situation (definite and taking place at the moment of speech, but not explicated in the sentence) has to be interrupted only for a while. The speaker offers to use this short span of time to improve it with some additional means; asyndetic clause, which follows *postoj*, explicates the speaker's suggestion.

USE OF CONTRAST AND EMPHASIS FOR CONVEYING IMPLICIT MEANINGS

Savinitch L. V. (savinitch@iitp.ru), A. A. Kharkevich Institute for Information Transmission Problems RAS, Moscow, Russia

The paper analyzes contrast and emphasis, modifiers of communicative meanings, their semantics and accentual structure in the sentences examined. We argue that contrastive and emphatic highlighting of one of the utterance components in the given examples are made by the speakers strategically, in order to convey occasional implicit meanings. All examples are illustrated with graphs displaying tone fluctuations, sound intensity, modulation of sound, and other prosodic features.

ON THE REGULAR AMBIGUITY: 'PARAMETER VS. HIGH VALUE OF PARAMETER'

Semenova S. Yu. (sonya_sem@mail.ru), Institute of Scientific Information on Social Sciences RAS, Russian State University for the Humanities, Moscow, Russia

The paper is concerned with regular ambiguity of the type 'parameter — high value' for Russian quantitative parametric nouns like *glubina* (depth), *davlenie* (pressure), etc. This type of ambiguity is shown to be heterogenous. For some dimensional nouns the ambiguity is caused by the metonymic shift from the meaning of a magnitude to the meaning of a spatial area where the value of this magnitude is high. For most parametric nouns this ambiguity is revealed in combinations with the verbs of surprise like *udivljat'sja* ('be surprised'). The ambiguity has some analogs in non-quantitative parametric nouns, e.g. 'parameter — Bon [the lexical function]' for the non-quantitative parameter *kachestvo* ('quality').

DISCURSIVE MARKERS OF NON-TRIVIAL LEXICAL CHOICE

Shilikhina K. M. (shilikhina@gmail.com), Voronezh State University, Voronezh, Russia

The paper discusses discursive functions of three Russian constructions: "*esli mozhno tak skazat*" [if I can say so], "*esli mozhno tak vyrazit'sya*" [if I can express it this way] and "*s pozvolenija skazat*" [if I'm allowed to call it X]. These constructions play the role of metalinguistic tools that structure the information flow. Functioning as parenthesis, these constructions mark the speaker's attitude towards his/her own speech actions and attract the attention of the addressee to the non-trivial form of expression. These non-trivial forms include unexpected lexical choice, metaphoric nomination and breaking the norms of word formation. By using "*esli mozhno tak skazat*" or "*esli mozhno tak vyrazit'sya*" the speaker can also introduce the process of searching for the most optimal way of expressing an idea. Non-trivial lexical choice or ungrammatical forms introduced by the constructions *esli mozhno tak skazat* 'if I may say so' or *esli mozhno tak vyrazit'sya* 'if I may express myself so'are signals of the speaker's stance towards the object or the situation. Another possible goal of unusual verbal behavior is switching from bona fide to non-bona fide mode of communication. Along with the negative evaluation this switch can lead to the ironic interpretation of the utterance. The third construction — "*s pozvolenija skazat*" 'if I am allowed to say so' — functions as a signal of linguistic categorization process. By using it the speaker shows that the object cannot belong to a particular category due to the lack of necessary properties.

PROCESSING OF QUANTITATIVE EXPRESSIONS WITH UNITS OF MEASUREMENT IN SCIENTIFIC TEXTS AS APPLIED TO BELARUSIAN AND RUSSIAN TEXT-TO-SPEECH SYNTHESIS

Skopinava A. M. (skelena777@gmail.com), **Hetsevich Yu. S.** (Yury.Hetsevich@gmail.com), **Lobanov B. M.** (Lobanov@newman.bas-net.by), United Institute of Informatics Problems of the NAS of Belarus, Minsk, Belarus

The article discusses problems of identification, analysis, classification (according to the International System of Units and separately according to word formation peculiarities), and processing of quantitative expressions (QE) with measurement units (MUs) as applied to text-to-speech synthesis by means of the linguistic processor NooJ and specially collected legal, scientific and technical text corpora for the Belarusian and Russian languages. In addition to a general description of algorithms and resources for finding QE in Belarusian and Russian texts, the paper gives an overview of QE with MUs with regard to how their components could be written, i.e. digital descriptors, and MUs proper (five different types). It is shown that QE with MUs can get the correct intonation marking only after they are properly generated, i. e. expanded into orthographical words.

PROCESSING OF CASE MORPHOLOGY: EVIDENCE FROM RUSSIAN

Slioussar N. (slioussar@gmail.com'), Utrecht Institute of Linguistics OTS, Utrecht, Netherlands; Saint-Petersburg State University, Saint-Petersburg, Russia,
Cherepovskaia N. (ajmi@yandex.ru), Saint-Petersburg State University, St. Petersburg, Russia

Many studies discuss how morphological ambiguity influences processing. In particular, it is well known that attraction errors in subject-verb agreement are produced more often and cause smaller delay in comprehension if the form of the intervening noun coincides with the Nominative case form. This is the case in the German example *die Stellungnahme gegen die Demonstrationen waren…* 'the position against the demonstrations (Acc.Pl=Nom.Pl) were' as opposed to *die Stellungnahme zu den Demonstrationen waren…* 'the position on the demonstrations (Dat. Pl≠Nom.Pl) were'. However, the explanation of this phenomenon is a matter of debate. How are such errors produced or missed in comprehension, how are ambiguous forms represented so that they can influence this process?.. We offer a novel perspective on this problem by looking at novel data. We conducted two self-paced reading experiments exploring how Russian adjective forms ambiguous for case influence processing of case errors on the following nouns. We compare sentences containing errors like *fil'my bez izvestnyh akterah* 'movie.$_{\text{NOM.PL}}$ without famous.$_{\text{GEN.PL=PREP.PL}}$ actor.$_{\text{PREP.PL}}$' and *fil'my bez izvestnyh akteram* 'movie.$_{\text{NOM.PL}}$ without famous.$_{\text{GEN.PL≠DAT.PL}}$ actor.$_{\text{DAT.PL}}$' to grammatically correct sentences. Errors of the first type are detected later and their effect is less pronounced. The results help answering several questions that arise in connection with attraction errors in subject-verb agreement.

WHAT KIND OF "SITUATIONS" UNDERLIE THE RUSSIAN VERBS «OTLICHIT' — OTLICHAT'» (DISCRIMINATE)

Sokolova E. G. (minegot@rambler.ru), Russian State University for the Humanities, Moscow, Russia, **Kononenko I. S.** (irina_k@cn.ru), Institute of Informatics Systems SB RAS, Novosibirsk, Russia

In the paper the method of Discourse Contexts is introduced to describe the semantics and use of one Russian verb pair that corresponds to situation of discrimination. Discrimination implies comparison resulting in differentiation and singling out. Discourse context is understood as complex entity including: (i) abstract Immanent Situation which underlies the use of verbs and represents a configuration of essential elements such as an idea of discrimination and entities involved including subject of discrimination and features of discrimination; (ii) Entity Situation in which the essential elements are classified according to concrete verb; (iii) Grammar Constituent. The analysis of the material of Russian National Corpus gives five types of discourse contexts for verbs *otlichit' — otlichat',* which are presented and exemplified in the paper. Discourse contexts are shown to help catch different meanings and explain semantic peculiarities of *otlichit' — otlichat'.*

DATABASE "LANGUAGES OF THE WORLD" AND IT'S APPLICATION. STATE OF THE ART

Solovyev V. D. (maki.solovyev@mail.ru), Kazan Federal University, Kazan, Russia, **Polyakov V. N.** (pvn-65@mail.ru), National University of Science and Technology "MISiS", Moscow, Russia

The article is dedicated to the largest digital resource in the world that contains a uniform description of language grammars — typological database "Languages of the World" ("Jazyki Mira"). There is information on the contents of the database, the programs for data procession. The database "Languages of the world" has three main areas of application: it can be used for quantitative researches, as a reference linguistic resource and for educational purposes. We give examples of database application in scientific researches in typology and areal linguistics. The examples demonstrate new opportunities of studying such questions as stability of grammatical features, liability to borrowing, typological and areal classification of languages. "Languages of the World" is compared with another famous typological database WALS.

GRAMMAR OF THE VERB AND DIALECTAL VARIATION

Tatevosov S. G. (tatevosov@gmail.com), Lomonosov Moscow State University, Moscow, Russia

The paper argues for a theory that accounts for the hierarchical structure of Russian verb. The theory assumes that possible derivations of verb stems are constrained by aspectual selectional characteristics of prefixes or by their position with respect to the "secondary imperfective" morpheme. Accordingly, two groups of prefixes can be identified, selectionally restricted and positionally restricted. The paper focuses on dialectal variation that determines class membership of individual prefixes and shows that this variation is conditioned by the same selectional and positional constraints. In that way, the dialectal variation provides further support for the proposed theory of the structure of Russian verb.

THE RUSSIAN ADVERB *VPORU* 'SUITING BEST': A PREDICATE WHICH DOES NOT COMBINE WITH NEGATION

Uryson E. V. (uryson@gmail.com), V. V. Vinigradov Russian Language Institute RAS, Moscow, Russia

The object of the paper is the Russian adverb VPORU 'suiting best'. In the 19 century the meaning of this word was less rich, so it was used in more types of contexts than now. At presence the adverb VPORU is freely used in three types of contexts: (a) *Pidzhak emu vporu* 'The coat is the right size for him'; (b) *Ej zamuzh vporu (a ona v kukly igraet)* 'She should marry (but not play with dolls)'; (c) *Zdes' tak temno — vporu na chetveren'kakh polzti* 'It is so dark here — one might as well crawl on his fours'. The adverb VPORU in (a) freely combines with negation in the 19 cent. language, but not in the the present day Russian. The reason is that in the 19 cent. language the meaning of the word VPORU in (a) is 'suiting' but not 'suiting best'. The latter meaning consists of two predicates. It is demonstrated that negation of such sense breaks a Grice maxim.

So, Grice maxims being applied to a meaning of an anomalous word combination can explain the reason of its anomaly. The adverb VPORU in (b) and (c) does not combine with negation. Contexts (b) are similar to (a). As for (c) the meaning of VPORU here has a rich modal frame. Being in the scope of negation the assertion of VPORU contradicts this modal frame; this reason of an anomaly of a word combination has been described by Ju. D. Apresjan [1978/1995].

SENTENCE INCOMPLETENESS VS. DISCOURSE INCOMPLETENESS: PITCH ACCENTS AND ACCENT PLACEMENT

Yanko T. E. (tanya_yanko@list.ru), Institute for Linguistics, Russian Academy of sciences, Moscow, Russia

The prosodic cues for discourse incompleteness may be either identical with the prosodic means expressing the topic or independent of marking the communicative constituents of a sentence: the topic or the focus. The autonomous prosodic marking of discourse incompleteness becomes possible in the context of tails. A tail is a fragment of a sentence placed after the accent-bearer of the focus. (Thus in the sentence *Malo ja smyslju v muzhskoj krasote* 'Little I know about men's attractiveness' with *malo* 'little' as the accent-bearer of the focus the fragment *ja smyslju v muzhskoj krasote* is the tail). A tail may be either deaccented or it may be used to carry the rise of discourse incompleteness. Generating a tail is conditioned by activation of entities within a sentence, contrast, emphasis, and verification expressed either by lexemes or by prosody, or both. In Russian, a tail can also result from a specific word order transformation with the focus accent-bearer being shifted to the left in front of the finite verb. The sentence-final verb, therefore, transforms into the tail to be specifically used as the bearer of discourse incompleteness pitch accent. (Thus in the sentence *Ja pidzhak snjal*… literally: 'I my coat took off…' with *pidhzak* 'coat' as the accent-bearer of the focus the sentence-final verb *snjal* 'took off' is the tail). Sentences with tails are able, therefore, to display a full set of communicative meanings including topics, focus and discourse incompleteness expressed by separate accent-bearers carrying the respective pitch accents.

COMPARISON OF OPEN INFORMATION EXTRACTION FOR ENGLISH AND SPANISH

Zhila A. (alisa_zh@mail.ru), **Gelbukh A.** (www.gelbukh.com), Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico

Open Information Extraction (IE) is the task of extracting relational tuples representing facts from text, with no prior specification of relation, no pre-specified vocabulary, or a manually tagged training corpus. Part-of-speech based systems are shown to be competitive with parsing-based systems on this task and work faster for large-scale corpora. Nevertheless, implementation of such a system requires language-specific information. So far, all work has been done for English. We present a relation extraction algorithm for Open IE in Spanish, based on POS-tagged input and semantic constraints. We provide a description of its implementation in an Open IE system for Spanish ExtrHech. We compare its performance with Open IE systems for English, including a comparison on a parallel English-Spanish dataset, and show that the performance is comparable with the state-of-the-art systems, while the system is more robust to noisy input. We give a comparative analysis of errors in extractions for both languages.

RANSITIVE IMPERSONALS IN SLAVIC AND GERMANIC: ZERO SUBJECTS AND THEMATIC RELATIONS

Zimmerling A. V. (fagraey64@hotmail.com), Institute for Modern Linguistic Research, SMSUH, Moscow, Russia

The paper argues that transitive impersonals in Russian, Ukrainian and Icelandic can be accounted for in terms of Mel'čuk's zero lexemes reanalyzed here as pronouns in the nominative case acting as agreement controllers. An alternative analysis resorting to Burzio's Generalization stipulates defective *v*P for different classes of verbs licensing transitive impersonals but fails to make correct predictions. The distribution of impersonals in Russian and Ukrainian does not depend on the distinction of unaccusative vs unergative vs psych predicates. Most Russian verbs labeled 'psych' in the previous generative research are either semantic causatives or agentive verbs with an external argument and valency grid <Agent, Patient>.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
Международной конференции «Диалог»

Выпуск 12 (19). 2013

Том 1. Основная программа конференции

Ответственный за выпуск **А. А. Белкина**
Вёрстка **К. А. Климентовский**

Подписано в печать 14.05.2013
Формат 152 × 235
Бумага офсетная
Тираж 250 экз. Заказ № 553

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06

Отпечатано с готового оригинал-макета в типографии
ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9

Яндекс

Чтобы создавать современные поисковые технологии, нужны глубокие знания в области математики, лингвистики, анализа данных, программирования и других дисциплин. Благодаря давним традициям российской науки и образования, Яндексу удалось создать команду сильных специалистов, которые и сделали Яндекс одной из ведущих IT-компаний в России. Присылайте свое резюме, если хотите присоединиться к команде.

Все вакансии Яндекса: http://company.yandex.ru/job/vacancies/

Аналитик отдела веб-поиска

В обязанности входит анализ веб-страниц и запросов пользователей, а также их поискового поведения. Необходимо отличать предпочтения пользователей и качество поиска от статистического шума и артефактов. Стать аналитиком может человек с высшим техническим или математическим образованием, свободно читающий профессиональную литературу на английском. Пригодится также общее понимание поисковых технологий.

Пройти тестовое задание: http://clck.ru/AnX5

Прикладной исследователь по направлению Data Mining / Information Retrieval

Работа для начинающих и опытных исследователей (data scientists, applied researchers). В обязанности входит разработка новых методов обработки информации, повышающих качество поиска, и описание их в статьях уровня ведущих академических конференций: SIGIR, WSDM, CIKM и т. п.

- Идентификация проблем поиска
- Анализ существующих решений
- Проведение экспериментов
- Написание научных статей

Пройти тестовое задание: http://clck.ru/8deUG

Стажировка в Яндексе

Мы приглашаем студентов, аспирантов и выпускников вузов в московский и петербургский офисы Яндекса. Вы сможете своими глазами увидеть, как в Яндексе создают интернет-сервисы, поработать над «боевыми» задачами, поучиться у знатоков своего дела. Проявите себя, и, вполне возможно, мы пригласим вас на постоянную работу.

- Опыт работы не важен
- Работу в Яндексе можно совмещать с учёбой
- Стажёры получают зарплату и бесплатные обеды

Заполните анкету: http://company.yandex.ru/job/intern/