ST PETERSBURG STATE UNIVERSITY
INSTITUTE FOR LINGUISTIC STUDIES (RAS)
HERZEN STATE PEDAGOGICAL UNIVERSITY OF RUSSIA

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
«CORPUS LINGUISTICS–2019»

June 24–28, 2019, St. Petersburg

SAINT PETERSBURG UNIVERSITY PRESS
2019

Организационный комитет конференции
«Корпусная лингвистика–2019»

В. П. Захаров (председатель), Е. Л. Алексеева,
Л. Н. Беляева (зам. председателя), Е. С. Гвоздёва, А. О. Гребенников,
А. А. Захарова, О. Н. Камшилова, О. Н. Крылова, О. В. Митренина,
О. А. Митрофанова, И. С. Николаев (зам. председателя),
В. И. Рубинер, М. В. Хохлова

Программный комитет конференции
«Корпусная лингвистика–2019»

В. П. Захаров (председатель), И. В. Азарова, Е. Л. Алексеева,
Л. Н. Беляева, В. Бенко (Словакия), С. Ю. Богданова, Н. В. Борисов,
В. В. Бочаров, Р. Вальденфельс (Германия), Л. А. Вербицкая,
Р. Гарабик (Словакия), А. Горак (Чехия), Т. Елинек (Чехия),
А. В. Зубов (Беларусь), Л. Л. Иомдин, Н. Н. Казанский, Е. Каллас (Эстония),
М. В. Копотев (Финляндия), Д. А. Кочаров, М. Кршен (Чехия),
М. А. Куниловская (Великобритания), У. Лоу (Великобритания),
О. Н. Ляшевская, В. Матоушек (Чехия), О. А. Митрофанова,
И. С. Николаев, Х. Нэси (Великобритания), К. Пала (Чехия),
В. Петкевич (Чехия), А. Ч. Пиперски, В. А. Плунгян,
Р. Рейнольдс (США), Л. В. Рычкова (Беларусь), С. О. Савчук,
В. П. Селегей, Д. В. Сичинава, О. Скривнер (США), В. Д. Соловьев,
А. Стефанович (Германия), Ю. Тао (Китай), М. В. Хохлова,
А. Я. Шайкевич, С. А. Шаров (Великобритания), Т. Ю. Шерстинова,
М. Шимкова (Словакия), С. Эйден (Франция), М. Якубичек (Чехия)

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ РАН
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ
ИМ. А. И. ГЕРЦЕНА

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2019»

24–28 июня 2019 г., Санкт-Петербург

ИЗДАТЕЛЬСТВО САНКТ-ПЕТЕРБУРГСКОГО УНИВЕРСИТЕТА
2019

ББК 81.1
Т78

Ответственный редактор издания
В. П. Захаров

Труды международной конференции «Корпусная лингвистика–
Т78 2019». — СПб.: Изд-во С.-Петерб. ун-та, 2019. — 448 с.

Сборник содержит материалы докладов, представленных на Международной научной конференции «Корпусная лингвистика-2019» 24–28 июня 2019 г. в Санкт-Петербурге.

Создание корпусов текстов является одним из приоритетных направлений в современной лингвистике. Проведение конференции по данной тематике знакомит ученых с современными разработками и новыми технологическими решениями в этой области, а также способствует обобщению опыта научных исследований по корпусной лингвистике.

ББК 81.1

ОГЛАВЛЕНИЕ

ПЛЕНАРНЫЕ ДОКЛАДЫ

KEYNOTE TALKS

M. Hnátková, T. Jelínek, M. Kopřivová,
V. Petkevič, A. Rosen, H. Skoumalová, P. Vondřička

LEXICAL DATABASE OF MULTIWORD EXPRESSIONS IN CZECH[1]

Abstract. This abstract describes basic features of the representative lexical database of multi-word expressions in Czech, called LEMUR. The paper (i) sketchily depicts the content of a database entry based on a multifaceted typology of Czech MWEs and (ii) technical aspects of a database entry in more detail.
Keywords. Multiword expressions, multifaceted typology, fixedness, flexibility, idiomaticity.

1. Introduction

In this paper, we briefly describe a multifaceted typology of Czech multiword expressions (MWE) and a representative lexical database (lexicon), named LEMUR, currently (as of June 2019) including more than 7,000 entries. Both the typology and the database are important for many reasons concerning:

(a) lexicography
(b) key NLP tasks such as part-of-speech tagging and parsing, word sense and semantic disambiguation
(c) theoretical issues of MWEs as partially fixed units standing between lexicon and grammar in general and often having a specific meaning
(d) identification and search of MWEs in corpus data
(e) teaching Czech as a foreign language
(f) other.

In addition to the necessary flexibility in morphology, word order and other specific features of Czech as an inflectional language, the typology and

[1] This paper was supported by the Grant Agency of the Czech Republic reg. No. 16-07473S.

the database are designed to be robust so as to account not only for standard forms of MWEs but also for their modifications/fragments as a reflection of creativity of language users.

In part 2 we describe the adopted typology of MWEs being reflected in database entries, and in part 3, a core part, the structure of a database entry is outlined.

2. MWE typology reflected in LEMUR

In this section, we very sketchily describe the typology of MWE, i.e. the features on various levels of linguistic description we account for in a database entry (cf. Hnátková et al. 2017a, Hnátková et al. 2017b for more detail).

MWEs can be defined as lexical items that

(a) can be decomposed into multiple lexemes
(b) display lexical, syntactic, semantic pragmatic and/or statistical idiomaticity.

Especially the second property is complex and deserves a more detailed explanation. However, idiomaticity — in addition to its role in the diagnostics of MWEs — and its types can be used to distinguish MWEs as a part of their taxonomy.

On the basis of the proposal described in Baldwin et al. 2002, the PARSEME project (http://typo.uni-konstanz.de/parseme) categorizes MWEs simultaneously according to their:

- syntactic structure
- fixedness and flexibility
- idiomaticity.

We adopt and extend this taxonomy. Most of the extensions are motivated by the goal to design a lexical template for MWEs useful for a human user as well as NLP applications, but some extensions reflect the properties of Czech as an inflectional language with a significant degree of main constituent order freedom, relevant also for MWEs. The extensions concern the following aspects of MWEs: definition, examples, usage type, valency patterns, use of fragments and variants, register/stylistic markedness and a more detailed specification of some types (e.g. morphological idiomaticity).

We divide the description of every MWE in the database into two parts (cf. a detailed description in Hnátková et al. 2017b with characteristic examples):

- global properties, describing the MWE as a whole
- local properties, describing single positions (words) in the MWE.

The description of global properties consists of:

(i) MWE lemma, typically as a sequence of individual MWE components
(ii) MWE definition explaining the MWE's meaning
(iii) relevant examples found in corpora of Czech
(iv) basic part-of-speech pattern as a sequence of extended part-of-speech codes
(v) syntactic structure expressed as a dependency tree and a phrase-structure tree
(vi) word order variability

Type of MWE according to three different categorizations:

- *Usage*/*global type* characterizing a MWE as being one of the following kinds: term, proverb, saying/locution, citation, comparison/simile, other
- *Idiomaticity* describing the degree of MWE's idiosyncrasis on the morphological, syntactic, semantic, lexical and pragmatic level, and also mere statistical idiomaticity (concerning fixed and semantically compositional collocations)
- *Syntactic type* concerning a categorization of MWEs as to their core syntactic structure (noun-headed phrase, adjectival phrase…) Within the MWE syntactic description we also account for
  (i) possible syntactic *transformations* of MWEs:
    - passivization: MWE can/cannot be passivized
    - topicalization: MWE can/cannot be topicalized
    - nominalization: MWE can/cannot be nominalized
    - adjectivization: MWE can/cannot be adjectivized
    - Reflexivization
  (ii) *valency* of the entire MWE: (*vzít zavděk* 'be happy with') + NP_{Ins}
  (iii) *MWE fragments and roots:* Some MWEs occur in texts as fragments, therefore MWEs should be recognizable via two or more characteristic words, called *roots*.

The description of local properties concerns features of every MWE word/component:

(i) fully morphologically disambiguated word forms where appropriate

(ii) morphological variability — every morphologically variable word form in a MWE is described via morphological categories and their values
(iii) lexical variants
(iv) style/register
(v) internal syntactic modifiability
(vi) negation/affirmation

3. Database entry

In this part we describe technical aspects of a database entry, i.e. an implementation of abovementioned properties we account for (cf. Vondřička (in press)).

3.1 Slots and fillers

A MWE consists of two or more components, i.e. words. They may be more or less fixed since components may be realized:

- by one particular word/lexeme, or
- by a selection of several different words/lexemes.

They can be formed by

- any lexeme
- a whole phrase

of a particular type.

Moreover, they may be

- freely inflected/modified, or
- subject to various restrictions.

Therefore, each MWE is defined via its components and their various possible realizations.

The entry pattern is specified by means of *slots* and *fillers*. Each slot represents a single component of the MWE (pattern), which constitutes the syntagmatic dimension of the MWE. Slots consist of fillers and slot-specific features, with fillers representing the paradigmatic dimension of the components: the possible variants which may be used to realize a particular component. The primary role of fillers is to represent actual (terminal) tokens to be matched in the data. As the tokens in the corpora of Czech are annotated by a combination of positional attributes (*lemma* and *tag*) and their values, the fillers are defined by them as well.

Examples are shown in Table 1:

Table 1. **Example definitions of positional attributes for different types of fillers**

lemma="flinta" tag="NNF[SP]1"	noun *flinta* 'gun' in nominative singular or plural
lemma="hodit" tag="V"	verb *hodit* 'throw / suit' in any form
tag="AA"	any common adjective in any form

The fillers may actually declare just any arbitrary positional attributes used to identify the matching tokens. Other restrictions (e.g. syntactic ones), such as possible word order, modifications or transformations, are defined by additional features.

The attributes to be matched may also be underspecified: the tag value may contain just a prefix referring to the part-of-speech or a regular expression to match a custom choice of acceptable morphological forms. Specification of the lemma may be completely avoided if any lexeme of some particular part of speech or morphological category may fill the position — an *open slot*, but its presence is still necessary (or typical) for the identification of the MWE (cf. examples in Table 1). Moreover, the filler may provide its own additional features.

For strictly fixed expressions, a slot mostly contains only one possible filler defining a particular type of token to be matched. More flexible expressions may contain a list of several synonymous/alternative fillers. As the fillers may also have their own features, it is possible to document e.g. their actual relative usage or further individual effects on the other slots or on the MWE as a whole. Such slots can be classified as *fixed* (closed). In case of relatively open slots, the fillers may be underspecified as mentioned above. They may also represent only the most typical representatives of a relatively open semantic class. This is relevant in cases where such a group of acceptable fillers cannot be fully defined formally in an explicit way: this third kind of slots is called *semi-open*. Such incomplete description can currently only be of limited use for a NLP parser, but it will still remain a useful hint for human users of the database.

If a slot/filler is to represent a whole phrase of some type (e.g. in the case of valency elements), a combination of positional attribute values to match one single token cannot be used any more. Specific features to define the

phrase type (restriction) are necessary instead. Such a description can be useful for human users and later also for possible higher-level parsers operating also at the syntactic level.

3.2 Features and their classification

Features are generic pairs of type (name) and value. For easier organization and systematization of various types of features, we use a hierarchical system of specification of the type by means of a path in an arbitrary hierarchy of features, using colon as the separator. At the top level, features are classified as morphological, syntactic, semantic, statistical, related to the form, purely user-oriented or editor-oriented notes, etc. Further levels are divided as needed: as specific groups of features, source of data, by particular theory, etc. This also makes it possible to store multiple similar features from different sources (or for different purposes) at the same time.

If multiple alternative values of some type of feature are to be included, custom subspecification may be used. This concerns primarily user notes, examples from real texts or statistical values. For example, the basic type of features for absolute frequency :stat:fq:abs is expected to be extended by additional custom subspecification of the corpus (and possibly subcorpus) used to acquire the frequency value, e.g. :stat:fq:abs:CNC:fiction. Thus the database can be searchable by features both using underspecification of the type (by means of a path prefix) or its full (sub) specification as needed.

3.3 Tree structures

In the database, tree structures such as dependency and constituency structure of a MWE expression are accounted for as well. Dependency relations between the components can easily be recorded in the form of slot features. One single feature is needed as a reference to the parent slot and another one to identify the syntactic function of the component. Such relations can easily be projected into a resulting tree structure.

Constituency trees, however, need non-terminal nodes and we need to be able to refer and assign features to them as well. Therefore, they are represented by standalone objects in the database, equivalent to the slots. That is the reason why just the grouping of components by means of features would not be a satisfactory solution.

The flat structure of slots and fillers excludes nesting. Previously, we have suggested to use recursive structures for lexical descriptions, where fillers can branch into further sequences of "subslots." However, indexing and querying recursive data structures is still a very demanding task not well sup-

ported by the current database and search engines. Therefore the idea was abandoned. Instead, we decided to keep to the flat structure, but to allow fillers to refer to a sequence of other slots by means of their identifiers (labels). This makes it possible to add non-terminal fillers (and their respective non-terminal slots).

Various advantages and disadvantages emerge from this design: indexing and searching for both terminal and non-terminal nodes is equally simple, but traversing relations between them in a single query is not supported by the search engine.

This means that searching for MWEs via their structure — e.g. by syntactic (or other) relations — would be difficult to implement. Currently, we do not expect the need to search the database by tree structures, but if necessary, the structures can be reconstructed for all entries, encoded into some kind of searchable patterns and indexed separately by the same or a more appropriate engine. Another advantage is the possibility to record several independent tree structures within a single entry, which corresponds well to the requirement of multifunctionality. A partial disadvantage is the potential need for treatment of possible partial trees, overlapping trees and orphan nodes.

4. Conclusion

In this paper, a complex typology of MWEs in Czech and MWE representative database were briefly presented. The stress was laid on the description of a database entry (part 3). The database is constantly developed, i.e.

(i) it is enhanced with new MWEs identified in SYN-series corpora of contemporary Czech (Czech National Corpus Project: http://korpus.cz), and
(ii) the content of database entries is refined and enhanced with new features.

References

1. *Baldwin T., Kim S. N.* (2010): Multiword Expressions. In: Indurkhya N., Damerau F. J. (eds.), *Handbook of Natural Language Processing*. Boca Raton: CRC Press, 2nd edn., 267–292.
2. *Hnátková M., Jelínek T., Kopřivová M, Petkevič V., Rosen A., Skoumalová H., Vondřička P.* (2017a): Eye of a Needle in a haystack. Multiword Expressions in Czech: Typology and Lexicon. In: Mitkov R. (ed.), *Computational and Corpus-Based Phraseology:*

Second International Conference, Europhras 2017, London, UK, November 13–14, 2017, Proceedings, Springer International Publishing, LNAI 10596, pp. 160–175, ISBN: 978-3-319-69805-2, DOI: 10.1007/978-3-319-69805-2_12, URL: https://doi.org/10.1007/978-3-319-69805-2_12.

3. *Hnátková M., Petkevič V., Skoumalová H.* (2017b): Multiword Expressions in Czech: Between Lexicon and Grammar. In: *Trudy meždunarodnoj konferencii „Korpusnaja lingvistika — 2017"* (*Proceedings of the International Conference „Corpus Linguistics — 2017"*). June 27–30, 2017. St. Petersburg: St.-Petersburg State University, Institute of Linguistic Studies (RAS), Russian State Herzen Pedagogical University, 36–42.
4. *Vondřička P.* (in press): *Design of a MultiWord Expressions Database*.

Milena Hnátková, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová, Pavel Vondřička
Faculty of Arts, Charles University
E-mail: vladimir.petkevic@ff.cuni.cz

R. Reynolds

RUSSIAN NLP FOR LANGUAGE LEARNERS: TECHNOLOGIES AND APPLICATIONS

Abstract. In this talk, I discuss my research in Russian natural language processing tools intended for applications in Computer-Assisted Language Learning. As should be expected, research on natural language processing is dominated by applications for native speakers using normative language, i.e. language that conforms to orthographic and grammatical standards. This means that the implicit assumptions in the design of mainstream tools can be ill-suited for applications intended for non-native speakers, whether processing normative language or learner language. These assumptions involve the content of system, the nature of the information that the system delivers, and the confidence with which the system delivers it. I present a two-level morphological analyzer and constraint grammar for Russian, UDAR (short for *udarenie*), discussing the explicit design decisions that make it more amenable to language-learning applications. I present the results corpus-based evaluation of two tasks: 1) automatic wordstress placement in unrestricted text, and 2) automatic detection/diagnosis/correction of learner errors. I also briefly showcase a number of language-learning applications which rely on UDAR. Among these are automatic wordstress placement in Russian running text, a web browser extension for automatically generating grammatical exercises in context, and a web search engine for Russian language teachers and learners.

Robert Reynolds
Office of Digital Humanities, Brigham Young University, Provo, Utah, USA
E-mail: robert_reynolds@byu.edu

ОБЩИЕ ВОПРОСЫ КОРПУСНОЙ ЛИНГВИСТИКИ

GENERAL ISSUES OF CORPUS LINGUISTICS

А. Евдокимова, Ю. Николаева
A. Evdokimova, Yu. Nikolaeva

НЕМАНУАЛЬНЫЕ ДВИЖЕНИЯ В КОММУНИКАЦИИ И ИХ КОМПЛЕКСНОЕ ОПИСАНИЕ

NONMANUAL COMMUNICATION MOVEMENTS AND THEIR INTEGRATED ANOTATION

Аннотация. В статье предлагается система описания движений локтей и их взаимосвязь с движениями плеч, корпуса и ног. Статья призвана дополнить существующие подходы к описанию жестов в коммуникации и уточнить систему аннотации мануальных жестов. Предлагается единый подход, учитывающий формальные признаки рассматриваемых движений, их коммуникативную роль и взаимосвязь с другими движениями; кроме того, рассматриваются индивидуальные особенности кинетического поведения.
Ключевые слова: мультиканальная аннотация, жесты, движения плеч, движения локтя.

Abstract. The article presents a framework for annotation of elbow movements and their interconnections with shoulders, torso and legs movements. We propose a united approach considering formal features of these movements, their function in communication and alignment with movements of other body parts. In addition, we consider individual characteristics of kinetic behavior in communication.
Keywords: multichannel annotation, gesticulation, elbow movements, shoulder movements.

1. Введение

С середины XX века наблюдается рост интереса к жестам со стороны лингвистики (хотя уже со времен античность роль жестов, сопровождающих речь, привлекала внимание разных авторов). При этом под жестикуляцией понимаются в основном или даже исключительно жесты рук, а еще точнее, движения кистей рук [см. напр. McNeill 1992]. В нашей работе мы рассматриваем и другие кинетические каналы:

движения головы, локтей, плеч, туловища и ног. Данная статья анонсирует метод, благодаря которому в рамках единого подхода могут быть описаны все движения в процессе естественной коммуникации, на примере анализа комплекса движений, возникающих в кинетических каналах в момент движений локтей.

2. Метод и материал

Мы использовали для исследования корпус [www.multidiscourse.ru], «Рассказы и разговоры о грушах» [Кибрик, Федорова 2018]; видеозаписи этого ресурса содержат пересказы «Фильма о грушах» У. Чейфа [Chafe 1980]. В каждой сессии участвовало 4 человека: Рассказчик (N), смотревший фильм, должен был передать его содержание Пересказчику (R), не видевшему фильма. Комментатор (C) тоже видел фильм и после рассказа N мог добавлять и поправлять его, а R мог задавать им любые вопросы. После этих двух этапов (рассказ и обсуждение) наступал третий, когда к участникам присоединялся Слушатель (L), и R описывал ему содержание фильма.

Рассмотренный нами материал состоял из 3 записей (04, 22 и 23). Каждая запись включала данные трех видеокамер, направленных на трех первых участников (N, C, R) и одной общей видеокамеры для последнего (L), таким образом общая длительность изученного видео составила 240 минут. Для всех участников были размечены движения кистей рук, локтей, головы, плеч, корпуса и ног в программе ELAN. Анализ указанных аннотаций и послужил материалом для выводов и наблюдений, предложенных ниже.

3. Движения локтя

Традиционно движения локтей, в большей или меньшей степени, включались в движения рук, т. е. не различались движения кисти руки и руки от плеча [см. напр. McNeill 1992]. Однако это не всегда было правильно с позиций точности описываемых явлений: иногда движение локтя не сопровождалось заметным движением кисти рук.

В соответствии с принятым нами подходом движения локтя делятся на жесты и адапторы в зависимости от того, выполняли движения коммуникативную функцию, или их целью был комфорт говорящего; кроме этого, движение может быть отнесено к сменам положения.

Помимо этих трех типов движений, инициированных в рассматриваемом канале, могут быть движения, инициированные в другом

кинетическом канале, называемые **эхо** (например, короткие кивки головой при смехе, или подскок руки вверх, когда она слишком резко опустилась на колени) и **перемещения** (например, смещение головы, вызванное движением корпуса, когда не были никак задействованы мышцы шеи). Применительно к рукам эхо и перемещения считаются несущественными с точки зрения их роли в коммуникации и не отмечаются систематически, однако они играют большую роль, когда рассматриваются движения головы. В описании движений локтя мы будем придерживаться того же подхода, что и при аннотировании жестов рук, выделяя только движения, инициированные внутри этого канала.

Также важно разграничить движения локтя и движения рук, описываемые в мануальной аннотации. Мы выделяем только такие движения локтя, амплитуда которых больше амплитуды смещения кисти на том же временном отрезке, при этом они не являются частью вспомогательной фазы для жеста руки.

4. Результаты

Как следует из изученного материала, существенная часть движений локтя выполнена левой рукой, хотя все участники записи правши. Еще одно наблюдение, опровергающее стереотипный взгляд на жесты, состоит в том, что нередко жест локтя продолжается, когда говорящий уже закончил свою реплику и выступает в роли слушающего, что крайне нетипично для жестов рук.

При том, что выразительность таких движений гораздо меньше, чем у жестов рук, жесты локтя также могут передавать некоторую дополнительную информацию: например, прагматическую (подчеркивая замечание о завершении монолога, или призывая другого участника высказаться).

У каждого участника есть свой достаточно стереотипный набор движений (например, участник с кодом 04C1 опирается локтями на колени, его самые частые движения — опереться локтями на бедра, при этом локти иногда соскальзывают; участник 04N делает движение локтем от себя, сопровождая или подчеркивая жесты).

Если сопоставить количество движений локтей и движений рук у каждого участника, получается следующая картина (см. табл. 1).

[1] Код участника состоит из номера записи (04, 22 или 23) и буквы, обозначающей роль участника.

Таблица 1. **Количество жестов и адапторов локтей и рук у каждого участника**

	Локти адапторы	Руки адапторы	%	Локти жесты	Руки жесты	%
04C	8	172	4,7 %	1	105	1,0 %
04N	3	200	1,5 %	12	502	2,4 %
04R	6	234	2,6 %	44	557	7,9 %
22N	2	64	3,1 %	0	362	0,0 %
22R	2	180	1,1 %	4	145	2,8 %
23C	1	89	1,1 %	8	167	4,8 %
23N	1	122	0,8 %	5	405	1,2 %
23R	5	135	3,7 %	54	297	18,2 %

При анализе сопутствующих жестам и адапторам локтей движений, происходящих в других кинетических каналах (цефалическом, канале туловища, плеч и канале ног) были выявлены следующие закономерности. Для большинства испытуемых движение локтя дает эхо в соответствующем плече, и плечо изменяет свое положение в пространстве. В некоторых случаях (около 15 %) инициатором движения локтя выступает плечо, и формально это либо перемещение, либо эховое движение локтя. В цефалическом канале может происходить наложение на движение локтей собственно жестов головы и эховых движений или перемещений, в зависимости от особенностей цефалического портрета испытуемого. В канале туловища встречается либо подстройка туловища под движения локтей, чаще всего наклоны вперед или назад, либо эховые движения, которые в свою очередь могут наложится на собственные жесты корпуса. В канале ног при опоре локтей на них у многих испытуемых отмечается синхронизация ног друг с другом и отзеркаливание движений локтей. Так, например, 4C то сдвигает ноги, то раздвигает, а 4R то поднимает колени, то опускает. Возможны и другие стратегии, которые будут более подробно разобраны в докладе с соответствующими корреляциями общих поз испытуемых.

Приведем некоторые примеры из анализа наших материалов. Так большая часть движений локтя у участника 4C — это адапторы в следующей позе: локти опираются на ноги, поставленные на ширине

плеч. В момент, когда физиологически опора перестает быть удобной, и левый локоть меняет свое положение, в других каналах происходит следующее: правая нога синхронизируется с левой и делает зеркальные движения, левая нога подстраивается под локоть, чаще всего двумя движениями, левое плечо поднимается вслед за локтем, перемещение и эхо в корпусе, в голове — эхо от корпуса. Если меняет позу правый локоть, то корпус смещается сначала вперед и затем назад, голова перемещается, в ногах сначала эхо, потом правая нога подстраивается под локоть, а левая синхронизируется с ней. Когда локти не опираются на колени, и руки сложены на груди, то при смене позы из всех каналов двигается только соответствующее плечо.

Для 4N типической оказывается поза, когда кисть левой руки опирается на коленку, что приводит к эху в голове от сопутствующих движений корпуса. При этом чаще всего плечи идут вверх и участвуют в движении. В некоторых случаях движения локтя сопровождаются жестом плеча и разворотом головы и корпуса. 23R сидит, держа ноги на ширине плеч, локти упираются в бедра. У него часто наблюдается наклон корпуса вперед или назад, вместе с корпусом перемещается голова, иногда она подстраивается под корпус и наклоняется вместе с ним.

Согласно нашим наблюдениям, очень часто движения локтя входят в вертикальные кластеры жестов, которые собираются в разных каналах, подобно описанным нами на примере взаимодействия цефалического и мануального каналов [см. Евдокимова, Николаева, Сухова 2019].

Литература

1. *Chafe W.* (1980), The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. Norwood, New Jersey: Ablex.
2. *McNeill, D.* (1992), Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.
3. *McNeill, D.* (2005), Gesture and thought. Chicago: University of Chicago Press.
4. *Евдокимова А. А., Николаева Ю. В., Сухова Н. В.* (2019), Мультиканальные кластеры: цефалическое и мануальное взаимодействие в устном дискурсе (в печати).
5. *Кибрик А. А., Федорова О. В.* (2018), An empirical study of multichannel communication: Russian Pear Chats and Stories. Психология. Журнал Высшей Школы экономики, 15 (2), с. 191–200.

References

1. *Chafe W.* (1980), The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. Norwood, New Jersey: Ablex.
2. *McNeill, D.* (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.
3. *McNeill, D.* (2005). Gesture and thought. Chicago: University of Chicago Press.
4. *Evdokimova A., Nikolaeva Y., Sukhova N.* (2019) Multichannel clusters: cephalic and manual interaction in face-to-face discourse (in press).
5. *Kibrik A. A., Fedorova O. V.* (2018), An empirical study of multichannel communication: Russian Pear Chats and Stories. Psychology. Journal of the Higher School of Economics, 15 (2), pp. 191–200.

Николаева Юлия Владимировна
Московский государственный университет им. М. В. Ломоносова (Россия)
Nikolaeva Yulia
Lomonosov Moscow State University (Russia)
E-mail: julianikk@gmail.com

Евдокимова Александра Алексеевна
Институт языкознания РАН (Россия)
Evdokimova Alexandra
Institute of linguistics of Russian academy of sciences (Russia)
E-mail: arochka@gmail.com

А. В. Зубов
A. V. Zubov

СОЗДАНИЕ БОЛЬШОГО КОРПУСА ТЕКСТОВ БЕЛОРУССКОГО ЯЗЫКА И ЕГО ИСПОЛЬЗОВАНИЕ ДЛЯ ИЗУЧЕНИЯ БЕЛОРУССКОГО ЯЗЫКА И ЕГО СВЯЗИ С ДРУГИМИ ЯЗЫКАМИ ЕВРОПЫ

THE CREATION OF THE LARGE CORPUS OF BELARUSSIAN LANGUAGE AND THE USE OF IT FOR THE INVESIGATION THE BELARUSSIAN LANGUAGE AND ITS CONNECTION WITH THE DIFFERENT LANGUAGES OF EUROPEAN

Аннотация. Созданный корпус текстов включает письменные источники двух жанров: художественную литературу (проза, драма, стихи) и публицистику (газеты, журналы). Этот корпус состоит из тегированного корпуса текстов белорусского языка, содержащего 1 млн словоупотреблений, и трех параллельных тегированных подкорпусов текстов: русско-белорусского, англо-белорусского и немецко-белорусского. Каждый из них содержит по 300 000 словоупотреблений. Для возможности извлечения из этих подкорпусов различной лексической, грамматической и структурной информации, и ее автоматической обработки созданы 4 компьютерные программы для обработки белорусского корпуса текстов и 8 компьютерных программ для работы с параллельными корпусами текстов.
Ключевые слова. Художественный, публицистика, белорусский корпус, тегированные подкорпусы, текст, компьютерные программы.

Abstract. The organized corpus of texts included the texts of belles letters and political. The corpus of Belarussian languages included 1 mln words and three podcorpuses have on 300 000 words. In order to extract of linguistic informations from these corpuses was make 12 computer programs.
Keywords. Corpus, text, Belarussian language, words, computer program.

В последние годы компьютеры все больше используются для решения различных лингвистических задач. На этом пути ярко выявились те проблемы, которые компьютер не может решить, имея в памяти отдельные тексты или словари. Все чаще взоры исследователей обращаются к корпусам текстов, содержащим в себе множество текстов различного типа и определенную информацию о каждом тексте, предложении и слове. По этим проблемам издаются специальные журналы, монографии и сборники статей, ежегодно проводятся научные конференции.

Понятие «корпус текстов», как и большинство лингвистических понятий, не имеет единого общепринятого определения. Авторы

впервые созданного в 1963 году корпуса текстов («Брауновский корпус») У. Френсиз и Г. Кучера употребили это понятие в значении «совокупность текстов, считающаяся представительной для данного языка, диалекта или другого подмножества языка, предназначенная для лингвистического анализа». Этот корпус состоял из 500 отрывков разных текстов печатной прозы США, каждый из которых содержал 2000 словоупотреблений. Они представляли 15 наиболее массовых жанров англоязычной печатной прозы 60-х годов.

При составлении этого корпуса текстов его авторы выдвинули ряд специфических критериев отбора текстов в корпус текстов [5]:

- наличие определенных требований к авторам текстов, отбирая тексты, они выбирали таких авторов, которые были носителями американского варианта английского языка;
- определенные требования к физической структуре отбираемых отрывков текстов, например, если в тексте были диалоги, то они должны занимать менее половины объема отрывка;
- сохранение временного интервала написания текстов (У. Френсиз и Г. Кучера отбирали тексты, изданные только в 1961 году);
- наличие специальной методики количественного отбора жанров и отдельных текстов в жанры, при создании Брауновского корпуса отбор отрывков текста проводился с учетом вероятностной процедуры;
- специальная предварительная обработка текстов с целью возможности их дальнейшей компьютерной обработки.

Обобщая эти критерии, последующие исследователи выдвинули пять минимальных базовых качеств, делающих некоторое множество текстов — корпусом текстов. К их числу относят следующие качества [6]:

- расположение множества текстов на магнитном носителе;
- наличие определенной процедуры отбора текстов в корпус текстов;
- единая методика представления сведений о текстах и их единицах на магнитном носителе;
- конечный размер корпуса текстов;
- репрезентативность множества текстов, входящих в корпус.

Цикл исследований, связанных с правилами организации текстов в корпус, разработкой алгоритмов анализа таких текстов в рамках не-

которой научной методологии получил название **корпусная лингвистика** [1].

В отличие от других приемов и способов проведения лингвистических исследований, корпусной анализ характеризуется следующими особенностями:

- он является исключительно эмпирическим, так как опирается на анализ реальных примеров, использованных в естественных текстах;
- его основой является, как было отмечено выше, специальным образом построенное большое собрание текстов естественных языков;
- он широко использует компьютерный анализ, в том числе автоматический, и интерактивные приемы;
- он опирается на количественный и качественный приемы.

Достаточно большой опыт работы с корпусами текстов показывает, что с их помощью можно по-новому решить целый ряд лингвистических задач:

- в лексикографии и лексикологии (для составления различных словарей, определения значений многоязычных слов, выявления ассоциативных связей слов в тексте, выявление терминов и терминологических словосочетаний и т. д.);
- в грамматике (для определения частоты употребления грамматических морфем в текстах различного типа, выявление наиболее употребляемых типов словосочетаний и предложений, определения частоты употребления классов слов и т. д.);
- в лингвистике текста (для дифференциации типов текста, создание конкордансов, выявление связи между предложениями в абзацах и между абзацами и т. д.);
- при автоматическом переводе текстов (для поиска контекстов слов, имеющих несколько переводных эквивалентов, поиска переводных эквивалентов терминологических и фразеологических словосочетаний в параллельных текстах и т. д.);
- в учебных целях (для выбора цитат, отдельных произведений и примеров, используемых в процессе создания учебников и учебных пособий и т. д.).

Известны данные о корпусах текстов английского, немецкого, французского, испанского языков, создаваемых в рамках таких круп-

ных научных объединений как ELRA (European Language Resources Association) и TELRI (Trans-European Language Resources Infrastructure). Участники этих объединений подготовили сотни различных корпусов, как письменных текстов, так и устной речи. Менее известна информация о корпусах текстов славянских языков. Такие корпуса создаются в России, в Германии, Финляндии, на Украине и в других государствах [3].

Корпусная лингвистика в Республике Беларусь стала быстро развиваться с 2007 года в результате активного сотрудничества Института языка и литературы Национальной Академии наук Беларуси и Минского государственного лингвистического университета.

Для создания большого корпуса текстов белорусского языка были отобраны тексты 15-и белорусских писателей и поэтов. Публицистика в создаваемом корпусе белорусских текстов была представлена статьями из таких газет и журналов как «Звязда», «Літаратурная Беларусь», «Новы час» и некоторых других за 2006–2010 годы.

Как показали результаты использования различных корпусов текстов для их практического лингвистического анализа, многие лингвистические задачи с их помощью не могут быть решены, если тексты и их составляющие не имеют специальных индексов (тегов). В создаваемом корпусе текстов были приняты следующие группы признаков для тегирования:

1) признаки для кодирования всего корпуса текстов;
2) признаки для кодирования отдельных текстов;
3) признаки для кодирования предложений;
4) признаки для кодирования словосочетаний;
5) признаки для кодирования слов;
6) признаки для кодирования морфем в словах.

Каждая из этих групп включала от 4 до 12 различных конкретных признаков [2].

Возможны 3 различных способа тегирования письменных корпусов:

A. ручной;
B. полуавтоматический;
C. автоматический.

В первом случае каждое словоупотребление текста, каждое словосочетание и предложение вручную снабжаются необходимым набором индексов.

При полуавтоматической разметке компьютерная система по результатам работы заложенных в систему программ автоматического морфологического и синтаксического анализа дает словоформам и предложениям текста определенные индексы. Затем результат такой автоматической разметки высвечивается на экране, и лингвист проводит корректировку полученной разметки.

При полностью автоматической разметке весь набор необходимых индексов приписывается единицам текстов автоматически. Разработан целый ряд систем для такой автоматической разметки единиц текста [6].

Для возможности в последующем обмена информацией с другими странами при разработке корпуса текстов белорусского языка принят стандарт CES (Corpus Encoding Standard), который широко используется при разработке европейских проектов MULTEXT 135 и EAGLES (Expert Advisory Group on Language Engineering Standard) в сотрудничестве с американским партнером Vassar College и французским партнером CNRS (Centre National de la Recherche Scientifique).

Этот стандарт еще удобен тем, что он специально создан для автоматического решения задач прикладной лингвистики, машинного перевода, лексикографии и т. п.

«Приписывание» тегов всем словоупотреблениям белорусских текстов проводилось автоматически с опорой на существующий базовый словарь белорусского языка.

В результате был создан тегированный корпус текстов современного белорусского языка объемом в 1 млн словоупотреблений.

Помимо описанного корпуса текстов белорусского языка, кафедра информатики и прикладной лингвистики МГЛУ создала 3 параллельных тегированных подкорпусов текстов: русско-белорусский, англо-белорусский и немецко-белорусский, каждый объемом в 300 000 словоупотреблений. Тегирование словоупотреблений используемых при этом белорусских текстов, а также их переводов на русский, английский и немецкий языки проводилось в следующей последовательности:

A. для каждого текста, включаемого в подкорпус, компьютер строил его алфавитный словарь словоформ;
B. специалисты по белорусскому, русскому, английскому и немецкому языкам «приписывали» словоформе весь необходимый набор признаков, в тех случаях, когда конкретная словоформа не могла быть описана однозначным набором признаков, ей

«приписывается» два таких набора признаков, отделяемых друг от друга наклонной чертой (/).

Для возможности использования белорусского корпуса текстов для извлечения различной лингвистической информации нами были созданы 4 компьютерные программы. 8 компьютерных программ были созданы для автоматического извлечения лингвистической информации из параллельных корпусов текстов. Так, они могут стать основой для совершенствования систем машинного перевода текстов с различных языков на белорусский язык:

- для постоянного пополнения автоматического двуязычного словаря;
- при автоматическом разрешении лексической неоднозначности;
- для автоматического разрешения синтаксической неоднозначности;
- при создании систем автоматического перевода с переводческой памятью;
- при автоматическом сопоставлении переводного конкорданса;
- при автоматическом переводе терминологических словосочетаний, фразеологизмов и идиом.

Созданный тегированный корпус текстов белорусского языка и 3 параллельных белорусско-иноязычных подкорпуса текстов отвечают самым последним требованиям, предъявляемым к корпусам текстов. Для возможности обмена информацией, содержащейся в созданных корпусах текстов, с информацией, представленной в корпусах текстов других стран, в наших корпусах принят стандарт тегирования CES (Corpus Encoding Standart), который был широко использован при разработке аналогичных европейских проектов MULTEXT 135 и EAGLES (Expert Advisory Group of Language Engineering Standart) в сотрудничестве с американским партнером Vassar College и французским партнером CNRS (Centre National de la Reckosche Scientifique) [4].

Литература

1. *Захаров В. П., Коваль С. А.* (2002), Корпусная лингвистика НТИ. Серия 2 Информационные процессы и системы, № 7, с. 34–49.
2. *Зубов А. В.* (2005), Корпусная лингвистика: возможности и проблемы. Актуальные проблемы компьютерной лингвистики: Сборник научных статей, Минск, МГЛУ, с. 60–77.

3. *Копотев В., Мустойоки А.* (2008), Современная корпусная русистика. Инструментарий русистики: корпусные подходы, Хельсинки, с. 7–24.
4. *Кошчанка У. А., Капылоў І. А.* (2009), Актуальны стан і перспектывы развіцця корпуснай лінгвістыкі і камп'ютэрнай лексікаграфіі ў Інстытуце мовы і літаратуры НАН Беларусі. Беларуская мова ў культурнай і моўнай прасторы Славіі: Матэрыялы Міжнароднай навуковай канферэнцыі, Мінск, 24–25 лістапада 2009 г.
5. *Francis W. N., Kucera H.* (1979), Manual of Information to Accompany a Standart Corpus of Present Day. Brown University Providence.
6. *McEnery T., Wilson V.* (1997), Corpus linguistics. Edinburg: Edinburg University.

References

1. *Zaharov V. P., Koval S. A.* (2002), Corpusnaja Lungwistica. NTI. Seria 2 Informationnii processi i sistemi, № 7, s. 34–49. [The Corpus Linguistics. NTI. Series 2 Information processings and systems. № 7, pp. 34–49.
2. *Zubov A. V.* (2005), Corpusnaja lingwistika: vozmojnosti i problemi. Actualnie problemi compjuternoi lingwistiki: Sbornik nauchnih statey. [The Corpus of linguistic: possibilities and problems. The actual of problems of Computational Linguistics: The collection of Scientific Articeles]. Minsk.
3. *Kopotev V., Mustajoki A.* (2008), Sovremennaja corpusnaja rusistika. Instrumentary rusistiki: korpusnie podhodi, Helsinky, s. 7–24. [The contemporary of russistiky: the corpus approaches]. Helsinky, pp. 7–24.
4. *Koshchenko V. A., Kapilov I. A.* (2009), Aktualnoe sostojanie i perspektivi razvitija korpusnoy lingvistiki i komputernoy leksikografii v Institute jazika i literature NAN Belarusi. Belarusskiy jazik v kulturnom i jazikovom prostranstve Slavii: Materialy Mechdunarodnoy nauchnoy konferenzii, Minsk, 24–25 nojabrja, 2009 g. [The contemporary state and perspectives of development of corpus linguistics and computer lexicography in Institute of language and literature National Academy of Science Belarus. The Belarussian language in cultural and language space of Salvia: The material of international beatific conference]. Minsk, November, 24–25, 2009.
5. *Francis W. N., Kucera H.* (1979), Manual of Information to Accompany a Standart Corpus of Present Day. Brown University Providence.
6. *McEnery T., Wilson V.* (1997), Corpus linguistics. Edinburg: Edinburg University.

Зубов Александр Васильевич
Минский государственный лингвистический университет (Беларусь)
Zubov Alexander
Minsk State Linguistic University (Belarus)
E-mail: proscien@mslu.by

М. Копотев, А. Катинская, С. Иванова, Р. Янгарбер
M. Kopotev, A. Katinskaia, S. Ivanova, R. Yangarber

REVITA: ИЗУЧЕНИЕ ЯЗЫКА НА ОСНОВЕ КОРПУСНЫХ ПОДХОДОВ

REVITA: CORPUS-BASED LANGUAGE TEACHING TOOL

Аннотация. Статья посвящена описанию системы Revita, которая создается в Хельсинском университете. Система представляет собой новаторский подход к проведению индивидуальных тестов и индивидуализированных упражнений, для создания которых активно используются корпуса и инструменты автоматического анализа текста. Данные, собранные в процессе использования системы, открывают путь к индивидуальному подходу в изучении языка, к описанию индивидуальной грамматики ученика.
Ключевые слова. корпусные методы в преподавании языка, компьютерные средства обучения, тестирование.

Abstract. This article describes Revita, a system for assisting language learners, being developed at the University of Helsinki. The system employs a novel approach to progress assessment of learners of foreign languages, and uses both corpus data and NLP tools to automatically generate randomized exercises targeting the learner's level of competency. The data collected from L2 learners offers opportunities and challenges in studying individual grammar from both applied and theoretical perspectives.
Keywords. computer-assisted language learning (CALL), intelligent tutoring systems (ITS), NLP tools, corpus-based

0. Изучение языка с помощью компьютера (англ. computer-aided language learning, CALL) было предложено уже в 50-ые годы прошлого века и тех пор существенно продвинулось благодаря стремительному развитию информационных технологий (Hart 1981; Carol&Jamieson 1983). В настоящее время существуют сотни программ для изучения языка — бесплатных и коммерческих, простых и сложных, для начинающих и экспертов. Некоторые программы не просто используют компьютер как средство обучения, но опираются на современные достижения в области корпусной лингвистики и автоматической обработки языка (Nagata 2002, Heift 2001). Такие системы могут быть названы интеллектуальными обучающими системами. Одной из таких систем посвящена настоящая статья.

1. REVITA (revita.cs.helsinki.fi)

Revita является открытой интернет-платформой, предназначенной прежде всего для поддержки исчезающих языков путем включе-

ния учащегося в активное освоение или развития языковых навыков (Katinskaia et al. 2018). Инструмент предназначен прежде всего для людей, которые уже обладают определенной компетенцией в языке, и не подходит для начинающих изучать язык. Модуль русского языка является самым продвинутым, на его примере мы представим всю систему. Модуль состоит из двух частей, тесно взаимосвязанных и создающих основу для контроля и развития индивидуальной грамматики: система проверки языковой компетенции и языковые упражнения.

1.1. Проверка языковой компетенции

Тесты предназначены для контроля прогресса языковых навыков. Первоначально они были разработаны на Отделении современных языков Хельсинкского университета для использования в процессе обучения финских студентов (Мустайоки 2001, Копотев 2010). В проекте Revita они адаптированы к нуждам студентов с любым родным языком. В своей методической части тесты опираются на процессинговую модель в обучении (англ. processing model) и восходят к так называемым тестам прогресса (англ. progress tests), которые давно применяются, например, в учебных программах подготовки врачей. Основные особенности созданного нами теста, которые выделяют его из многочисленных способов тестирования языковых знаний, описаны ниже.

Задания для студентов выбираются из объемной базы данных, содержащей в настоящий момент около 3,5 тыс. заданий по разным темам: от склонения и спряжения до правописания. Перед тестированием преподаватель и/или студент может определить, какое количество заданий и по какой теме следует выбрать; существует возможность контролировать их общее количество, время, отведенное на ответы, и некоторые другие параметры. Выбор конкретных вопросов из базы осуществляется компьютером автоматически. В настоящий момент тест позволяет проверять знания по следующим разделам:

склонение прилагательных,
склонение существительных,
склонение числительных,
спряжение глаголов,
употребление глагольного вида,
глаголы движения,
глагольное управление,
употребление форм глагольных времен,

устойчивые фразы,
порядок слов,
ударение,
лексическая сочетаемость,
синтаксические конструкции,
орфография.

Каждый раздел, в свою очередь, представлен развернутым списком тем. Все задания в тесте представляют собой вопросы с несколькими вариантами ответов — так называемый *multiple choice* — самая распространенная на сегодняшний день форма, применяемая во множестве тестов (Kehoe 1995). Каждая сессия содержит определенное преподавателем количество заданий, на выполнение каждого отводится по умолчанию 15 секунд (последнее сделано для того, чтобы учащийся не имел возможности проверить ответ).

В ходе тестирования задания даются в случайном порядке, однако в конце теста студент получает распределенные по темам и уровням сложности результаты, которые сохраняются в базе данных, что позволяет студенту уже во время второго тестирования получить сведения о прогрессе своих языковых навыков. Результат содержит как обобщенную оценку — процент правильных ответов, так и детальную картину ответов по всем темам, включенным в сессию.

Безусловно, на выполнение теста влияет и темп выполнения заданий, и внимательность, и усталость от теста. Именно поэтому по каждой микротеме предлагается не менее трех вопросов, что дает возможность получить более надежные результаты. Наконец, важно отметить, что тест можно сдать в любом месте в любое удобное время. Конечно, в таких условиях существует возможность для разного рода манипуляций, но, как показала практика, ограничение по времени и ясная мотивировка приводят к тому, что студенты проходят тесты самостоятельно и без подсказок. Одним из главных преимуществ разработанной системы тестов является возможность продолженной оценки уровня знаний. Такой тонкий контроль динамики позволяет не только оценить прогресс в освоении, но и предложить индивидуальный набор заданий, за которые отвечает второй модуль нашей системы.

1.2. Генерация индивидуальных упражнений

1.2.1. Основная идея этой части системы — стимулирование активного использования языка на основе текстов. Под этим мы подразумеваем активное продуцирование языковых форм в контексте, а не пас-

сивное впитывание языковых примеров или правил. Система предназначена для изучения и грамматики, и лексики, и орфографических правил. Упражнения, которые предлагает система, включают тесты на заполнение пропусков, выбор правильного ответа, кроссворды, задания на карточках и т. д. Ключевой особенностью системы является возможность загрузить тексты по выбору и создать собственную библиотеку. Мы считаем, что индивидуальный выбор текстов значительно повышает мотивацию студента, его вовлеченность в процесс обучения. В то же время студенты и учителя могут делиться текстами или использовать заранее подготовленную библиотеку.

По сути, система дает возможность создать собственный корпус текстов разной степени сложности и автоматически обрабатывает его с помощью стандартных инструментов языкового анализа: морфологического, синтаксического и коллокационного. Для этого используются различные инструменты компьютерной обработки текста, основная обработка русских текстов осуществляется с помощью морфологического анализатора Crosslator (Klyshinsky et al. 2011). На основе выполненного анализа система автоматически создает упражнения: карточки, кроссворд, грамматические упражнения и т. д., которые мы опишем ниже.

1.2.2. Первый тип заданий — **грамматические упражнения**. Для создания упражнений для всех токенов в тексте делается автоматический морфологический анализ и лемматизация. После этого система извлекает из текста токены, которые становятся основой для упражнений. Темы для упражнений могут быть заданы учителем или самим учащимся, однако гораздо более полезной является возможность их генерации на основе тех результатов, которые студент получил в ходе выполнения теста. Таким образом студент получает индивидуальный набор упражнений, которые создаются на основе фрагмента текста объемом в 10–30 слов, или 1–5 предложений. В полученном упражнении часть слов убрана и заменена одним из двух типов заданий: выбор ответа из списка или генерация правильной формы на основе показанной начальной формы.

При создании грамматических заданий существует одно существенное ограничение: токены, которые не могут быть однозначно приписаны одной лемме, на данный момент игнорируются: например, форма *жил* омонимична и имеет два морфологических разбора: прошедшее время глагола *жить* и родительный падеж множественного числа существительного *жи́ла*. Такие случаи игнорируются и не вы-

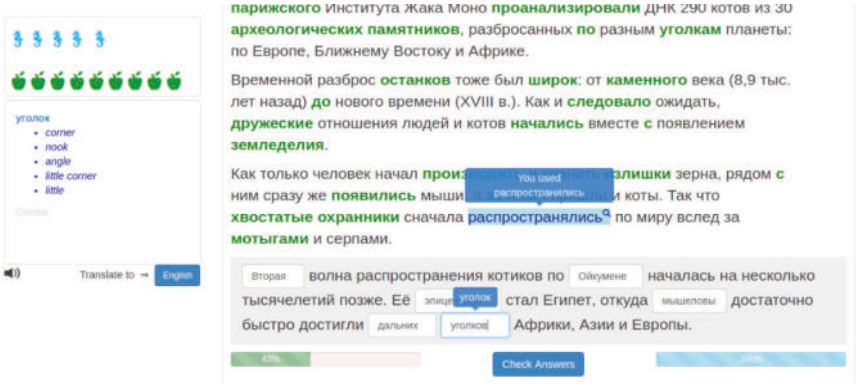


Рис. 1. **Пример упражнения на заполнение пропусков**

бираются для дальнейших упражнений. Система анализирует выполнение упражнений и, основываясь на индивидуальных результатах, предлагает упражнения разной степени сложности, например, выбор правильной формы из предложенных или создание формы при заданной лемме, или создание правильной формы без подсказок. Все ответы, правильные и неправильные, сохраняются в системе. Персональная история учащегося служит основой для выбора заданий для него: например, упражнения, на которые студент никогда не отвечает правильно, отмечаются, но не предлагаются студенту в текущую сессию; упражнения, на которые студент отвечает иногда правильно, иногда нет, получают приоритет и т. д. Для каждого нового задания создаются новые наборы упражнений, так что у студента нет возможности просто вызубрить формы в контексте.

1.2.3. Второй тип — **карточки**, предназначенные для изучения новых слов. Выполняя все виды упражнений, студент может нажать на слово и получить его перевод на выбранный язык. Слова, перевод которых запрашивает ученик, рассматриваются как незнакомые и требующие дополнительного внимания. Все они включаются в набор карточек, связанных текстом, в котором они встретились, а также в набор всех карточек для изучаемого языка. Каждая карточка содержит слово с одной стороны и его сохраненный перевод с другой. Revita позволяет просто листать карточки, а также выполнять упражнения, а именно: ученик должен правильно вставить слово на основе полученного перевода или правильно вставить перевод полученного слова. Пример второго типа упражнений с карточками можно увидеть на Рисунке 2.

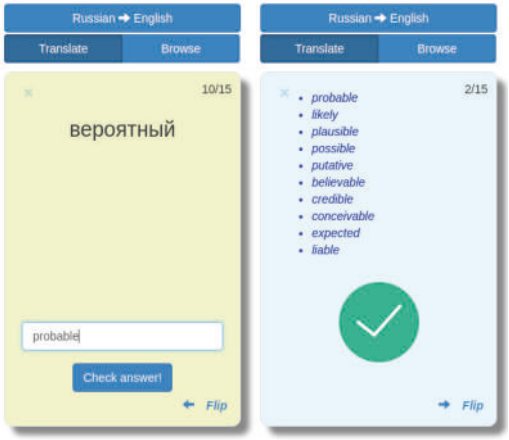


Рис. 2. **Упражнение с карточками**

1.2.4. Третий тип заданий, **кроссворд**, представляет собой лексические упражнения, автоматически создаваемые на основе выбранного текста. Кроссворд состоит из 40–50 слов, выбранных из изучаемого текста. Задача учащегося — вписать в клеточки кроссворда слова в правильной форме. Если учащийся вводит слово правильно, оно добавляется в текст. Если студент испытывает затруднения, система предлагает ему подсказку в виде перевода на известный ему язык. Пример кроссворда можно увидеть на Рис. 3.

Рис. 3. **Кроссворд**

На любом этапе, обычно после изучения темы, учащийся снова проходит тест (целиком или только ту тему, которую он изучил) и получает сведения о своем прогрессе в освоении языка. Результаты представляются в виде диаграмм на специальной странице: чем больше круг на диаграмме, те больше заданий выполнил студент. Чем больше правильных ответов дал студент, тем более ярким становится этот круг.

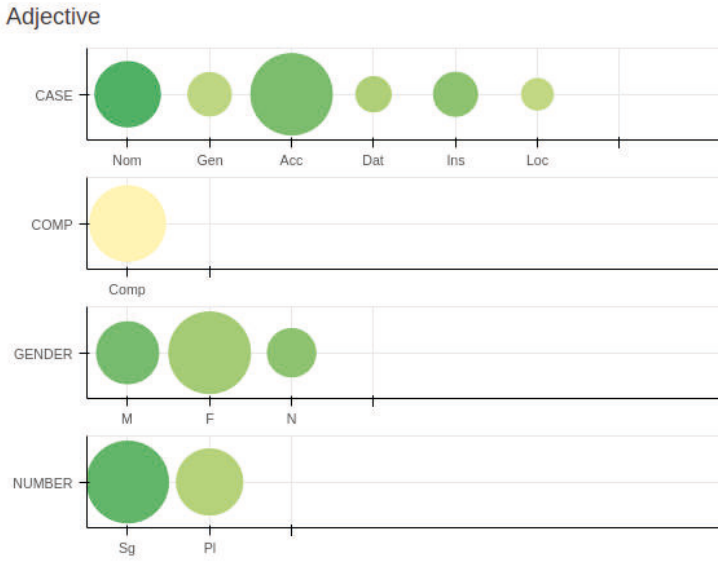


Рис. 4. **Прогресс языковых навыков**

На Рисунке 5 показана общая схема тонкой настройки заданий, исходя из действий конкретного пользователя. После прохождения теста (assessment module) строится модель индивидуальных навыков (student module) с учетом конкретных языковых тем (domain module). Это становится основой для индивидуализированных заданий (exercise module), результаты которых хранятся в истории (student history). На каждом следующем шаге система подстраивается под нужды студента, учитывая его прогресс в выполнении конкретных заданий. Система в принципе работает автономно, однако у преподавателя есть возможность настраивать ее в ручном режиме (instruction module), задавая темы, тексты, параметры тестирования и т. д.

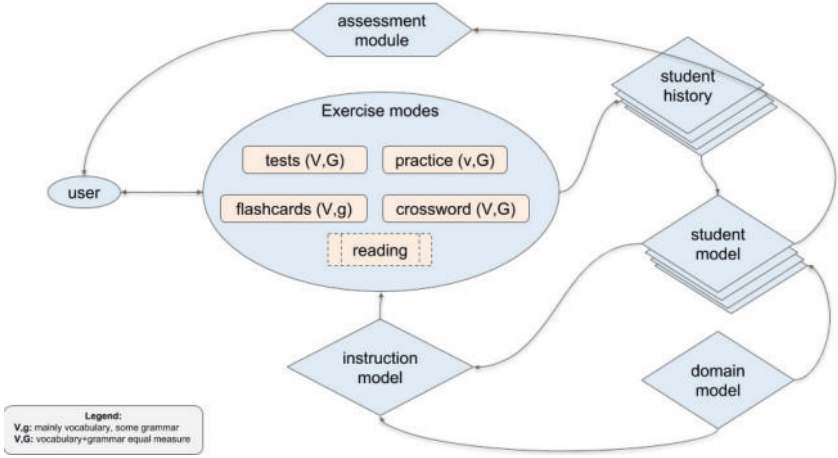


Рис. 5. **Общая схема работы системы**

2. Выводы

На сегодняшний день в лингвистике накоплен солидный опыт по обработке текстов и созданию разнообразных корпусов. Существуют коллекции текстов и инструменты автоматического анализа для множества языков. Следующий шаг состоит в том, чтобы создавать ресурсы, опирающиеся на эти достижения, один из которых представлен в настоящей статье. Система Revita находится в стадии разработки. Однако уже сейчас ясно, что использование достижений компьютерной лингвистики открывает потенциал для изучения и преподавания языка в индивидуальном режиме. Изучающему Revita дает возможность создавать бесконечное количество заданий с учетом собственных потребностей, преподавателю — возможность гибкого динамичного контроля результатов как одного студента, так и всей группы, объединенной общей грамматической темой. Кроме того, постепенное накопление знаний о том, какие темы являются сложными или легкими для всех изучающих язык открывают путь к созданию учебных материалов, исходящих из реальных проблем, с которыми сталкивается студент-иностранец или носитель эритажного языка.

Список литературы

1. *Chapelle, C. and Jamieson, J.* (1983), Language lessons on the PLATO IV system, System, 11(1), pp. 13–20.
2. *Klyshinsky, E. S., Kochetkova, N. A., Litvinov, M. I., Maximov, V. Yu.* (2011), Method of POS disambiguation using information about words cooccurrence (for Russian), Proc. of GSCL, pp. 191–195.
3. *Hart, R.* (1981), Language study and the PLATO system, Studies in Language Learning, 3(1), pp. 1–24.
4. *Heift, T.* (2001), Intelligent language tutoring systems for grammar practice. Zeitschrift fur Interkulturellen Fremdsprachenunterricht, 6(2).
5. *Katinskaia, A, Nouri, J., Yangarber, R.* (2018), Revita: a language learning platform at the intersection of ITS and CALL. 11th edition of the Language Resources and Evaluation Conference, 7–12 May 2018, Miyazaki (Japan).
6. *Kehoe J.* (1995), Writing multiple-choice test items, Practical Assessment, Research & Evaluation. Vol. 4 (9).
7. *Nagata, N.* (2002). Banzai: An application of natural language processing to web-based language learning, CALICO journal, pp. 583–599.
8. *Копотев М. В.* (2010), Система прогрессивного тестирования KARTTU (описание и первые результаты), Русский язык за рубежом, 3, с. 23–29.
9. *Мустайоки А.* (2001), Упражнения по русской грамматике: таксономия и база данных, Русский язык на рубеже тысячелетий, СПб, с. 122–138.

References

1. *Chapelle, C. and Jamieson, J.* (1983), Language lessons on the PLATO IV system, System, 11(1), pp. 13–20.
2. *Hart, R.* (1981), Language study and the PLATO system, Studies in Language Learning, 3(1), pp. 1–24.
3. *Heift, T.* (2001), Intelligent language tutoring systems for grammar practice. Zeitschrift fur Interkulturellen Fremdsprachenunterricht, 6(2).
4. *Katinskaia, A, Nouri, J., Yangarber, R.* (2018), Revita: a language learning platform at the intersection of ITS and CALL. 11th edition of the Language Resources and Evaluation Conference, 7–12 May 2018, Miyazaki (Japan).
5. *Kehoe J.* (1995), Writing multiple-choice test items, Practical Assessment, Research & Evaluation. Vol. 4 (9).
6. *Klyshinsky, E. S., Kochetkova, N. A., Litvinov, M. I., Maximov, V. Yu.* (2011), Method of POS disambiguation using information about words cooccurrence (for Russian), Proc. of GSCL, pp. 91–195.
7. *Kopotev M. V.* Sistema progressivnogo testirovaniia KARTTU (opisanie i pervyje rezul'taty) [The progress test KARTTU: description and first results] // Russkij jazyk za rubezhom. 2010. 3, pp. 23–29.
8. *Mustajoki A.* Uprazhnenija po russkoj grammatike: taksonomija i baza dannyx [Russian grammar exercises: taxonomy and database] // Russkij jazyk na rubezhe tysjacheletij. Saint Peterspurg, 2001. 122–138

9. *Nagata, N.* (2002). Banzai: An application of natural language processing to web-based language learning, CALICO journal, pp. 583–599.

Копотев Михаил
Хельсинкский университет (Финляндия)
Mikhail Kopotev
University of Helsinki (Finland)
mihail.kopotev@helsinki.fi

Анисия Катинская,
Хельсинкский университет (Финляндия)
Anisia Katinskaia
University of Helsinki (Finland)
anisia.katinskaia@cs.helsinki.fi

Сардана Иванова
Хельсинкский университет (Финляндия)
Sardana Ivanova
University of Helsinki (Finland)
sardana.ivanova@helsinki.fi

Роман Янгарбер
Хельсинкский университет (Финляндия)
Roman Yangarber
University of Helsinki (Finland)
roman.yangarber@cs.helsinki.fi

D. Lukeš, M. Kopřivová, Z. Komrsková, P. Poukarová

CREATING A SOCIOLOGICALLY BALANCED SPOKEN CORPUS[1]

Abstract. The article presents the corpora of spoken Czech. They capture private spontaneous dialogues, therefore they were compiled according to the sociological criteria of each speaker. These corpora have been balanced from the beginning in the binary categories of gender, age and highest achieved level of education. Later, dialect regions were added, in which the speaker spent his/her childhood. It is quite difficult to combine these criteria when recording longer interviews. Full balancing of all categories is accomplished in the ORTOFON corpus.
Keywords. spoken corpus, sociological balance, spontaneous spoken language, publicly available data.

1. Introduction

This paper aims to explore data from publicly accessible corpora of spoken Czech built at the Institute of the Czech National Corpus (ICNC) with the emphasis on their sociological balance. The process of spoken corpora creation usually begins with the preparation of the corpus design, i.e. at least the selection of relevant criteria for balancing [cf. Oostdijk 2002]. The sociological balancing includes both the sociological variables and their operationalization. At this point, there is important to mention that this paper concentrates only on the corpora used for linguistic research and not primarily for sociological research. The criteria for balancing mentioned below refers to this "limitation".

The corpora mentioned in this paper provide access to transcripts of prototypical spoken language [Čermák 2009:118], which is defined as informal conversation between well-acquainted parties in a casual setting [Grishina 2006:122–123]). The speakers know each other well and appear in their usual roles, the speech events take place in familiar environments (in private, among friends etc.) and the situations are not experimentally induced. We only record the speech of adult speakers (18+ years old). This type of spoken language covers the cases in which the speaker is minimally self-conscious about the formal attributes of his/her speech. In general, it can be distinguished both from written language as well as from spoken language as used in formal situations, which poses greater demands on speakers in terms of the form and precision of their utterances.

[1] This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

2. Overview of spoken corpora of Czech

2.1. Prague Spoken Corpus (PSC), Brno Spoken Corpus (BSC)

The first Czech spoken corpora were created on the basis of recordings surveying the speech in large cities, where residents from different parts of the Czech Republic (i.e. Moravia, Silesia) are mixed.

The first of them is the Prague Spoken Corpus (PSC) (see [Čermák et al. 2007, 11f.]). It is balanced with respect to four sociolinguistic categories characterizing the speakers and their interactions: sex, highest level of education attained (tertiary × nontertiary), age group (under 35 years old × over 35 years old) and the formality of the situation (informal × formal, i.e. controlled dialogue). The transcription and design criteria pioneered by the PSC formed the basis for several subsequent corpora of spoken Czech (see below).

The same design criteria were also used for the Brno Spoken Corpus (BSC) (see [Hladká 2002]). The third corpus began to be established in Olomouc, but is not yet publicly accessible [Pořízka 2009].

2.2. The ORAL series corpora

The ORAL series corpora include four spoken corpora of Czech which were made according the same design criteria, but there were some changes during the time of the data collection (the first was that it started recording outside of Prague throughout Bohemia [Kopřivová — Waclawičová 2005]).

The first two spoken corpora of the series are the **ORAL2006** [Kopřivová — Waclawičová 2006, Waclawičová 2007] and **ORAL2008** [Waclawičová et al. 2010] corpora. On the contrary to the ORAL2006 corpus, the ORAL2008 corpus is balanced across the binary division of three sociolinguistic categories: age, sex, the highest achieved level of education (see above). The same characteristic for both the corpora is the restriction to the region of Bohemia. The four corpora mentioned up to this point (PSC, BSC, ORAL2006, and ORAL2008) contain only transcriptions. The third corpus, **ORAL2013**, differs in more characteristics than the two previous mentioned corpora of the ORAL series [Válková et al. 2012]. It is also the first to have been collected in all areas of the Czech Republic. The biggest difference is a tone-alignment with the corresponding transcript. This corpus is not balanced in the same sense as the ORAL2008 corpus. During the data collection, the four sociolinguistic categories (sex, age, level of education, and region) were taken into account, because the main idea was to collect the same proportion within the each binar-divided sociolinguistic cate-

gory, i.e. the same number of words from male and female. However, there was not the final step of data selection to keep only the same proportions as it was during creation of the ORAL2008 corpus. The main idea of the ORAL2013 corpus was to publish as much data as possible. The last spoken corpus within this series does not belong properly among the previous because it is not intended to be the follow-up collection to the ORAL2013 corpus. The **ORAL** corpus combines ORAL2006, ORAL2008, ORAL2013 with newly transcribed material (ORAL-Z) into a single conveniently accessible and more richly annotated resource (tagged and lemmatized), about 6 million running words in length.

Table 1. **Available corpora of informal spoken Czech mentioned in this paper**

corpus	Size		time span	region	total hours of audio
	tokens	positions			
PSC	674,992	819,267	1988–1996	Prague	N/A
BSC	500,460	596,009	1994–1999	Brno	N/A
ORAL2006	1,000,798	1,312,282	2002–2006	Bohemia	111
ORAL2008	1,000,097	1,349,536	2002–2007	Bohemia	115
ORAL2013	2,785,189	3,285,508	2008–2011	Bohemia, Moravia, Silesia	292
ORAL	5,368,392	6,361,707	2002–2011	Bohemia, Moravia, Silesia	582
ORTOFON	1,014,786	1,236,508	2012–2017	Bohemia, Moravia, Silesia	103

2.3. The ORTOFON corpus

The most recent spoken corpus of everyday communication has been published in 2017, the data was collected during 2012–2017. The recordings were acquired in Bohemia, Moravia and Silesia [Komrsková et al 2017]. Because one-tier transcription is always a compromise between simplicity and accuracy, multi-tier transcription is used in this corpus: orthographic, which works better in lemmatization and tagging, and phonetic, which better captures the specifics of pronunciation An important innovation is full balancing, which is described in the following section.

3. Balancing the ORTOFON corpus

The previous ORAL2008 and ORAL2013 corpora are balanced according to three sociolinguistic variables: gender, age, and the highest achieved level of education. Each variable was split into two levels (female × male, 18–34 years old × 35+ years old, non-tertiary × tertiary education) to avoid excessive fragmentation and to enable comparability with the PSC.

But what does balancing mean, specifically? In the case of the ORAL2008 and ORAL2013 corpora, it means that the overall proportions of words contained in the corpus, when divided into groups based on the balancing variables, are roughly equal. In other words, roughly half the corpus is by male speakers and half by female, half is by younger speakers and half by older, etc. Crucially though, no steps were taken to ensure that balance is maintained for groups defined by fully crossed variables, i.e. it is not guaranteed that roughly the same amount of material was collected from e.g. 18–34 y.o. tertiary-educated females as from 35+ y.o. non-tertiary-educated males (and so on, with every other possible combination of the factor levels). Indeed, it is not even guaranteed that the specific combination of levels "18–34 y.o. tertiary-educated female" is represented in the corpus at all. This means that in theory, a corpus consisting exclusively of one half 18–34 y.o. tertiary-educated females and one half 35+ y.o. non-tertiary-educated males would be considered balanced by these standards, which does not correspond to an intuitive notion of balance. In practice of course, the ORAL2008 and ORAL2013 are much more diverse than this extreme theoretical case, just not consistently so.

When it came to the ORTOFON corpus as the next iteration of the CNC spoken corpora, the natural evolution was to take on the challenge of the more demanding notion of balancing hinted at above, where each group defined by a different combination of levels of the fully-crossed variables of interest would be represented by roughly the same amount of words. To make it even more challenging, we added a fourth sociolinguistic variable, childhood region of residence, which divides into ten dialect regions. Their borders were refined according to several dialect studies, so they have been slightly modified compared to ORAL2013[2]. While the previous ORAL series corpora only used the criterion of territory to a certain extent to make the data as representative as possible, ORTOFON treats the criterion of childhood region of residence on par with the other balancing variables. The aim was to make the final corpus both representative (i.e. include speakers rep-

[2] The map is available at: https://wiki.korpus.cz/lib/exe/detail.php/cnk:o13.png

resenting all possible combinations of the sociolinguistic variables, and as many different speakers as possible), and as balanced as possible (i.e. make the proportions of all the fully-crossed categories roughly equal).

Considering the target size of the corpus and the number of levels per the four variables, we get $2 \times 2 \times 2 \times 10 = 80$ categories, i.e. 1m tokens $\div$ 80 = 12,500 tokens ideally for each combination of levels, e.g. for 35+ y.o. tertiary-educated female speakers from West Bohemia. We strove for a minimum of five different speakers per combination [9], which reduces the risk of a category being excessively tied to a single idiolect and maintains variability[3].

The implementation of this process starts already during data collection. Once we have assembled a redundant amount of material that covers all target categories (but in unequal proportions), how to come up with a selection algorithm that will pick a subset of the recordings that fulfills these criteria, or at least comes close? Brute force search is not realistic — 2^N different subsets can be chosen from a set of size N. In our case, there were about $N \approx 600$ recordings, which means we would have to rank 2^{600} corpus candidates and select the one with the most balanced representation of the target categories for the target size. For comparison, the number of atoms in the known universe is on the order of 2^{115}, so this is clearly not computationally feasible.

Ideally, we would use an established optimization procedure, which goes about achieving the same result but in a much more clever (and therefore feasible) way. In theory, if the perfect solution existed given our input data, such a procedure should be able to find it. However in practice, this selection problem is so highly constrained that it is in fact very likely to be overconstrained in a real-life setting such as ours.

We ended up opting for a heuristic algorithm which picked the subset of recordings that went into the final ORTOFON corpus using the following steps:

- choose an initial "seed" recording at random
- add another recording to the corpus from the pool of candidates by picking the one which maximally differs from the current composition of the corpus in terms of the relevant sociolinguistic variables[4]
- repeat step 2 until the target size of the corpus (1m words) is reached

[3] More details at http://wiki.korpus.cz/doku.php/cnk:ortofon

[4] The general idea is to compute a diversity score (for each recording) which is higher when the recording adds words in categories that are not well-represented in the corpus so far, and/or adds new speakers. Recordings which would increase the word count of a cat-

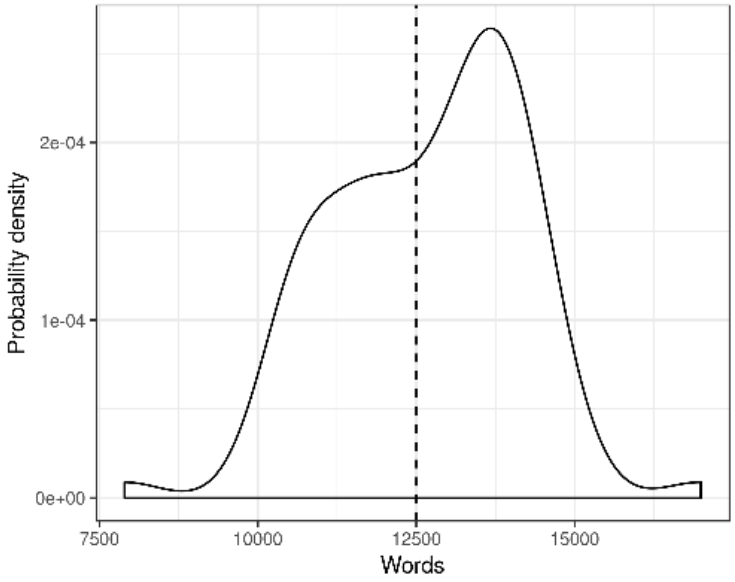


Fig. 1. **Distribution of actual category sizes in the ORTOFON corpus. 12,500 tokens, indicated by the dashed line, is the ideal target category size. As can be seen from the plot, most of the 80 categories in the corpus contain between 10,000 and 15,000 words**

- generate as many variants of the corpus as possible (thousands) by repeating steps 1–3, each time with a different random seed recording
- define an objective function which quantifies how badly any given corpus variant fails at fulfilling the stipulated balancing criteria
- using the objective function from step 5, determine the best / least bad solution from the pool of corpus variants generated in step 3, and select that as the final composition of the corpus

In the case of the first (current) version of the ORTOFON corpus, this procedure yielded a solution which has at least 5 speakers per category in all but 4 of the 80 target categories; in these 4, we have 4 speakers per category, which comes close. The most speakers per category we have is 13. As for word counts, there is not enough space here to provide the full table of word counts per category in the final corpus, and even if there were, it would probably go into unnecessary detail. What is interesting however is

egory over a specified hard limit should be completely ruled out. Given this general idea, the specific definition of the score may then vary depending on circumstances.

to convey a general idea of how well the heuristic algorithm did in selecting a corpus where each of the 80 categories is roughly 12,500 tokens in size. We try to provide such a general idea by plotting the distribution of category sizes in Fig. 1.

4. Conclusion

Another possible notion of balance for a language corpus is having the ratios between the different varieties of the language contained in the corpus correspond to their relative representation in reality. However, this is not only difficult to achieve (the ratios may be hard or even impossible to determine accurately), but also arguably undesirable: it makes underrepresented varieties harder to study, because proportionately less material is collected and made available. At the CNC, spoken corpora have never followed this path, and perhaps more tellingly, recent written corpora have also diverged from it, after careful consideration [Cvrček et al. 2016].

In the case of Czech spoken corpora, the question of balance has always been taken into account and the notion has been supplemented and refined over time. The decisive factor for balancing here is not the relation to reality, but the relationship between predetermined criteria — in this particular case, among the various sociolinguistic factors. As mentioned, the criteria have slightly varied over time: while the two oldest corpora (PSC and BSC) incorporate information on the (in)formality of the recording situation among the balancing criteria, the newest one, ORTOFON, pays attention to regional provenance. The balancing algorithm has also evolved, towards the more sophisticated fully-crossed balancing employed in ORTOFON. However, three categories have remained relevant to balancing (in their simplified binary form) throughout the history of spoken corpora of Czech, thus ensuring a basic level of continuity: gender, age, and highest achieved level of education.

References

1. Balhar, J. et al. (1992): *Český jazykový atlas 1*. Praha: Academia.
2. Cvrček, V., Čermáková, A., Křen, M. (2016). Nová koncepce synchronních korpusů psané češtiny. *Slovo a slovesnost*, 77, 83–101.
3. Čermák, František. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics* 14 (1), 113–23.
4. Grishina, E. (2006). Spoken Russian in the Russian National Corpus (RNC). In *Proceedings of the fifth international conference on language resources and evaluation* (LREC'06). Genova.

5. Hladká, Zdeňka. (2002): Creating Corpuses of the Czech Language at the Faculty of Arts, Masaryk University, Brno. In *The 2nd International Conference on Applied Mathematics and Informatics at Universities.* Trnava: Univerzita Trnava, 115–119.
6. Kopřivová, M., Waclawičová, M. (2005): Construction of spoken corpus based on the material from the language area of Bohemia. In *Computer Treatment of Slavic and East European Languages.* Bratislava: Veda; 137–140.
7. Kopřivová, M., Waclawičová, M. (2006): Representativeness of Spoken Corpora on the Example of the New Spoken Corpora of the Czech Language. In *Trudy mezdunarodnoj konferencii "Korpusnaja lingvistika".* Saint Petersburg: SPGU, 174–181.
8. Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus — Gramatika — Axiologie*, 15, 47–67.
9. Kopřivová M., Goláňová H., Klimešová P., Komrsková Z., Lukeš D. (2014): *Multi-tier Transcription of Informal Spoken Czech: The ORTOFON Corpus Approach.* In *Complex Visibles Out There.* Olomouc: Univerzita Palackého v Olomouci, 529–544.
10. Oostdijk, N. (2002). The Design of the Spoken Dutch Corpus. In Pam Peters — Peter Collins — Adam Smith (eds), *New Frontiers of Corpus Research*. Amsterdam, pages 105–112.
11. Pořízka, P. (2009): Olomouc Corpus of Spoken Czech: characterization and main features of the project. In: *Linguistik online* 38, 2/2009.
12. Válková, L., Waclawičová, M., Křen, M. (2012): Balanced Data Repository of Spontaneous Spoken Czech. In *LREC 2012: Eighth International Conference on Language Resources and Evaluation,* 3345–49.
13. Waclawičová, M. (2007): "Spoken Corpus ORAL2006: Information It Provides and General Characteristics of Spoken Text." In J. Levická and R. Garabík: *Computer Treatment of Slavic and East European Languages: Fourth International Seminar.* Brno: Tribun, 283–289.
14. Waclawičová, M., Křen, M., Válková, L. (2010): Balanced Corpus of Informal Spoken Czech: Compilation, Design and Findings." In *10th Annual Conference of the International Speech Communication Association*, 2009 (INTERSPEECH 2009). Brighton: Curran, 1819–1822.

David Lukeš
Institute of the Czech National Corpus, Charles University, Czech Republic
E-mail: david.lukes@ff.cuni.cz

Marie Kopřivová
Institute of the Czech National Corpus, Charles University, Czech Republic
E-mail: marie.koprivova@ff.cuni.cz

Zuzana Komrsková
Institute of the Czech National Corpus, Charles University, Czech Republic
E-mail: zuzana.komrskova@ff.cuni.cz

Petra Poukarová
Institute of the Czech National Corpus, Charles University, Czech Republic
E-mail: petra.poukarova@ff.cuni.cz

Н. А. Коротаев
N. Korotaev

ПАУЗЫ ХЕЗИТАЦИИ В РАССКАЗЕ И РАЗГОВОРЕ: СОПОСТАВИТЕЛЬНЫЙ КОЛИЧЕСТВЕННЫЙ АНАЛИЗ[1]

HESITATION PAUSES IN NARRATIVES AND CONVERSATIONS: A QUANTITATIVE COMPARISON

Аннотация. На материале корпуса «Рассказы и разговоры о грушах» анализируется использование говорящими заполненных пауз в двух коммуникативных режимах. Рассмотрено речевое поведение восьми участников, каждый из которых на различных этапах записей выступает и как единоличный говорящий, и как один из собеседников в разговоре. Показано, что у всех говорящих при переходе от монологического режима к диалогическому резко понижается доля заполненных пауз относительно общего времени говорения и относительно суммарного количества произнесенных слов.
Ключевые слова. Устная речь, устные корпуса, диалог, монолог, речевые затруднения, частотность, русский язык.

Abstract. In this paper, I analyze the distribution of hesitation pauses (fillers) across two modes of spoken communication. Using data from the "Russian Pear Chats and Stories" corpus, I demonstrate that the frequency of fillers varies dramatically from narration to conversation. Eight speakers were studied, and all of them used less hesitation pauses during the conversation than when retelling a film.
Keywords. Spoken discourse, spoken corpora, dialogue, monologue, disfluencies, word frequency, Russian.

1. Постановка задачи

Объект настоящего исследования — заполненные паузы, т. е. вокальные отрезки, состоящие из тех или иных «долексических» звуков: эканья, мэканья и проч. (см. [Кибрик, Подлесская (ред.) 2009: 67–72]). Такие единицы служат специализированными маркерами хезитации, посредством которых говорящие сигнализируют о том, что, с одной стороны они не готовы прямо сейчас приступить к реализации своего коммуникативного замысла или продолжить ее, с другой стороны, все-таки намерены осуществить вербализацию и, соответственно, работают над возникшей проблемой ([Clark, Fox Tree 2002: 81–92]).

Хезитация — одно из базовых явлений устной речи, и это проявляется в том числе и в частотности заполненных пауз. Так, в устном подкорпусе Британского национального корпуса (BNC) единицы *er и erm*

[1] Исследование поддержано грантом РФФИ № 19-012-00626.

занимают соответственно 18-е и 29-е места в частотном списке лемм, а их суммарный ipm превышает 14 500; в в устной части Американского национального корпуса позиции единиц *uh и um* еще выше: соответственно 10-е и 28-е места с суммарным ipm более 27 000[2]. Предварительные данные для русского языка приводятся в [Шерстинова 2016]: в 230-тысячной выборке из корпуса «Один речевой день» (ОРД) единица (*э*) занимает 21-е место в частотном списке словоформ, с ipm равным 7 350.

Задача данного исследования — проанализировать, как говорящие на русском языке используют паузы хезитации в двух противопоставленных режимах коммуникации: монологическом рассказе и разговоре. Принципиальный исследовательский вопрос заключается в том, как коммуникативный режим влияет на частотность и характер использования заполненных пауз.

2. Материал и метод исследования

Хотя заполненные паузы, как уже было отмечено, входят в число наиболее частотных единиц устной речи, их разметка не всегда проводится последовательно. Например, в устном подкорпусе НКРЯ обнаруживаются лексемы *ээ*, *аа* и др., но их частотность совсем не велика — и это связано с тем, что паузы хезитации размечены далеко не на всем массиве текстов. Поэтому для решения нашей задачи удобно обратиться к данным «малых» устных коллекций, снабженных подробной дискурсивной разметкой. Одной из таких коллекций является корпус «Рассказы и разговоры о грушах» (http://multidiscourse.ru), содержащий набор однотипных коммуникативных эпизодов. Общая продолжительность всех записей корпуса составляет около 15 часов, суммарное число словоупотреблений — около 160 тысяч. Для настоящей работы был проанализирован подкорпус из четырех записей общей продолжительностью около 1,5 часа и объемом немногим более 15 тысяч словоупотреблений. Выбор подкорпуса обусловлен состоянием разметки: для этих записей уже реализована подробная вокальная аннотация, подразумевающая, помимо прочего, установление точных временны́х границ всех «обычных» слов, заполненных и абсолютных пауз, а также ряда дополнительных вокальных действий. Подробнее о принципах разметки см. [Коротаев 2019]; здесь достаточно

[2] По данным Sketch Engine (https://app.sketchengine.eu), дата обращения 11.03.2019.

отметить, что временна́я привязка осуществляется на полуавтоматической основе, а ее результат хранится в файлах формата textgrid, используемом в программе акустического анализа Praat (http://www.fon.hum.uva.nl/praat/).

С поставленной задачей согласуется и общий характер входящих в корпус записей. В каждой из них принимают участие четыре человека с фиксированными ролями. Двое участников (Рассказчик и Комментатор) до начала записи смотрят «Фильм о грушах» — хорошо известный стимульный материал, разработанный в 1970-е годы группой под руководством У. Чейфа (http://www.linguistics.ucsb.edu/fa-culty/chafe/pearfilm.htm). На первом этапе Рассказчик в монологическом режиме рассказывает содержание фильма третьему участнику — Пересказчику, который фильм не видел. На втором этапе Комментатор уточняет рассказ, а Пересказчик задает вопросы обоим своим собеседникам. Этот этап носит преимущественно диалогический характер и может рассматриваться как достаточно обычный разговор. На третьем этапе в комнате появляется Слушатель, который также не видел фильм. Пересказчик пересказывает ему содержание фильма, ориентируясь на то, что узнал во время первых двух этапов. В конце Слушатель должен представить свое понимание услышанного в письменной форме. Подробнее см., в частности, [Кибрик 2018].

Несложно заметить, что в каждой записи есть два участника, реализующих как монологический, так и диалогический дискурс: Рассказчик и Пересказчик. Это позволяет анализировать речевые действия одних и тех же говорящих в разных коммуникативных режимах, т. е. выполнять контролируемое сопоставление. Таким образом, непосредственным материалом исследования послужили вокальные действия Рассказчиков на этапах рассказа и разговора и вокальные действия Пересказчиков на этапах пересказа и разговора. Для каждого из рассмотренных восьми говорящих были подсчитаны количество и суммарная продолжительность заполненных пауз, после чего эти числа были приведены к общему числу слов и продолжительности вокализации на нужных этапах.

3. Результаты и обсуждение

Результаты проведенных подсчетов представлены на рис. 1. Прежде чем переходить к обсуждению полученных данных, необходимо сделать следующие оговорки.

Говорящий	Число заполненных пауз				Доля заполненных пауз относительно времени вокализации	
	На 100 слов		На 100 секунд			
	Mono	Conv	Mono	Conv	Mono	Conv
04N	9.71	3.53	24.43	9.98	0.067	0.027
04R	10.54	3.54	25.84	10.34	0.109	0.028
16N	11.78	5.05	31.04	14.65	0.085	0.044
16R	9.91	4.54	23.39	11.04	0.120	0.060
22N	4.44	1.19	12.29	3.81	0.038	0.009
22R	8.79	4.21	23.87	8.57	0.069	0.025
23N	11.35	2.83	29.00	9.78	0.105	0.020
23R	5.22	0.69	15.98	2.88	0.045	0.006
Суммарно по 8 говорящим	***9.00***	***3.63***	***23.54***	***10.43***	***0.082***	***0.036***

Рис. 1. **Меры частотности и временны́е доли заполненных пауз в речи Рассказчиков и Пересказчиков размеченного подкорпуса в (пере)рассказе (Mono) и в разговоре (Conv)**

- При разметке корпуса выделяются следующие типы заполненных пауз: эканье (обозначается как (*ə*)), аканье ((*ɐ*)), мэканье ((*ш*)), гортанный скрип ((*ʔ*)), а также различные комбинации этих вариантов ((*ɐш*), (*шə*) и т. д.). Содержательные различия между способами заполнения составляют отдельный вопрос, нуждающийся в дополнительном изучении (см., в частности, подробный анализ противопоставления английских *uh и um* в [Clark, Fox Tree 2002]). В рамках данного исследования все типы заполненных пауз рассматриваются как представители одного класса.
- Вопрос о принадлежности заполненных пауз к словам также является дискуссионным. Здесь принято техническое решение: к словам, помимо «безусловных» лексических единиц, отнесены

также оборванные фрагменты, заполненные паузы и смех (если он не накладывается на произнесение других сегментных единиц).

- При оценке частотности заполненных пауз используется число вхождений на 100 слов. Это обусловлено малым общим объемом материала: использовать для него более стандартную меру ipm было бы опрометчиво. При этом, разумеется, в высшей степени условный ipm можно получить из использованной меры путем умножения на 10 000.
- При подсчете суммарного времени вокализации не учитывались паузы молчания, расположенные между элементарными дискурсивными единицами (ЭДЕ) или другими единицами верхнего уровня сегментации (см. [Кибрик, Подлесская (ред.) 2009: 55–72; Коротаев 2019: 11–13]). В то же время паузы молчания внутри ЭДЕ включались в общее время вокализации.

Перейдем к непосредственному анализу полученных результатов. Как следует из рис. 1, для всех восьми говорящих наблюдаются следующие тенденции.

1) В рассказе / пересказе частотность заполненных пауз существенно выше, чем в разговоре. Наиболее ярко эта тенденция проявляется в речи Пересказчика в записи 23, наименее ярко — в речи Пересказчицы в записи 22, но и у этой говорящей в пересказе заполненные паузы встречаются в два раза чаще, чем в разговоре. (Отметим, что даже в разговоре частотность заполненных пауз все равно очень высока: 36 300 по всем говорящим в пересчете на ipm. В целом же заполненные паузы встречаются в исследованном подкорпусе с частотой 62 500 ipm.)
2) При альтернативной оценке частотности (число вхождений на 100 секунд вокализации) картина в целом остается неизменной: в монологической речи паузы хезитации появляются с очевидно бóльшей частотой, чем в разговоре. Наблюдается и индивидуальное варьирование, но, за исключением Пересказчика в записи 23, границы этого варьирования не слишком широки (см. длины закрашенных областей в столбце Conv на рис. 1).
3) В эту же сторону различаются и доли произнесения заполненных пауз относительного общего времени вокализации. Суммарно в монологических этапах на долю заполненных пауз приходится более 8 % всей вокализации, а на этапе разговора — ме-

нее 4%. Индивидуальное варьирование тут несколько более заметно, чем в предыдущем параметре, но все равно по крайней мере почти двукратное различие наблюдается у всех говорящих.

Таким образом, все использованные оценки дают согласованный результат: в рассказе и пересказе говорящие существенно чаще прибегают к паузам хезитации, чем в разговоре. Напрашивающаяся интерпретация этого результата такова: реализация монологического дискурса, в котором говорящий несет почти единоличную ответственность за успешность коммуникации, требует бóльших когнитивных усилий и в целом более сложна, чем участие в интерактивном диалоге. Этот вывод, как кажется, согласуется с результатами, полученными на другом материале. Так, согласно [Kilgarriff 1995], в той части устного подкорпуса BNC, которая основана на записи разговоров (demographic part), единицы *er u erm* примерно в два раза менее частотны, чем в части, содержащей существенную долю более трудных для исполнения монологических жанров (context-governed part). В работе [Шерстинова 2016], в свою очередь, было показано, что единица (*э*) наиболее частотна в образовательной коммуникации и наименее частотна в бытовых разговорах.

Возможно и альтернативное объяснение: в разговоре говорящие могут стремиться к тому, чтобы реже сигнализировать о своих затруднениях, и таким образом понижать риск потери хода. Иными словами, влияние на частотность хезитационных пауз может оказывать как кооперативный, так и, напротив, конкурентный характер диалога. Для проверки этого предположения необходимо расширить эмпирическую базу исследования, включив в него средства хезитации, отличные от заполненных пауз, а также другие типы речевых сбоев.

Литература

1. *Кибрик А. А.* (2018), Русский мультимодальный дискурс. Часть II. Разработка корпуса и направления исследований. Психологический журнал 39(1), с. 70–80.
2. *Кибрик А. А., Подлесская В. И.* (ред.) (2009), Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
3. *Коротаев Н. А.* (2019), «Рассказы и разговоры о грушах»: принципы вокальной аннотации. Версия 10.01.2019. URL http://multidiscourse.ru/data/ann/pears_vocal_annotation.pdf (Дата обращения 11.03.2019)
4. *Шерстинова Т. Ю.* (2016), Наиболее употребительные слова повседневной русской речи (в гендерном аспекте и в зависимости от условий коммуникации). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2016», с. 616–631.

5. *Clark G., Fox Tree J.* (2002), Using *uh* and *um* in spontaneous speaking, Cognition, 84, pp. 73–111.
6. *Kilgarriff A.* (1995), BNC database and word frequency lists, URL https://www.kilgarriff.co.uk/bnc-readme.html (Дата обращения 11.03.2019)

References

1. *Clark G., Fox Tree J.* (2002), Using *uh* and *um* in spontaneous speaking, Cognition, 84, pp. 73–111.
2. *Kibrik A. A.* (2018), Russkij mul'tikanal'nyj diskurs. Čast' II. Razabotka korpusa i napravlenija issledovanij [Russian multichannel discourse. Part II. Corpus development and avenues of research], Psixologičeskij žurnal [Psychological Journal], 39(1), pp. 70–80.
3. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009), Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: JaSK.
4. *Kilgarriff A.* (1995), BNC database and word frequency lists, available at: https://www.kilgarriff.co.uk/bnc-readme.html (Requested 11.03.2019)
5. *Korotaev N.* (2019), "Russian Pear Chats and Stories": Vocal annotation guide. Version 10.01.2019. Available at: http://multidiscourse.ru/data/ann/pears_vocal_annotation_en.pdf (Requested 11.03.2019)
6. *Sherstinova T.* (2016), Naibolee upotrebiteľnye slova povsednevnoj russkoj reči (v gendernom aspekte i v zavisimosti ot uslovij kommunikacii) [The most frequent words in everyday spoken Russian (in the gender dimension and depending on communication setting)], Komp'juternaja lingvistika i intellektual'nye texnologii: Trudy meždunarodonoj konferencii «Dialog-2016» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog-2016»], pp. 616–631.

Коротаев Николай Алексеевич
Российский государственный гуманитарный университет (Россия)

Korotaev Nikolay
Russian State University for the Humanities (Russia)
E-mail: n_korotaev@hotmail.com

А. М. Лаврентьев, Ф. Н. Соловьев, А. М. Чеповский
A. M. Lavrentiev, F. N. Solov'ev, A. M. Chepovsky

ВНЕДРЕНИЕ В TXM ДОПОЛНИТЕЛЬНЫХ ИНСТРУМЕНТОВ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА[1]

IMPLEMENTATION IN THE TXM PLATFORM OF ADDITIONAL INSTRUMENTS OF AUTOMATIC TEXT PROCESSING

Аннотация. В докладе представлен опыт расширения возможностей платформы TXM за счет инструментов автоматической обработки текста (выделение псевдооснов, именных групп, анализ глагольного управления). В сочетании со стандартными функциями TXM (факторный анализ соответствий, специфичность и т.д.) они позволяют более эффективно осуществлять анализ специализированных корпусов, нацеленных, в частности, на выявление противоправного дискурса.
Ключевые слова. автоматический анализ текстов, платформа TXM, псевдоосновы, именные группы, глагольное управление.

Abstract. This paper presents an experience of extending the capacities of the TXM platform by adding tools of automatic text processing (allocation of pseudo-bases by stemming technique that uses a word structural pattern method, noun phrases, the analysis of verbal dependencies). Combined with the standard TXM functions (the factorial correspondence analysis, specificity, etc.) they allow the users to improve the performance of analysis of specialized corpora, such as those aimed at the detection of unlawful discourse.
Keywords. automated text analysis, TXM platform, stemming, noun phrases, verbal dependencies.

Введение

В настоящей работе мы опираемся на программный комплекс — платформу TXM (http://textometrie.org). Платформа TXM является эффективным средством корпусного анализа, позволяющим проводить комплексный анализ корпусов (анализ соответствий, кластеризация, построение лексических таблиц, поиск сложных лексических конструкций, выделение подкорпусов по различным параметрам). Платформа TXM интегрирована с расширением TreeTagger [Schmid 1994], позволяющим проводить лишь морфологический анализ и лемматизацию словоупотреблений. Она использует словоупотребления в качестве структурных единиц анализа.

Для повышения эффективности таких используемых TXM методов, как анализ специфичности и анализ соответствий, целесообразно ввести в рассмотрение новые единицы анализа, опирающиеся на

[1] Работа выполнена при финансовой поддержке РФФИ в рамках научных проектов № 16-29-09546, № 18-00-00606(18-00-00233) и № 19-07-00806.

процедуры автоматизированной обработки текстов на естественных языках, описанные в [Чеповский 2015].

Мы предлагаем ряд расширений, позволяющих дополнить и усложнить анализ корпусов, включающий: автоматический морфологический анализ словоформ и приведение их к канонической форме, выделение псевдооснов, выделение именных и глагольных групп и комбинирование результатов работы предлагаемых расширений. Конечной целью дополнений к платформе ТХМ является создание механизмов для исследования применимости различных дифференцирующих признаков при решении задачи классификации текстов и создания тематических корпусов текстов.

В [Лаврентьев и др. 2018] мы провели эксперименты по использованию псевдооснов и именных групп для выявления экстремистской направленности текстов. В данной работе к этим характеристикам добавлены возможности учета глагольного управления.

Псевдоосновы

Для определения дифференцирующих признаков коротких текстов сети интернет, характеризующимися особыми тематическими и психолингвистическими свойствами, текстов, содержащих неологизмы и жаргонизмы, большой интерес представляет использование аналитического метода выделения псевдооснов, так как он позволяет обрабатывать отсутствующие в стандартных словарях формы.

Используемый способ выделения псевдооснов представляет собой метод структурных схем, описанный подробно в [Egorova и др. 2016]. Суть метода состоит в получении псевдоосновы словоформы путем рассмотрения и отбрасывания ее словоизменительных аффиксов. Словообразовательные аффиксы считаются в рамках этого метода элементом корневой части и не отбрасываются. Далее под аффиксами мы будем понимать исключительно словоизменительные аффиксы. Каждому слову можно сопоставить отвечающую ему последовательность аффиксов. Такие последовательности называются структурами некорневой части слова. Отсюда происходит название метода. Как и в традиционном морфологическом анализе, аффиксы подразделяются на префиксы и суффиксы в соответствии с их позицией относительно корня слова. Псевдоосновой называется часть слова, не содержащая суффиксов и префиксов. Способ автоматического выделения псевдооснов состоит в сопоставлении рассматриваемой словоформы

с множеством допустимых в языке структур некорневой части слова. Псевдооснова слова выделяется отбрасыванием всех соответствующих определенной структурной схеме аффиксов (то есть допустимой в данном языке максимальной комбинации префиксов и суффиксов). У глаголов, в частности, отбрасываются показатели лица, числа, рода, времени, причастной формы. Видовые префиксы не отбрасываются, так как они могут влиять на лексическое значение слова.

Псевдооснова не всегда совпадает с основой слова в традиционном понимании. Например, в словоформе *людьми* единственным аффиксом, который можно отбросить согласно продуктивной структурной схеме, является *–и*, поэтому выделяется псевдооснова *людьм.*

Данный подход позволяет анализировать текстовые конструкции, опираясь не только на точные словоформы и тем самым повышает полноту и гибкость корпусного анализа.

Морфологические характеристики

Возможность привести словоформу к канонической форме позволяет анализировать различные элементы словоизменительной парадигмы как одну и ту же структурную единицу текста. Это, в свою очередь, позволяет более корректно проводить содержательный статистический анализ текста, например, путем рассмотрения частот лексем вместо частот отдельных словоформ.

При предобработке всех русскоязычных текстов мы осуществляем автоматический морфологический анализ словоформ на основе словарной компьютерной морфологии, описанной в [Чеповский 2015]. Используемая стандартная в отественной компьютерной лингвистике морфологическая модель относит каждое слово к одному из 24 морфологических классов, включающих, помимо частей речи в традиционном понимании, такие разряды, как «неизменяемое слово», «аббревиатура», «топоним». Каждый из этих морфологических классов характеризуется набором грамматических характеристик: род, падеж, число, наклонение и др. В программной реализации словарной морфологии русского языка применяется специализированная структура данных, позволяющая осуществлять поиск словоформ за линейное по числу букв словоформы время. Каждая словоформа содержит свои грамматические характеристики и её каноническую (начальную) форму.

В настоящей работе мы также использовали интегрированный в TXM программный пакет TreeTagger [Schmid 1994], предоставляю-

щий возможность совместного морфологического анализа слов предложения на основе статистической модели, путем сопоставления словоупотреблений, снабжённых специальными метками, кодирующими морфологические характеристики. Преимуществом данной процедуры разметки является однозначность морфологического анализа, но при таком анализе существует риск ошибок, который возрастает, если текст содержит большое количество неологизмов и нестандартных написаний слов. Все виды морфологической разметки использовались в дальнейшем для сопоставительного анализа текстов корпуса.

Выделяемые из текста конструкции

Дополнительную информацию о специфическом содержании текста можно почерпнуть, анализируя не только словоформы, но и целые именные группы. Именная группа определяется нами как группа слов, у которой главное слово существительное, а другие слова связаны с ним подчинительными синтаксическими связями. Рассмотрение частотных именных групп и их сочетаний, в совокупности с анализом отдельных словоупотреблений позволяет получить более полную картину семантических и стилистических характеристик текста, релевантных для его содержания.

Определенную сложность при выделении именных групп представляет множественность морфологических разборов при омонимии. В ходе анализа слов в предложении наш метод предполагает рассмотрение всего множества возможных морфологических разборов каждого слова.

Используемый нами алгоритм подробно описан в [Чеповский 2015] и анализирует предложения русского языка в три этапа: 1) установление подчинительных синтаксических связей в предложении между парами слов; 2) установление синтаксических связей внутри конструкций с однородными членами; 3) выделение именных групп как цепочки последовательно связанных подчинительными связями слов.

Выделение глагольных групп (словосочетаний, главным словом которых является глагол), установление связей выделенных именных групп с глаголами представляет важную, необходимую составляющую синтаксического анализа предложения. Данные задачи решаются анализом глагольного управления в рамках коммуникативной грамматики. В рамках нашей работы был использован электронный словарь глагольного управления, в который вошли первые две тысячи наибо-

лее частотных глаголов русского языка по материалам Национального корпуса русского языка (ruscorpora.ru).

Словарь глагольного управления содержит набор ограничений, сопоставленных глаголу, и образует парадигму глагольного управления для данного глагола. Глагольным управлением является языковое явление, состоящее в проистекающих из семантики глагола требованиях, накладываемых последним на зависимые от него слова. Именно эти требования мы формализуем в виде указанных ограничений.

Результаты морфологического анализа и процедуры выделения именных групп позволяют, используя словарь глагольного управления, выявить синтаксические связи для определения глагольных групп. Выделение глагольных групп в предложении осуществляется путем анализа всех возможных пар (глагол, именная группа) предложения на предмет соответствия именной группы парадигме управления соответствующего глагола и принятия решения о наличии управления именной группы глаголом.

Анализ подкорпусов

Удобным инструментом количественной оценки «необычности» специального подкорпуса относительно всего корпуса является показатель специфичности [Lafon 1980]. Анализ специфичности позволяет составить своего рода «профиль» подкорпуса, выделенного на каких-либо внешних основаниях (например, автор, жанр, тематика или идеологическая направленность текста) путем выявления наиболее характерных или нехарактерных для него словоформ (лексем, псевдооснов, именных и глагольных групп и т. п.). Этот «профиль» может быть использован для диагностики нового текста.

Другим подходом к анализу разделенного на части (подкорпуса) по определенному критерию корпуса является анализ соответствий. Методика анализа соответствий, используемая ТХМ, была предложена Ж.-П. Бензекри [Benzecri 1979] и имплементирована в пакете FactoMineR для платформы R [Lê et al. 2008]. Анализ соответствий демонстрирует взаимную «близость» или «удаленность» подкорпусов на основе анализа частот совместного появления значений переменных (словоформ, начальных форм, псевдооснов, именных групп, морфологических тегов и т. д.).

Экспериментальный корпус был проанализирован с использованием двух обозначенных выше функций ТХМ — специфичность

и анализ соответствий. Детально были рассмотрены следующие лексические объекты: словоформы; начальные формы слов, полученные по словарной морфологии; начальные формы слов с морфологическими характеристиками, полученные с помощью TreeTagger; псевдоосновы слов; именные группы, составленные из словоформ; именные группы, составленные из начальных форм; именные группы, составленные из псевдооснов вместо отдельных словоупотреблений; глагольные группы.

Заключение

Проведенная работа по интеграции инструментов автоматической обработки текста и платформы корпусного анализа TXM показал, что такая интеграция позволяет расширить возможности статистического анализа текстов.

Детально были рассмотрены такие лексические объекты, как леммы, псевдоосновы, именные и глагольные группы различной структуры. Упомянутые средства были объединены в набор утилит, позволяющих вычислять для текстовых корпусов ряд характеристик языковых единиц, входящих в их состав. Корпуса с вычисленными характеристиками преобразуются нами в формат для импорта пакетом TXM.

Показано, что при делении текстов на подкорпуса, есть возможность интерпретировать близость, или разделенность значений рассматриваемых характеристик подкорпусов относительно друг друга как оценку, указывающую на сходство или различие маркированных подкорпусов между собой и по отношению к «нейтральному» подкорпусу.

В силу выявленных особенностей и противопоставленности нейтрального подкорпуса остальным, сформированный корпус может быть использован для машинного обучения в задачах классификации текстов на предмет выявления заданного содержания с целью их углубленного экспертного анализа.

В ходе дальнейших исследований мы планируем провести широкие исследования влияния различных дифференцирующих признаков и их комбинаций для формирования специализирующих подкорпусов текстов, наборов применяемых методов статистического и качественного анализа корпусов, формирования обучающих выборок и решения задач классификации текстовых массивов.

Литература

1. *Лаврентьев А. М., Смирнов И. В., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.* (2018), Создание специальных корпусов текстов на основе расширенной платформы TXM, Системы высокой доступности, 14 (3), с. 76–81.
2. *Чеповский А. М.* (2015), Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.
3. *Benzécri J.-P.* (1979), L'analyse des données: l'analyse des Correspondances. 2nd ed., vol. 2. Paris.
4. *Egorova E., Chepovskiy A., Lavrentiev A.* (2016), A structural pattern based method for automated morphological analysis of word forms in a natural language, Journal of Mathematical Sciences, 214 (6), pp. 802–813.
5. *Lafon P.* (1980), Sur la variabilité de la fréquence des formes dans un corpus, Mots, 1, pp. 127–165.
6. *Lê S., Josse J., & Husson F.* (2008), FactoMineR: an R package for multivariate analysis, Journal of statistical software, 25 (1), pp. 1–18.
7. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, available at http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf

References

1. *Benzécri J.-P.* (1979), L'analyse des données: l'analyse des Correspondances. [Data Analysis: Correspondence Analysis] 2nd ed., vol. 2. Paris.
2. *Chepovskiy A. M.* (2015), Informatsionnye modeli v zadachakh obrabotki tekstov na estestvennykh yazykakh [Information models in natural language text processing problems], 2nd ed. Moscow.
3. *Egorova E., Chepovskiy A., Lavrentiev A.* (2016), A structural pattern based method for automated morphological analysis of word forms in a natural language, Journal of Mathematical Sciences, 214 (6), pp. 802-813.
4. *Lavrentiev A. M., Smirnov I. V., Solov'ev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.* (2018), Sozdanie specialnikh corpusov tekstov na osnove rasshirennoy platformi **TXM** [Creating text corpora for special purposes on the basis of extended TXM platform], Systemi Visokoy Dostupnosti [Highly Available Systems], 14 (3), pp. 76–81.
5. *Lafon P.* (1980), Sur la variabilité de la fréquence des formes dans un corpus [On the Variability of Word-Form Frequencies in a Corpus], Mots, 1, pp. 127–165.
6. *Lê S., Josse J., & Husson F.* (2008), FactoMineR: an R package for multivariate analysis, Journal of statistical software, 25 (1), pp. 1–18.
7. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, available at http://www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf

Лаврентьев Алексей Михайлович
Институт истории представлений и идей нового времени НЦНИ и Высшей нормальной школы Лиона (Франция)
A. M. Lavrentiev
IHRIM Research Lab, CNRS & ENS de Lyon (France)
E-mail: alexei.lavrentev@ens-lyon.fr

Соловьев Федор Николаевич
Институт физико-технической информатики (Россия)
F. N. Solov'ev
Institute of Physical and Technical Informatics (Russia)
E-mail: the0@yandex.ru

Чеповский Андрей Михайлович
Национальный исследовательский университет «Высшая школа экономики» (Россия)
A. M. Chepovskiy
National Research University «Higher School of Economics» (Russia)
E-mail: achepovskiy@hse.ru

А. Н. Лапошина, Т. С. Веселовская, О. Ф. Купрещенко
A. N. Laposhina, T. S. Veselovskaya, O. F. Kupreshchenko

ИЛЛЮСТРАТИВНО-ТЕКСТОВЫЙ КОРПУС УЧЕБНИКОВ РУССКОГО ЯЗЫКА ДЛЯ ДЕТЕЙ МЛАДШЕГО ШКОЛЬНОГО ВОЗРАСТА: КОНЦЕПЦИЯ И МЕТОДИКА СОЗДАНИЯ[1]

TEXT-IMAGE CORPUS OF RUSSIAN LANGUAGE TEXTBOOKS FOR PRIMARY SCHOOL: CONCEPT AND METHOD OF CREATION

Аннотация: В статье описаны цели и методика создания корпуса учебников русского языка для детей младшего школьного возраста, а также приведены первые результаты исследований на этом материале. Разработана методика разметки, которая учитывает как текстовый, так и визуальный компонент обучающих материалов, а также особенности их взаимосвязи. Такая разметка позволяет решать ряд исследовательских задач: анализ сложности учебных текстов, соответствие лексического состава учебников возрасту и интересам современных школьников, влияние иллюстраций и элементов верстки на восприятие и понимание учебного материала.
Ключевые слова: русский язык, учебник русского языка, корпус учебников, иллюстративно-текстовый корпус, учебные тексты.

Abstract: This paper presents the goals and methods of creating a corpus of Russian language textbooks for primary school children. The first results of research on this data are also given. Corpus annotation included both the textual and visual components of the educational materials, as well as the peculiarities of their interlinkages. This annotation allows us to solve a number of research problems: analysis of the texts complexity in a coursebooks, correspondence between the textbook`s lexic and the age and interests of modern schoolchildren, the role of illustrations and page design in the text perception and understanding.
Keywords: Russian language, Russian language textbook, textbook corpus, text-image corpus, educational text.

Введение

Системное знакомство русских детей с родным языком и речью начинается с учебника по русскому языку. Кроме того, освоение русского языка в начальной школе является необходимой базой для успешной работы с учебными текстами по другим предметам. Эти факты не оставляют сомнений в необходимости тщательного анализа содержания учебников.

Одним из самых эффективных способов получения объективных данных о текстовом наполнении учебников является корпусной анализ. В настоящее время этот метод широко используются в исследовании и создании учебных материалов [Boulton 2017; McEnery et al.

[1] Работа выполнена при финансовой поддержке РФФИ, проект 17-29-09156

2010]. Среди примеров коллекций учебников с целью сравнительного анализа стоит отметить корпус учебников английского языка [Islam 2014], японского языка [Sato 2008], сравнение методических школ Южной и Северной Кореи на материале учебников английского для детей этих стран [Kim 2017] — все они исследуют учебники английского как иностранного. Однако внимание корпусных исследователей ещё не было направлено на учебные материалы по русскому языку для детей младшего школьного возраста.

В настоящей работе представлена концепция создания и методика разметки корпуса школьных учебников родной речи для комплексного сравнительного анализа учебных материалов и приведены примеры исследований, ставших возможными благодаря собранному материалу.

1. Концепция иллюстративно-текстового корпуса

Современная коммуникация характеризуется переходом от одномерных вербальных текстов к мультимодальным, смысл которых складывается из знаков различных семиотических систем (визуальные образы, дизайн, вербальные компоненты письменного языка и другие семиотические средства) [Kress, van Leeuwen 1996]. Эти изменения коснулись и учебной литературы, что подтверждается большим количеством исследований о влиянии иллюстраций на восприятие учебных текстов [Carney, Levin 2002; Jewitt 2008; Guo et al. 2018; Schneider et al 2018]. Таким образом, при анализе школьных учебников важно учитывать не только текстовую, но и визуальную составляющую материалов: могут ли иллюстрации изменить восприятие текста, повлиять на его сложность, как верстка страниц учебника сказывается на восприятии и т. д.

Корпус, содержащий такую информацию, является мультимодальным по своим свойствам, так как содержит информацию о текстах, смысл которых передается с помощью нескольких модусов. Однако поскольку в мировом научном сообществе термин «мультимодальный» в определении корпуса прочно закрепился за корпусами звучащей речи, было принято решение ввести более конкретный термин «иллюстративно-текстовый корпус» по аналогии с англоязычным термином «text-image corpus» [Tirilly et al. 2010; Mohd Yasin et al. 2012]. Для краткости и удобства изложения далее в этой статье мы обозначим его как *корпус*.

2. Состав и объем корпуса

При отборе учебников мы руководствовались несколькими факторами: распространенность учебника, соответствие ФГОС [Приказ 2009], новизна и оригинальность методических школ, результаты опроса учителей и родителей школьников. Для исследования были отобраны 6 наиболее репрезентативных линеек учебников по русскому языку для младшей школы[2]. С одной стороны, были отобраны учебники, входящие в традиционные программы (например, «Школа России», «Перспективная начальная школа»), которые используются в большинстве российских школ. С другой стороны, в фокусе исследования оказались учебники развивающих программ (например, система Л. В. Занкова). Кроме того, корпус был расширен экземплярами, исключенными из Федерального перечня учебников из-за несоответствия ФГОС, но представляющими интерес для научного исследования (например, учебники под редакцией М. С. Соловейчик и Н. С. Кузьменко; учебники под редакцией Г. Г. Граник и В. В. Рубцова).

К настоящему моменту объем аннотированного вручную корпуса составляет 26 учебников (около 380 тыс. токенов), более подробная информация об объеме корпуса представлена в табл. 1. Разметка остальных отобранных учебников продолжается.

Таблица 1. **Объем корпуса учебников русского языка для младшей школы**

	1 класс	2 класс	3 класс	4 класс
токенов	76 737	89 966	103 738	120 444
предложений	9 989	13 638	15 385	16 720
текстовых блоков	4 772	4 975	4 871	4 283
изображений	2 172	3 560	3 796	3 905

3. Принципы разметки корпуса

Простейшим элементом данного корпуса является законченный, визуально отделяемый блок текста, размеченный по служебным (класс, автор, страница, ссылка на изображение страницы) и основ-

[2] Русский язык: учеб. 1–4 кл.: В. П. Канакина, В. Г. Горецкий (2013–2014); Н. А. Чуракова, М. Л. Каленчук (2013–2017); Т. Г. Рамзаева (2008-2013); Н. В. Нечаева (2013–2014); М. С. Соловейчик, Н. С. Кузьменко (2014–2017), Г. Г. Граник и В. В. Рубцов (2012–2015).

ным метатекстовым параметрам. Остановимся подробнее на основных параметрах разметки.

3.1. Аппарат учебника

Этот параметр иллюстрирует отнесенность текстового блока к одному из структурных элементов учебника: текст, наполнение упражнения, формулировка задания, справочная информация и др. Благодаря такой разметке можно проводить исследования как на полном текстовом материале учебника (например, подсчёт базовых метрик текста, таких как рост средней длины предложения, увеличение количества терминов, общего объема лексики в учебниках разных классов), так и на отдельных блоках. Например, корпусные исследования формулировок заданий позволяют сделать выводы о методических особенностях пособия (в одних учебниках встречается больше заданий на развитие устной речи, в других — письменной).

3.2. Текстовый компонент

Текстовый материал учебника размечается по способу организации текста (проза, поэзия) и степени его аутентичности (сконструированный методистами, адаптированный, аутентичный). Так, табл. 2 иллюстрирует количество аутентичных текстовых фрагментов в учебниках для 1 класса разных авторских коллективов. Для удобства описания в работе мы будем называть все учебники (1–4 класс) по фамилии первого автора.

Таблица 2. **Использование авторских текстов в учебниках 1 класса**

	Канакина	**Чуракова**	**Рамзаева**	**Нечаева**
Всего текстов	95	111	101	96
Доля аутентичных текстов	71 %	23 %	58 %	48 %
Количество авторов	37	9	21	26

Данные табл. 2 позволяют сделать вывод, что среди авторов нет единого мнения по включению аутентичных фрагментов в учебник: разброс составляет от 23 % до 71 % от общего числа текстов. Максимальное количество таких текстов и разнообразие авторов наблюдается в учебнике Канакиной. В учебнике Чураковой, наоборот, преобладают сюжетные сконструированные авторами тексты, где герои

Маша и Миша вместе с читателями знакомятся с миром русского языка. Также интересно проследить предпочтения авторов касательно выбора источников аутентичных фрагментов. Так, во всех учебниках для 1 класса присутствуют тексты устного народного творчества — пословицы, поговорки, загадки, считалки, однако в разном количестве: если в учебниках Канакиной и Нечаевой широко используются тексты УНТ (25 и 21 вхождений соответственно), учебники Чураковой и Рамзаевой используют такие тексты значительно реже (1 и 3 раза соответственно). Лидерами по встречаемости и воспроизводимости во всех 4 анализируемых учебниках являются традиционные детские авторы прошлого столетия: С. Маршак (43), К. Чуковский (6), Б. Заходер (5), А. Барто (5). Однако в учебниках первого класса можно встретить и авторов XIX в.: В. Одоевского, А. Фета, К. Ушинского.

3.3. Визуальный компонент

Иллюстративный материал в корпусе размечен на основании функций, которые выполняет каждая иллюстрация в учебнике (навигация, декорация и т. д.) и информации, которая «считывается» с изображения (степень культурной маркированности). По данным параметрам размечаются не только рисунки, фотографии, репродукции и схемы, но и декоративные элементы на странице, специфичный шрифт, различные выделения. Всё это в совокупности погружает школьника в мир изучаемого предмета. С одной стороны, иллюстрации отражают действительность, а с другой — формируют картину мира ученика, в том числе выполняя социокультурную функцию.

Закономерно самым распространенным типом иллюстрации по корпусу является навигация (49 %), так как все учебники снабжены специальными картинками-иконками для удобства ориентирования в материале и структуре. В линейке под ред. Канакиной частотны репрезентация содержания (32 %) и репрезентация лексики (28 %), а в учебниках под ред. Рамзаевой и Чураковой — декорация (38 % и 34 % соответственно).

С одной стороны, иллюстративный материал способствует пониманию текста, дополняет его, а также упрощает восприятие нестандартной информации, с другой стороны, может наоборот «уводить» внимание, отвлекая от текста, требуя при этом от учеников дополнительных когнитивных усилий по обработке изображения. Например, в учебнике Чураковой в качестве декорации к справочной информации представлена летучая мышь (см. рис. 1).

Рис. 1. **Иллюстрация из учебника под ред Н. А. Чураковой 2 кл. 2 ч. 2013, С. 80.**

Выбор иллюстративного материала для «заданий по картинке» также представляет интерес. Зачастую такие задания предполагают работу с репродукциями. Однако в учебниках Канакиной используются только классические художественные произведения (например, «Девочка с персиками» В. А. Серова), в то время как в учебнике Чураковой более разнообразный материал, включающий картину импрессиониста Клода Моне «Прогулка» и сатирические рисунки Х. Бидструпа).

Пример разметки корпуса по описанным выше параметрам представлен в табл. 3.

Таблица 3. **Пример разметки корпуса учебников**

Пункты разметки	Текстовый блок 1	Текстовый блок 2
Текстовый блок	Прочитайте стихотворение. Какие слова в нём рифмуются?	Теперь, — Сказал мне человек, — Пора и нам сыграть. Я буду Доставать слова, Ты должен рифмовать…
Аппарат учебника	Формулировка задания	Наполнение упражнения. Текст
Тип текста	—	Поэзия
Авторство	—	Аутентичный
Имя автора	—	Д. Чиарди
Иллюстрация	—	Репрезентация содержания
Культурная маркированность	—	низкая

Заключение

Описанный формат разметки открывает возможности для большого количества исследований. Опишем самые перспективные с нашей точки зрения области применения корпуса учебников.

1. Исследование сложности текстов учебников, их доступности для соответствующего возраста, роль иллюстративного компонента в восприятии информации, сравнение сложности учебника русского языка с текстами других предметов.
2. Анализ лексического состава учебников: воспроизводимость лексики и процесс формирования словарного запаса школьника, тематическое распределения текстов учебника, принадлежности к эпохе, релевантности современным школьникам.
3. Поиск общего терминологического ядра учебников по русскому языку.
4. Исследование социокультурной информации, транслируемой посредством учебника: формирование национальных и гендерных стереотипов.

Литература

1. Вятютнев М. Н. (1984), Теория учебника русского языка как иностранного (методические основы). М.
2. Микк, Я. А. (1981), Оптимизация сложности учебного текста. В помощь авторам и редакторам. М.
3. Приказ Минобрнауки России от 06.10.2009 N 373 (ред. от 31.12.2015) «Об утверждении и введении в действие федерального государственного образовательного стандарта начального общего образования» (Зарегистрировано в Минюсте России 22.12.2009 N 15785).
4. Bakar, Kesuma A., Zarina Othman, and Fuzirah Hashim. «Making Representational Meanings of Gender Images in Malaysian School English Textbooks: The Corpus Way.» SSRN Electronic Journal, 0ADAD.
5. Behnke, Yvonne. (2016), How textbook design may influence learning with geography textbooks. Nordidactica: Journal of Humanities and Social Science Education. no 2016: 38–62.
6. Carney, R. N., & Levin, J. R. (2002), Pictorial Illustrations Still Improve Students' Learning From Text. Educational Psychology Review, 14(1), 5–26.
7. Guo, D., Wright, K. L., & McTigue, E. M. (2018), A Content Analysis of Visuals in Elementary School Textbooks. The Elementary School Journal, 119 (2), 244–269.
8. Jewitt, C. (2008), Multimodality and Literacy in School Classrooms. Review of Research in Education, 32(1), 241–267.
9. Kress G. Reading images: the grammar of visual design / G. Kress, van Leeuwen T. (1996), New York: Routledge.

10. Kuznetsova, P. Ordonez, V., Berg, A., Berg, T., Choi, Y. Generalizing image captions for image-text parallel corpus. ACL 2013 — 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference 2, pp. 790–796.
11. Mohd Yasin, M. S., Yasin, M. S. M., Hamid, B. A., Othman, Z., Bakar, K. A., Hashim, F., & Mohti, A. (2012), A Visual Analysis of a Malaysian English School Textbook: Gender Matters. Asian Social Science, 8(12).
12. Schneider, S., Dyrna, J., Meier, L., Beege, M., Rey, G. D. (2018), How affective charge and text-picture connectedness moderate the impact of decorative pictures on multimedia learning. Journal of Educational Psychology, 110 (2), pp. 233–249.
13. Tirilly, Pierre & Claveau, Vincent & Gros, Patrick. (2010), News image annotation on a large parallel text-image corpus. 7th Language Resources and Evaluation Conference, LREC'10.

References

1. Prikaz Minobrnauki Rossii ot 06.10.2009 N 373 (red. ot 31.12.2015) «Ob utverzhdenii i vvedenii v dejstvie federal'nogo gosudarstvennogo obrazovatel'nogo standarta nachal'nogo obshhego obrazovaniya» [Order of the Ministry of Education and Science of Russia 06.10.2009 N 373 "On approval and implementation of the federal state educational standard of primary general education"].
2. Bakar, Kesuma A., Zarina Othman, and Fuzirah Hashim. «Making Representational Meanings of Gender Images in Malaysian School English Textbooks: The Corpus Way.» SSRN Electronic Journal, 0ADAD.
3. Behnke, Yvonne. (2016), How textbook design may influence learning with geography textbooks. Nordidactica: Journal of Humanities and Social Science Education. no 2016:. 38–62.
4. Carney, R. N., & Levin, J. R. (2002), Pictorial Illustrations Still Improve Students' Learning From Text. Educational Psychology Review, 14(1), 5–26.
5. Guo, D., Wright, K. L., & McTigue, E. M. (2018), A Content Analysis of Visuals in Elementary School Textbooks. The Elementary School Journal, 119 (2), 244–269.
6. Jewitt, C. (2008), Multimodality and Literacy in School Classrooms. Review of Research in Education, 32(1), 241–267.
7. Kress G. Reading images: the grammar of visual design / G. Kress, van Leeuwen T. (1996), New York: Routledge.
8. Kuznetsova, P. Ordonez, V., Berg, A., Berg, T., Choi, Y. Generalizing image captions for image-text parallel corpus. ACL 2013 — 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference 2, pp. 790–796.
9. Mikk, J. A. (1981), Optimizatsiya slozhnosti uchebnogo teksta: v pomoshch' avtoram i redaktoram [Optimizing the complexity of an educational text: to help authors and editor]. Moscow.
10. Mohd Yasin, M. S., Yasin, M. S. M., Hamid, B. A., Othman, Z., Bakar, K. A., Hashim, F., & Mohti, A. (2012), A Visual Analysis of a Malaysian English School Textbook: Gender Matters. Asian Social Science, 8(12).

11. Schneider, S., Dyrna, J., Meier, L., Beege, M., Rey, G. D. (2018), How affective charge and text-picture connectedness moderate the impact of decorative pictures on multimedia learning. Journal of Educational Psychology, 110 (2), pp. 233–249.
12. Tirilly, Pierre & Claveau, Vincent & Gros, Patrick. (2010), News image annotation on a large parallel text-image corpus. 7th Language Resources and Evaluation Conference, LREC'10.
13. Vyatyutnev, M. N. (1984), Teoriya uchebnika russkogo yazyka kak inostrannogo (metodicheskie osnovy) [Theory of the textbook of Russian as a foreign language (methodological foundations)]. Moscow.

Веселовская Татьяна Сергеевна
Государственный институт русского языка им. А. С. Пушкина (Москва, Россия)
Veselovskaya Tatyana
Pushkin State Russian Language Institute (Moscow, Russia)
E-mail: TSVeselovskaya@pushkin.institute

Лапошина Антонина Николаевна
Государственный институт русского языка им. А. С. Пушкина (Москва, Россия)
Laposhina Antonina
Pushkin State Russian Language Institute (Moscow, Russia)
E-mail: ANLaposhina@pushkin.institute

Купрещенко Ольга Федоровна
Государственный институт русского языка им. А. С. Пушкина (Москва, Россия)
Kupreshchenko Olga
Pushkin State Russian Language Institute (Moscow, Russia)
E-mail: OFKupreshchenko@pushkin.institute

Э. Мелконян
H. Melkonyan

СООТНОШЕНИЕ КОРПУСНОЙ ЛИНГВИСТИКИ И ТИПОЛОГИИ

RELATIONSHIP BETWEEN CORPUS LINGUISTICS AND TYPOLOGY

Аннотация. Современная лингвистика выбрала курс на прикладные дисциплины, которые, по существу, направлены на приближение теории к практике, на решение практических задач. В числе современных прикладных лингвистических дисциплин заслуженно выделяется корпусная лингвистика. Эта молодая отрасль лингвистики, можно сказать, связана со всеми другими лингвистическими дисциплинами. Она черпает материал из всех дисциплин и, обрабатывая его, возвращает им. Можно заметить особую взаимосвязь между корпусной лингвистикой и лингвистической типологией, характеристике которой и посвящена данная публикация.
Ключевые слова. корпусная лингвистика, типология, соотношение, статистика, частотность.

Abstract. Modern linguistics has chosen the course of applied disciplines, which are essentially aimed at bringing the theory closer to practice, at solving practical problems. Among modern applied linguistic disciplines corpus linguistics is deservedly distinguished. This young linguistic branch can be said to be connected with all other linguistic disciplines. It takes the material from everyone and returns it back after processing. A special relationship can be observed between corpus linguistics and linguistic typology, the characteristics of which our publication is directed to.
Keywords. corpus linguistics, typology, relationship, statistics, frequency.

1. Цель и задачи этих дисциплин

1.1. «Миссия» корпусной лингвистики

Корпусная лингвистика, являясь одной из сравнительно молодых лингвистических дисциплин, становится более востребленной с точки зрения практики.

Объектом исследования и ее «продуктом» является **корпус**[1] — набор электронных текстов, с поисковыми возможностями и разными фильтрами. «*Корпус — это информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме*» [НКРЯ].

Можно сказать, что в корпусной лингвистике совмещаются метод анализа непосредственных составляющих и метод статистики, один — чисто лингвистический, а другой — междисциплинарный метод, заимствованный из социологии.

[1] Исторически, предшественником корпусов считается брауновский, с объемом в 2000 слов, с 500 текстовыми фрагментами разных стилей (Брауновский университет, 1962–1963, США) [Грудева 2012: 34].

Кто и с какой целью может создать корпус? Такую роль может взять на себя как исследовательская группа или организация, так и отдельный специалист в этой области. Цели могут быть самыми разными, а корпусы — варьироваться в объеме. Самый объемный корпус того или иного языка называется «национальным», который, как правило, оснащен интернет-доступностью (напр.: российский — http://www.ruscorpora.ru/, британский — http://www.natcorp.ox.ac.uk/, армянский — http://www.eanc.net/am/composition/). Для таких корпусов важными особенностями считаются [НКРЯ]:

- **Представительность**: сбалансированный состав текстов.
- **Разметка / аннотация**: дополнительная информация о свойствах входящих в него текстов (в Национальном корпусе русского языка есть 5 типов разметок — метатекстовая, морфологическая-словоизменительная, синтаксическая, акцентная и семантическая). Наличие разметок считается отличительной чертой по сравнению с другими «простыми коллекциями» или «библиотеками» текстов, представленными в интернете, а разнообразием и глубиной разметок измеряется научная и учебная ценность корпуса.

Считаем необходимым заметить, что по своей миссии корпусная лингвистика сходна, в первую очередь, со словареведением. Обе они представляют из себя прикладные-практические дисциплины, направленные на обработку языкого материала, но у них есть и другая сторона — теория, направленная на продукт обработки (словарь, корпус).

Что касается практической корпусной лингвистики, то она осуществляет очень важную миссию, а именно — обеспечивание речевым материалом исследователя. Ведь язык в своей полной и точной характеристике открывается именно в своем выражении — в речи, отрывки из которой под взором исследователя становятся текстами. Изучать сущность языка, а не его структуру, можно лишь путем изучения языка в действии — в речевых-текстовых материалах. В процессе перехода языка в речь можно обнаружить два взаимодействующих инструмента-фильтра — норму и узус. Первый контролирует речь с точки зрения нормативных правил, а второй — с точки зрения языковой традиции. Если языковая норма “ссылается” на нормативную лингвистическую литературу, то узус — на языковой обычай. Этот переход мы проиллюстрировали графически в нашей книге [Մելքոնյան 2017: 98] (см. рис. 1).

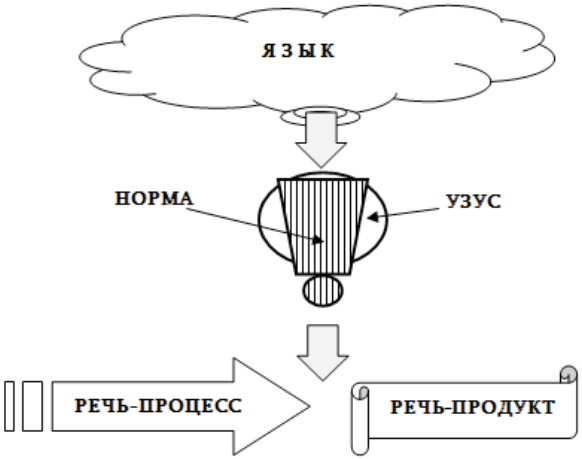


Рис. 1. **Переход от языка к речи**

Для отображения нормы и узуса мы выбрали разные фигуры, подчеркивая то, что они являются разными по существу. Располагая одну фигуру над другой, мы намеревались показать, что они могут иметь и соответствия (совпадающая часть фигур), и различия (фрагменты одной фигуры, выходящие за пределы другой), так как не все нормы могут быть приняты обществом в качестве языковых обычаев — узусов. С другой стороны, не все языковые обычаи зафиксированы как нормы. *Чем же можно измерить узуальность данного языкового явления?* Ответ на этот вопрос остался бы на уровне субъективизма (особенно для изучающих иностранный язык), если бы не пришел на помощь языковой корпус. Так вот миссией номер один для корпусной лингвистики мы считаем **разработку меры для узуальности языковых явлений**.

Адресатом корпуса, в первую очередь, является лингвист, но корпус может заинтересовать самых разных специалистов, и, в частности, из области общественных наук (литературовед, журналист, библиотековед, социолог, психолог, культуролог, юрист, политолог, экономист, педагог, PR менеджер и т. п.). Строго построенный корпус, вдобавок ко всему прочему, может предоставить информацию и о словарных сдвигах, о «биографии» того или иного слова, становясь дня него неким «свидетельством о рождении», и, что более важно, мерой для подтверждения авторских неологизмов.

1.2. «Миссия» лингвистической типологии

Типология является одной из классических лингвистических дисциплин, которая сформировалась еще в начале 19-ого веке, объектом которой является структура языка, а конечной целью — **типологическая классификация языков**. На пути к достижению этой цели ставятся такие промежуточные цели, как: выявление типичных черт отдельных языков, структурная характеристика отдельных языков и языковых групп, выявление универсалий и уникалий, создание универсальной лингвистической теории. Ключевым термином этой научной сферы является **тип**, который понимается в узком и широком значении:

- классовый тип языка, т. е. тип языка на его конкретном уровне (напр: фонетический, словообразовательный, морфологический, синтаксический и т. д.), для конкретной типизируемой области;
- общий тип языка (идеал типологии).
- Хотя первые типологические классификации были основаны на морфологическом принципе (аморфные, агглютинативные, флективные языки), но ставить знак равенства между морфологическими и языковыми типами, в широком значении последних, является уже устаревшим подходом, так как языковые типы проявляют себя на всех уровнях языка.

С другой стороны, языковой тип не бывает представлен в чистом виде: почти всегда с доминирующими чертами можно обнаружить сосуществование хотя бы слабо выраженных альтернативных проявлений. Тип языка при наличии таких разных языковых выражений определятся с помощью статистики текстовых показателей[2]. И именно здесь типолог нуждается в пропорционально строго построенном языковом корпусе.

2. Грани соприкосновения корпусной лингвистики и лингвистической типологии

Грань соприкосновения типологических и корпусно-лингвистических интересов можно привести к **тексту**: типолог обращается к тексту, чтобы с помощью статистических показателей найти ответ на вопрос: к какому типу принадлежит тот или иной язык, а корпусный лингвист берет на себя роль обработки этой базы текстов, предостав-

[2] Это может касаться лишь классового типа, поскольку типологическая классификация языков одновременно на всем уровням, приводит типологию к тупику.

ляя возможность для автоматической статистики. На первый взгляд можно предположить, что связь между обеими научными сферами односторонняя: т. е. типология только забирает, не отдавая ничего взамен. Но это не совсем так, а точнее — это должно быть не так, поскольку типология много чего может отдать корпусной лингвистике, содействуя ее развитию на совершенно новом уровне.

Современные глобализационные процессы, которые интенсивным образом проникают во все области человеческой деятельности, должны привести к круглому столу еще и лингвистов, ставя вопрос об актуальности создания универсальной теории языка. Универсализация лингвистики является основной задачей универсалогии — отрасли лингвистической типологии. А универсальные решения должны быть зафиксированы в языке-эталоне (человеческого языка)[3]. Эталон должен отвечать за информацию о всевозможных языковых выражениях, при этом охарактеризованных универсальным лингвистическим метаязыком. Имея такую информацию, корпусный лингвист может быть снабжен универсальными инструментами для маркировки единиц любого языка.

Это сможет послужить созданию параллельных текстов, что, по нашему мнению, заинтересует не только лингвистов, а специалистов самых разных областей, в первую очередь, переводчиков и лиц, изучающих иностранные языки. Так, например, универсальная теория, обрабатывая информацию о всевозможных моделях словообразования в языках, сможет применять единую модель для отображения структуры каждого слова каждого отдельного языка. Исследователь сможет иметь возможность нахождения многоязычной базы данных по искомой модели. То же самое можно сделать и для выявления структурных качеств других языковых явлений (длина слова, структура слога, структура предложения, структура словосочетаний или других конструкций, сочетаемость слов, структура абзаца, структура текста и т. д.). Переводчик же сможет сравнивать двуязычный или многоязычный переводческий материал с оригиналом.

3. Иллюстрация собственной типологической практики

Так как типология является областью наших научных интересов, мы часто сталкиваемся с задачей выявления текстовых показателей

[3] Плану обработки языка-эталона посвящена наша статья, уже готовая к публикации.

исследуемого явления. При решении этой проблемы, однако, мы не основывались на базе Национального корпуса армянского языка (ԱՐԵՎԱԿ). Хотя этот корпус имеет неплохой объем (110 млн. слов, тексты которого берут начало со второй половины 19-го века и являются текстами разных стилей, в том числе и тексты-переводы) и, в частности, удачные заметки, которые могут быть полезными при решении различных лингвистических задач, однако, для типологических исследований, как нам кажется, он не соответствует требованиям, так как:

- объем разнообразности текстов не равномерен: доминирует художественный материал;
- недостаточно текстов новейшего периода (2000-е годы);
- иногда тексты не датированы.

Для решении типологических задач мы создали свой корпус, объемом в 219 590 звуков, прерогативой которого стала строгая пропорциональная вовлеченность текстов разных стилей различных авторов. При помощи применения «ручных» корпусов мы нашли решения на следующие вопросы, касающиеся современного литературного восточно-армянского языка (официальный-государственный язык Армении):

- вокалическому или консонантному типу языков принадлежит язык?
- какие частотные показатели имеются у каждого звука?
- язык является аналитическим или синтетическим языком в именной системе?
- какие структуры слогов и с какими частотными показателями существуют в нем?
- какие текстовые показатели есть у дифтонгоидов?

Нетрудно заметить, что чем выше уровневая принадлежность изучаемого языкового явления, тем важнее иметь объемный текстовый материал.

4. Выводы

- Обе изучаемые дисциплины являются перспективными в области филологии.
- При их отношении друг к другу центр тяжести пока что находится в области корпусной лингвистики.

- Эти научные дисциплины могут войти в область равного взаимодействия, способствуя повышению взаимного научного уровня. Но для этого каждая из областей должна сначала решить свои собсвенные вышеперечисленные проблемы, которые могут препятствовать их продуктивному взаимодействию.

Литература

1. *Грудева Е.В.* (2012), Корпусня лингвистика, Учебное пособие, Москва, изд. ФЛИНТА, 2012, 165 с..
2. *ВАНК*, Национальный корпус восточноармянского языка, http://www.eanc.net/EANC/search/?interface_language=am.
3. *НКРЯ,* Национальный корпус русского языка, http://www.ruscorpora.ru/.
4. *Մելքոնյան Հ.Ա.* (2017), Լեզվաբանության հիմունքներ. Ուսումնական ձեռնարկ,Վանաձոր, «ՍԻՄ» տպ., 492 էջ.
5. *Մելքոնյան Հ.Ա.* (2014), Տիպաբանություն. հայերենի տիպաբանության հարցեր. Ուսումնական ձեռնարկ, Վանաձոր, «ՍԻՄ» տպ., 2014, 314 էջ:

References

1. *Grudjeva E. V.* (2012), Corpusnaja lingvistka, Uchebnoje posobie. [Corpus linguistics. Teaching manual]. Moscow.
2. *VANC* (*AREVAK*). Nacionalnij corpus vostochnoarmjanskogo jazyka. [Russian National Corpus]. http://www.eanc.net/EANC/search/?interface_language=am
3. *NCRJA,* Nacionalnij corpus russkogo jazyka, [Russian National Corpus]. http://www.ruscorpora.ru/.
4. *Melkonyan H. A.* (2017), Lezvabanuthjan himunqner: Usumnakan dzernark. [Basics of Linguistics. Teaching manual]. Vanadzor.
5. *Melkonyan H. A.* (2014), Tipabanuthjun: Hajereni tipabanutjan harcer: Usumnakan dzernark. [Typology. Typology issues of Armenian Language. Teaching manual]. Vanadzor.

Мелконян Эгине Азатовна
Ванадзорский государственный университет (Армения)
Heghine Azat Melkonyan
Vanadzor State University (Armenia)
E-mail: HeghineMelk@yandex.ru

А. Ю. Сиротина, Н. В. Лукашевич
A. Yu. Sirotina, N. V. Loukachevich

ОПЫТ СОЗДАНИЯ КОРПУСА ТЕКСТОВ В СФЕРЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ[1]

TOWARDS CONSTRUCTION OF AN ANNOTATED CORPUS IN CYBERSECURITY

Аннотация. В данной работе рассматривается задача разметки корпуса неструктурированных текстов на русском языке в сфере информационной безопасности. Были выявлены основные типы именованных сущностей, релевантных для данной предметной области, сформирован набор тегов, разработана подробная инструкция для разметки текстов. Также в данной работе обсуждаются сложные для разметки и неоднозначные контексты, специфичные для области информационной безопасности.
Ключевые слова. Информационная безопасность, извлечение именованных сущностей, корпусная лингвистика.

Abstract. In this paper we discuss the construction of a new corpus of unstructured Russian texts concerned with cybersecurity problems. Our analysis of the texts in question reveals a number of specific named entities' types that are relevant for the subject area. We also present a new instruction for human annotators. The instruction includes description of the most problematic cases and contexts.
Keywords. Cybersecurity, named entity extraction, corpus linguistics.

1. Введение

Для успешного предотвращения утечки и потери данных специалистам в сфере информационной безопасности (ИБ) необходимо максимально быстро получать актуальные сведения о хакерской активности, вирусах и уязвимостях. Одним из наиболее эффективных инструментов сбора этих сведений является система автоматического извлечения информации в данной предметной области. Для обучения и тестирования такой системы необходим размеченный корпус текстов в сфере ИБ. Этот корпус также может быть использован для лингвистического анализа текстов данной предметной области.

Актуальная информация об уязвимостях публикуется на специализированных интернет-ресурсах и форумах в виде **неструктурированных** текстов. Поэтому наиболее эффективной для сферы ИБ является система, способная извлекать информацию из **неструктури-**

[1] Работа частично поддержана РФФИ (проект 16-29-09606)

рованных текстов, и, следовательно, именно такие тексты должны составлять значительную долю от общего объема размеченного корпуса.

2. Обзор близких работ

Задача разметки корпуса текстов по ИБ рассматривалась в ряде работ. В работе [Bridges et al. 2013] была произведена автоматическая разметка корпуса. Для оценки её качества вручную были размечены несколько десятков текстов. В работах [Weerawardhana et al 2014; Joshi et al. 2013; Lim et al. 2017] рассматривается задача ручной разметки корпуса.

Большинство корпусов данной предметной области (в том числе корпуса, описанные в работах [Weerawardhana et al 2014; Bridges et al. 2013]) содержат частично структурированные тексты: например, бюллетени по безопасности Майкрософт или статьи из Национальной базы данных уязвимостей США (NVD, National Vulnerability Database). В работе [Joshi et al. 2013] корпус включает в себя как частично структурированные, так и неструктурированные тексты, однако доля последних в корпусе крайне мала. Корпус, описанный в работе [Lim et al. 2017], состоит из слабо структурированных формальных текстов — статей и докладов, которые могут содержать различные списки и таблицы.

Общепринятый набор тегов для разметки текстов по ИБ отсутствует, поэтому авторы работ используют различные наборы тегов в зависимости от задач исследования и степени структурированности текстов. В работе [Bridges et al. 2013] в качестве тегов использовались заголовки полей, которые заполняются при описании уязвимости в NVD. Набор тегов в работе [Joshi et al. 2013] был сформирован исследователями в результате анализа текстовой коллекции. В работе [Lim et al. 2017] в качестве источника тегов выступил словарь для описания уязвимостей MAEC[2]. Авторы работ [Joshi et al. 2013; Lim et al. 2017] отмечают, что многие разметчики испытывали трудности при выборе тега.

Во всех упомянутых выше работах описаны корпуса, содержащие тексты на английском языке. Для русского языка, насколько нам известно, размеченные корпуса текстов по ИБ отсутствуют.

[2] https://maecproject.github.io/

3. Исходный корпус и его разметка

В рамках данной работы источником текстов для корпуса послужили публикации и форумы сайта SecurityLab[3], которые, в отличие от статей из NVD и бюллетеней по безопасности, представляют собой **неструктурированные** тексты. В корпус не вошли публикации длиной меньше 500 слов, а также слишком длинные тексты.

На первом этапе работы была получена частичная автоматическая разметка коллекции. Для этого использовалась система извлечения именованных сущностей (ИС) на русском языке, описанная в работе [Можарова 2017]. В коллекции были автоматически размечены персоны, локации, организации и СМИ (теги **Person**, **Loc**, **Org и Media** соответственно).

Для ручной разметки был сформирован следующий набор тегов: **Hacker** (отдельные хакеры); **Hacker_Group** (группы хакеров); **Program** (программы, в том числе сайты, функции, части программ); **Device** (электронное оборудование); **Tech** (технологии, написанные с большой буквы); **Virus** (зловредное ПО разной природы); **Event** (различные события и мероприятия).

Тексты размечались четырьмя независимыми разметчиками, при этом не все разметчики являлись специалистами в сфере ИБ. Разметка производилась при помощи онлайн-инструмента для аннотации BRAT[4].

На данном этапе разметки аннотаторам было дано малое количество указаний о том, как следует размечать ИС различных типов. Предполагалось, что отсутствие изначальной инструкции по разметке позволит выявить наиболее часто встречающиеся ошибки разметки и сформировать в дальнейшем полноценный набор правил для разметчиков.

Всего было размечено 1124 публикаций. В корпус вошли только такие тексты, которые содержат хотя бы один из следующих тегов: **Hacker**, **Hacker_Group**, **Program**, **Device**, **Tech**, **Virus**. В результате объем корпуса составил 861 текст (406488 токенов).

Далее в целях устранения неточностей разметки был произведен вторичный анализ размеченных текстов. В ходе анализа было установлено, что разметчики склонны принимать разные решения при разметке одинаковых контекстов, что привело к большому количеству

[3] https://www.securitylab.ru/

[4] http://brat.nlplab.org/

ошибок и неточностей и общей непоследовательности разметки. Так, например, язык программирования *Java* в 30 случаях получил неверный тег **Program** и в 5 случаях — верный тег **Tech**; аббревиатура *СКЗИ* (средство криптографической защиты информации) в 13 случаях была неверно размечена как **Program**, в 43 случаях получила верный тег **Tech.**

3.1. Примеры непоследовательной разметки

В результате анализа размеченных текстов были выделены случаи непоследовательной разметки, в их числе:

- Включение или исключение из аннотации парных знаков препинания (скобок или кавычек), окружающих ИС: *главный ботмастер по прозвищу «**Netkairo**»*;
- Включение или исключение из аннотации русскоязычной части слова с дефисным написанием, где первая часть является ИС: *DDoS-атака*, *Andriod-устройство*;
- Выделение одной или двух ИС в контекстах, где название программы или технологии содержит название разработчика: *Apple iOS, Microsoft Internet Explorer 10*;
- Выделение или отсутствие ИС на номерах версий продукта в случае, если они перечислены вслед за названием продукта: *Android 7.1, 7.1.1 и 7.1.2*;

Кроме того, в текстах были обнаружены определенные типы ИС, выбор тега для которых был затруднен. К таким типам относятся: платежные системы; криптовалюта; языки программирования; команды, методы и классы объектов в различных языках программирования; библиотеки подпрограмм; средства разработки; параметры и факторы; программные ошибки; ссылки и директории. Также разметчики испытывали трудности при выборе тега для ИС с дефисным написанием: *Android-устройство* (**Program** или **Device**), *GSM-телефон* (**Tech** или **Device**).

Вместе с тем были зафиксированы такие группы ИС, выбор тега для которых зависит от контекста:

- **Организация/Локация**: *Кремль, Белый Дом, Пентагон*;
- **Организация/Программа**: *Яндекс, Google*;
- **Программа/Девайс**: различные программно-аппаратные средства и комплексы (например, межсетевые экраны);

Такое многообразие неоднозначных и сложных для разметки контекстов делает разметку корпуса текстов в сфере ИБ нетривиальной задачей. Наш опыт показывает, что разметчики, вне зависимости от того, являются ли они специалистами ИБ или нет, нуждаются в полноценной инструкции, включающей в себя подробное описание используемых тегов и правила выбора тега для ИС различных типов.

4. Новая инструкция и исправление разметки

Во избежание ошибок и неточностей при дальнейшей разметке и при внесении правок в существующую разметку, была разработана новая инструкция, в которой были учтены все сложные и неоднозначные контексты, обнаруженные в ходе анализа размеченных текстов. Так, в рамках новой аннотации были приняты следующие решения:

- Парные знаки препинания, окружающие ИС, включаются в аннотацию;
- В аннотацию ИС с дефисным написанием входят обе её части: и предшествующая дефису, и следующая за ним;
- В контекстах, где название программы или технологии содержит название ее разработчика, последнее не выделяется в качестве отдельной ИС;
- На версиях продукта, перечисленных после названия продукта, отдельные ИС выделяются, если названия версий содержат буквенные символы: *PowerPC G3, G4 и G5;*

Для каждого тега в инструкции указаны типы ИС, которым он присваивается. Наиболее актуальна эта мера была для тегов **Program**, **Device** и **Tech**, приписывание которых вызывало больше всего ошибок:

- **Program:** операционные системы (*iOS 9*); браузеры (*Google Chrome*); скачиваемые и устанавливаемые программы (*Adblock*); сайты (но не ссылки: *SlideShare*); файлы и процессы, названия которых записаны в виде «*имя.расширение*» (*Autorun.exe*, *ipfilter.dat*) и пр.;
- **Tech:** ряд русскоязычных аббревиатур (*СКУД, СЗПДн и* пр.); языки программирования (*JavaScript*); расширения (*XML*); протоколы и стандарты (*pptp*, *SSDP*);

Для каждого из вышеперечисленных тегов также указываются типы ИС, которым данный тег **не** присваивается. Так, например, было

принято решение **не** присваивать тег **Program**: ссылкам и директориям, непрерывным частям кода, ряду аббревиатур (*ПО*, *ОС*, *СПО* и пр.).

В соответствии с новой инструкцией были исправлены ошибки и неточности в разметке корпуса.

5. Заключение

В данной работе рассмотрена задача создания размеченного корпуса текстов в сфере ИБ. В результате анализа неструктурированных текстов были выделены основные типы ИС, которые встречаются в текстах по ИБ, а также перечислены трудности, которые возникают в ходе разметки. В целях создания последовательной разметки была разработана инструкция, включающая в себя подробное описание тегов и правила выделения и аннотации ИС в неоднозначных контекстах. Корпус текстов был размечен в соответствии с разработанной инструкцией и может быть в дальнейшем использован для обучения и тестирования моделей извлечения ИС в сфере ИБ.

Литература

1. *Можарова В. А.* (2017), Методы машинного обучения в задаче извлечения именованных сущностей на русском языке. Дипломная работа (магистр), МГУ имени М. В. Ломоносова.
2. *Jones C. L., Bridges R. A., Huffer K. M., Goodall J. R.* (2015), Towards a relation extraction framework for cyber-security concepts. In Proceedings of the 10th Annual Cyber and Information Security Research Conference, p. 11.
3. *Joshi A., Lal R., Finin T., Joshi A.* (2013), Extracting cybersecurity related linked data from text. In Semantic Computing (ICSC), 2013 IEEE Seventh International Conference, pp. 252–259.
4. *Lim S. K., Muis A. O., Lu W., Ong C. H.* (2017), MalwareTextDb: A database for annotated malware articles. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1557–1567.
5. *Weerawardhana S., Mukherjee S., Ray I., Howe A.* (2014), Automated Extraction of Vulnerability Information for Home Computer Security. In International Symposium on Foundations and Practice of Security, pp. 356–366.

References

1. *Mozharova V. A.* (2017), Metody mashinnogo obuchenija v zadache izvlechenija imenovannyh sushchnostej na russkom jazyke. [Approaches to Machine Learning for Named Entity Extraction in Russian]. Master's thesis, Moscow State University.

2. *Jones C. L., Bridges R. A., Huffer K. M., Goodall J. R.* (2015), Towards a relation extraction framework for cyber-security concepts. In Proceedings of the 10th Annual Cyber and Information Security Research Conference, p. 11.
3. *Joshi A., Lal R., Finin T., Joshi A.* (2013), Extracting cybersecurity related linked data from text. In Semantic Computing (ICSC), 2013 IEEE Seventh International Conference, pp. 252–259.
4. *Lim S. K., Muis A. O., Lu W., Ong C. H.* (2017), MalwareTextDb: A database for annotated malware articles. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1557–1567.
5. *Weerawardhana S., Mukherjee S., Ray I., Howe A.* (2014), Automated Extraction of Vulnerability Information for Home Computer Security. In International Symposium on Foundations and Practice of Security, pp. 356–366.

Сиротина Анастасия Юрьевна
Московский государственный университет им. М. В. Ломоносова (Россия)
Sirotina Anastasiia
Moscow State University (Russia)
E-mail: overnastuhed@yandex.ru

Лукашевич Наталья Валентиновна
Московский государственный университет им. М. В. Ломоносова (Россия)
Loukachevich Natalia
Moscow State University (Russia)
E-mail: louk_nat@mail.ru

BVS CORPUS: A MULTILINGUAL PARALLEL CORPUS OF BIOMEDICAL SCIENTIFIC TEXTS AND TRANSLATION EXPERIMENTS

Abstract. The BVS database (Health Virtual Library) is a centralized source of biomedical information for Latin America and Carib, created in 1998 and coordinated by BIREME (Biblioteca Regional de Medicina) in agreement with the Pan American Health Organization (OPAS). Abstracts are available in English, Spanish, and Portuguese, with a subset in more than one language, thus being a possible source of parallel corpora. In this article, we present the development of parallel corpora from BVS in three languages: English, Portuguese, and Spanish. Sentences were automatically aligned using the Hunalign algorithm for EN/ES and EN/PT language pairs, and for a subset of trilingual articles also. We demonstrate the capabilities of our corpus by training a Neural Machine Translation (OpenNMT) system for each language pair, which outperformed related works on scientific biomedical articles. Sentence alignment was also manually evaluated, presenting an average 96 % of correctly aligned sentences across all languages. Our parallel corpus is freely available, with complementary information regarding article metadata.
Keywords. Parallel Corpora, Biomedical, Translation, Spanish, English.

1. Introduction

The availability of cross-language parallel corpora is one of the basis of current Statistical and Neural Machine Translation systems (SMT and NMT). Acquiring a high-quality parallel corpus that is large enough to train MT systems, specially NMT ones, is not a trivial task, since it usually demands human curating and correct alignment. In light of that, the automated creation of parallel corpora from freely available resources is extremely important in Natural Language Processing (NLP), enabling the development of accurate MT solutions. Many parallel corpora are already available, some with bilingual alignment, while others are multilingually aligned, with 3 or more languages, such as Europarl [Koehn 2005], from the European Parliament, JRC-Acquis [Steinberget et al. 2006], from the European Commission, OpenSubtitles [Zhang 2014], from movies subtitles.

The extraction of parallel sentences from scientific writing can be a valuable language resource for MT and other NLP tasks. The development of parallel corpora from scientific texts has been researched by several authors, aiming at translation of biomedical articles [Wu et al. 2011; Neves et al. 2016], or named entity recognition of biomedical concepts [Kors et al. 2015]. Regarding Portuguese/English and English/Spanish language pairs, the FAPESP corpus [Aziz and Specia 2011], from the Brazilian magazine revista pesquisa FAPESP, contains more than 150,000 aligned sentences per language pair, constituting an important language resource.

In Latin America and Carib, the Pan American Health Organization (OPAS), in agreement with BIREME (Biblioteca Regional de Medicina), maintains the BVS database, which is an important source of biomedical texts in three main languages: English, Spanish, and Portuguese. Currently, BVS has more than 1 million texts indexed, and also provides integrated search capabilities with PUBMED.

In this article, we explore the BVS database as a source of parallel corpora for the 3 aforementioned languages. We developed a trilingual parallel corpus with the 3 languages, as well as parallel corpora of English/Portuguese and English/Spanish abstracts. In addition, we provided various metadata regarding the publications.

2. Licensing

Most articles in the BVS database are open access documents. In order to avoid any copyright issues, we included in our datasets only open access documents. To retrieve license information, we crawled the BVS website containing information about the indexed journals[1] as well as the Directory of Open Access Journals[2].

3. Materials and Methods

In this section, we detail the information retrieved from BVS website, the filtering process, the sentence alignment, and the evaluation experiments. Figure 1 shows the diagram of the steps followed for the development of the parallel corpora.

3.1 Document retrieval and parsing

BVS's website[3] offers simple and advanced search capabilities. We iteratively queried the database to retrieve all lists of results, which were then parsed and all relevant contents stored, such as authorship, title, and abstracts. We adopted the MongoDB database system, as it is document-oriented, and allows for the easy querying and storage of this type of data.

After the initial filtering, the resulting documents were processed for language checking[4] to make sure that there was no misplacing of abstract

[1] http://portal.revistas.bvs.br/

[2] https://doaj.org/

[3] http://bvsalud.org/

[4] https://github.com/Mimino666/langdetect

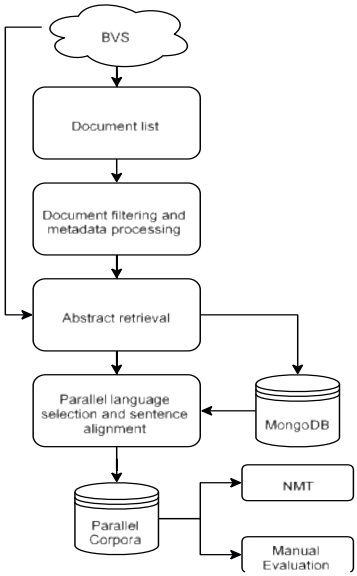


Fig. 1. **Method employed for corpora building**

language (e.g. English abstracts in the Portuguese field, or the other way around), removing the documents that presented such inconsistency. In addition, we also removed newline/carriage return characters (i.e \n and \r), as they would interfere with the sentence alignment tool.

3.2 Sentence alignment

For sentence alignment, we used the LF aligner tool[5], a wrapper around the Hunalign algorithm [Varga et al. 2005], which provides an easy to use and complete solution for sentence alignment, including pre-loaded dictionaries for several languages.

Hunalign uses Gale-Church sentence-length information to first automatically build a dictionary based on this alignment. Once the dictionary is built, the algorithm realigns the input text in a second iteration, this time combining sentence length information with the dictionary. When a dictionary is supplied to the algorithm, the first step is skipped. A drawback of Hunalign is that it is not designed to handle large corpora (above 10 thousand sentences), causing large memory consumption. In these cases, the

[5] https://sourceforge.net/projects/aligner/

algorithm cuts the large corpus in smaller manageable chunks, which may affect dictionary building.

The parallel abstracts were supplied to the aligner, which performed sentence segmentation followed by sentence alignment. After sentence alignment, the following post-processing steps were performed: (i) removal of all non-aligned sentences; (ii) removal of all sentences with fewer than three characters, since they are likely to be noise.

3.3 Machine translation evaluation

To evaluate the usefulness of our corpus for MT purposes, we trained an NMT model using the OpenNMT system [Klein et al. 2017] for all language pairs. The produced translations were evaluated according to the BLEU score [Papineni et al. 2002].

3.4 Manual evaluation

Although the Hunalign algorithm usually presents a good alignment between sentences, we also conducted a manual validation to evaluate the quality of the aligned sentences. We randomly selected 300 sentences, 100 for the triligual subset, and 100 for each subset of EN/PT and EN/ES. If the pair was fully aligned, we marked it as ”correct”; if the pair was incompletely aligned, due to segmentation errors, for instance, we marked it as ”partial”; otherwise, when the pair was incorrectly aligned, we marked it as ”no alignment”.

4. Results and Discussion

In this section, we present the corpus' statistics and quality evaluation regarding NMT system, as well as the manual evaluation of sentence alignment.

4.1 Corpus statistics

Table 1 shows the statistics (i.e. number of sentences) for the aligned corpus according to the 2 language pairs and the trilingual subset. The dataset is available[6]in TMX format [Rawat et al. 2016], since it is the standard format for translation memories. We also made available the aligned corpus in an SQLite database in order to facilitate future subset selection. In this database, we included the following metadata information: year, keywords

[6] 10.6084/m9.figshare.8067533

in the available languages, database of origin, country, authorship, and URL for the full-text when available.

Table 1. Corpus **statistics according to language pair**

Language Pairs	Sentences
EN/PT	711,475
EN/ES	789,547
EN/PT/ES	203,719

4.2 Translation experiments

Prior to MT experiments, sentences were randomly split in three disjoint datasets: training, development, and test. Approximately 14,000 sentences were allocated in the development and test sets, while the remaining was used for training. For the NMT experiment, we used the Torch implementation[7] to train a 2-layer LSTM model with 500 hidden units in both encoder and decoder, with 20 epochs. During translation, the option to replace UNK words by the word in the input language was used.

Table 2 presents the BLEU scores for both translation directions with the 3 language pairs for the development and test partitions. We also included the best scores from a similar parallel corpus from Scielo [5] as a benchmark.

Table 2. **BLEU scores for translation using OpenNMT for the development and test partitions. Previous related work by Neves et al.(2016) is also presented for comparison in the right-hand column as benchmarking**

Language Pairs		Dev	Test	Bench
EN-ES	EN→ES	34.80	34.96	32.75
	ES→EN	33.82	34.28	30.53
PT-ES	PT→ES	55.78	56.11	—
	ES→PT	56.26	56.50	—
EN-PT	EN→PT	35.62	36.03	33.37
	PT→EN	35.88	36.12	31.78

[7] http://opennmt.net/OpenNMT/

Our models achieved better performance than the benchmark for all language pairs and directions, with at least 2.21 percentage points (pp) higher for the EN/ES language pair, achieving a maximum of 4.34 pp for the EN/PT language pair. It is noticeable the high scores achieved in the ES/PT pair, which we expect to be due to the high similarity between both languages.

4.3 Sentence alignment quality

We manually validated the alignment quality for 300 sentences randomly selected from the parsed corpus and assigned quality labels according Section 3.4. From all the evaluated sentences, average 96 % were correctly aligned, while average 2 % were partially aligned. The trilingual subset was the one with the best alignment, achieving 97 % correct alignment. The small percentage of no alignment is probably due to the use of Hunalign algorithm with the provided dictionaries. Figure 2 shows the alignment accuracy for all language subsets.

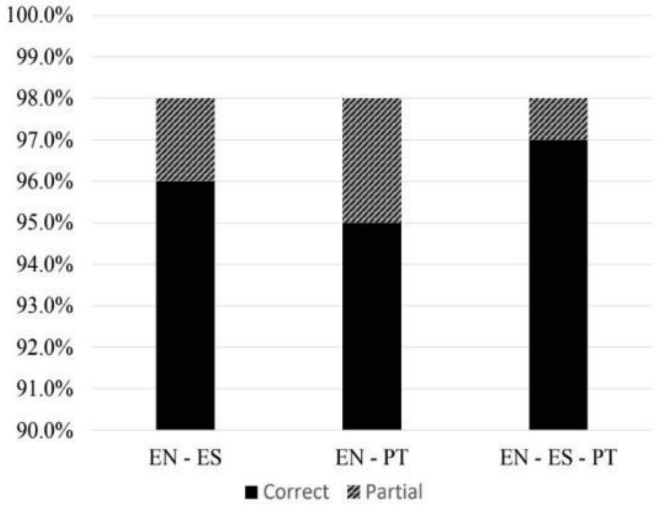


Fig. 2. **Alignment accuracy for the three language subsets**

5. Conclusion and future work

We developed a parallel corpus of biomedical abstracts in English, Spanish, and Portuguese. Our corpus is based on the BVS database, which contains biomedical texts from several sources in Latin America and Carib. The

corpus contains the EN/ES, EN/PT language pairs as well as a trilingual subset of EN/PT/ES sentences.

Our corpora were evaluated through NMT experiments with OpenNMT system, presenting superior performance regarding BLEU score than a related work with a similar corpus. The NMT model presented remarkable results for the PT/ES language pair, possibly due to the similarity between the languages. We also manually evaluated sentences regarding alignment quality, with average 96 % of sentences correctly aligned.

For future work, we foresee the use of the presented corpus in mono and cross-language text mining tasks, such as text classification and clustering. As we included several metadata, these tasks can be facilitated. Other machine translation approaches can also be tested, including the concatenation of this corpus with other multi-domain ones.

Acknowledgments

This work was supported by the Encargo de Gestion SEAD-BSC of the Spanish National Plan for the Advancement of Language technologies, the ICTUSnet INTERREG Sudoe programme, the European Union Horizon2020 eTransafe (grant agreemed 777365) project, and the Amazon AWS Cloud Credits for Research.

References

1. Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In MT summit, volume 5, pages 79–86, 2005.
2. Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. arXiv preprint cs/0609058, 2006.
3. Shikun Zhang, Wang Ling, and Chris Dyer. Dual subtitles as parallel corpora. 2014.
4. Cuijun Wu, Fei Xia, Louise Deleger, and Imre Solti. Statistical machine translation for biomedical text: are we there yet? In AMIA Annual Symposium Proceedings, volume 2011, page 1290. American Medical Informatics Association, 2011.
5. Mariana Neves, Antonio Jimeno Yepes, and Aurelie N´ ev´ eol. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may 2016. European Language Resources Association (ELRA).
6. Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. A multilingual gold-standard corpus for biomedical concept

recognition: the mantra gsc. Journal of the American Medical Informatics Association, 22(5):948–956, 2015.
7. Wilker Aziz and Lucia Specia. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In STIL 2011, Cuiaba, MT, October 2011.´
8. Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor Tron. Parallel corpora for medium density languages. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE, page 247, 2005.
9. G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: OpenSource Toolkit for Neural Machine Translation. ArXiv e-prints, 2017.
10. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings *of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
11. Sunita Rawat, M. B. Chandak, and Nekita Chauhan. *An Approach for Efficient* Machine *Translation Using Translation Memory*, pages 285–291. Springer Singapore, Singapore, 2016.

Felipe Soares
Barcelona Supercomputing Center (BSC) — Spain
E-mail: felipe.soares@bsc.es
Martin Krallinger
Barcelona Supercomputing Center (BSC) — Spain
E-mail: martin.krallinger@bsc.es

T. O. Shavrina, V. Benko

OMNIA RUSSICA: EVEN LARGER RUSSIAN CORPUS[1]

Abstract. This paper focuses on combining Russian open corpus resources into one single source. The article describes the motivation for a gradual integration of existing text resources to create a more general project and analyzes in detail the main steps to merge the existing data to formats based on *NoSketch Engine* corpus standards and interface.
Keywords. Corpus linguistics, Russian corpora, open source, corpus construction.

1. Introduction

Contemporary NLP-technology is inextricably linked with machine learning, and from 2010s on we are experiencing a new scientific revolution, where the latest technologies use generalizations on huge amounts of data. The methods include all sorts of vector models, both for words (like *word2vec* [Mikolov et al. 2013], *GloVE* [Pennington et al. 2014]), and symbols (*fasttext* [Bojanowski et al. 2017]); unsupervised machine translation systems [Lample et al. 2018]; as well as the newest trend — universal language models, pre-trained on huge corpora of the language and replacing the need for developers to have their own data for the task (these include primarily the works [Devlin et al. 2018, Peters et al. 2018, Radford et al. 2018]). Universal encoders show SOTA-quality in the most difficult tasks of natural language understanding — SQUAD, information retrieval, machine translation, etc. In this paradigm, modern corpora are considered as a "training set", exclusively a source of textual data, preferably representative and diverse.

However, not only machine learning tasks require large corpora. The traditional corpus research areas, such as lexicography and teaching of foreign languages are, also changing: for example, studying the differences in the distribution of rare words also require more and more data. Since the word frequency distribution of a language obeys the Zipf law [Zipf 1949], most words of interest to researchers are often quite rare and have a very small IPM[2] value. If IPM is less than one, it is quite difficult to get enough contexts of words to make a theoretical generalization on them, especially when the task concerns verbs, that usually have rather heterogeneous contexts due to complex actant structure and its arbitrary filling. To get 100 contexts

[1] This work has been, in part, funded by the Slovak KEGA and VEGA Grant Agencies, Project No. K-16-022-00, and 2/0017/17, respectively.

[2] IPM — instances per million

for a word with IPM 0.1, you need 1 billion words, and if these contexts are unique, then to get the most frequent you need at least 10 times more data — already 10 billion words, and so on. For comparison, to get at least 100 examples of each use of 3 Russian synonyms — "umen'shit'" ("*to reduce*", IPM 7[3]), "ubavit'" ("*to subtract*", IPM 1.36), and "skostit'" ("*to knock off*", IPM 0.02), you need a corpus of 50 billion words.

In terms of case studies, modern open corpora still do not contain enough data [Shavrina 2019]. We believe that in the dominant paradigm of the corpus as a training set, corpus linguistics now has a new, extremely important application — knowing your corpus, which is possible only when full corpus metadata, genre ratio and original texts is available for the user. Until now, corpus comparison metrics are rather poorly developed (the first attempts to develop such metrics can be found in [Kilgarriff 2012; 2001], where they are used for comparing contexts of corpora). An analysis of the bias and homogeneity of contexts in corpora can be key to the needs of vector models and universal encoders since the question of choosing the best data for training a model remains open.

2. Russian Corpus Resources

Russian belongs to languages with a long history of corpus linguistics, and many projects addressing open access data, national literature, manual and automatic annotation of different kinds are carried out at present. The most notable of them include:

- *Russian National Corpora* (*RNC* [Lyashevskaya et al 2005], about 400 million words), based on national literature and fiction;
- *General Internet Corpus of Russian* (*GICR*, [Selegey et al 2016], 20 billion words), unifying social networks and media sources;
- *OpenCorpora* ([Bocharov et al 2013], 3 million words), collecting data from news and blogs;
- *Araneuum Russicum*, a general web-crawled corpus ([Benko, 2014; Benko and Zakharov 2016], 20 billion words); and
- *Taiga*, open source corpus of Russian texts for machine learning tasks ([Shavrina, 2018], 6 billion words).

Most of these corpora are accessible via web interface but do not give access to the source data (*RNC*, *GICR*), yet some can be also downloaded

[3] Hereinafter, IPMs are given from General Internet-Corpus of Russian, VK subcorpora.

for independent research needs (*Aranea*, *OpenCorpora*, *Taiga*). Despite the differences in the theoretical standards of tokenization, morphology and syntax annotation, indexing algorithms, all of these corpora can potentially be reduced to a single format, as shown, e.g., in [Lyashevskaya et al. 2017].

Research based on corpus data requires an accurate understanding of the ratio of sources, genres of texts, their dates of origin, the ratio of authors and so on [Lyashevskaya and Sharoff 2008] — in this case, the researchers often compose a text sample by themselves, choosing materials in the necessary relationship from existing corpora, and adding new resources. Open resources like *Wikipedia*, news, *Common Crawl* data are often joining the task as they are easy to obtain.

To stop the repetition of the same work done by researchers, and to bring all data to a single standard for ease of processing and analysis, we propose to combine the existing open resources within a single project. The following sections will shortly describe the components envisaged as parts of a merged corpus resource.

3. Wikipedia

Wikipedia is a collaborative project of creating a free encyclopedic resource covering over one hundred different languages, where the Russian *Wikipedia* belong to the largest ones, totaling over 1.5 million articles. Due to its open license and rich metadata (including links among language versions), *Wikipedia* data is often used in NLP-related research and development activities.

Full and up-to-date *Wikipedia* materials are not available in a "corpus-ready" format and so they had to be prepared separately for our Project. For various reasons, we opted for downloading the data in a page by page manner via an API by means of the *wiki2corpus*[4] script. It was expected to be operational "out-of-the-box", i.e., without any additional programming necessary, and it produced output file directly compatible with the rest of our processing pipeline, providing also full metadata. The disadvantage of the solution was a somehow slow pace of getting the data. To prevent blocking the download process, a user-settable delay (one second by default) is imposed after downloading each article, making the average speed approx. 2,000 articles per hour.

The process yielded (after basic filtration and deduplication) approximately 500 million tokens of corpus data.

[4] http://corpus.tools/wiki/wiki2corpus

4. Taiga

Taiga is an open source corpus constructed in the way of maximum recall of the crawled resources. *Taiga* corpus unites such text sources as news, social media, fiction, poetry, chat logs and some relatively small amount of special datasets: fake news, authorship attribution collections, parallel corpora in English and German, resulting to 6 billion of words with open access[5] and full metadata available, i.e., date and web source of origin, info about the author, theme, genre, etc. Both morphological and syntactic annotation has been performed by *UDPipe*.

Taiga is the only Russian web corpus with billions of words with genre annotation tagged by authors of the respective texts themselves (sourcing *proza.ru* and *stihi.ru*). The corpus is based on the methodology of differential web corpus construction [Shavrina 2018], meaning all the unique resources having entered into the data composition are crawled entirely to reach full representativeness within the segment.

5. Araneum Russicum

The *Aranea* project[6] is targeted at creating a family of web-crawled corpora for approx. two dozens of languages using a so-called "Brno Pipeline"[7]. All corpora are processed by a unified set of tools and by uniform methodology and are available in two basic sizes (1 billion and 100 million tokens, respectively). For some languages, even larger corpora are attempted, with Russian being one of them. The size of the largest Russian corpus is approx. 25 billion tokens, yielding almost 20 billion tokens after paragraph and sentence level deduplication.

The corpus is available in a vertical format containing morpho-syntactic annotation by *TreeTagger* and light XML markup containing metadata provided by the *SpiderLing crawler*[8]. As the early versions of the software did not store all metadata items, IP-addresses and page titles are available only for data crawled during the latest sessions.

[5] Corpus is available for download under "fair-use" conditions.
[6] http://unesco.uniba.sk/guest/
[7] http://corpus.tools/
[8] http://corpus.tools/wiki/SpiderLing

6. Common Crawl

Common Crawl is an open source of petabytes crawled web pages, including mainly text data from various web sources — blogs, ads, news, automatically generated content, etc. Typically, about 8.5 % of all downloaded materials in the world in *Common Crawl* is Russian. On the Fig. 1 the size on various aggregation levels (host, domain, top-level domain / public suffix) is shown.

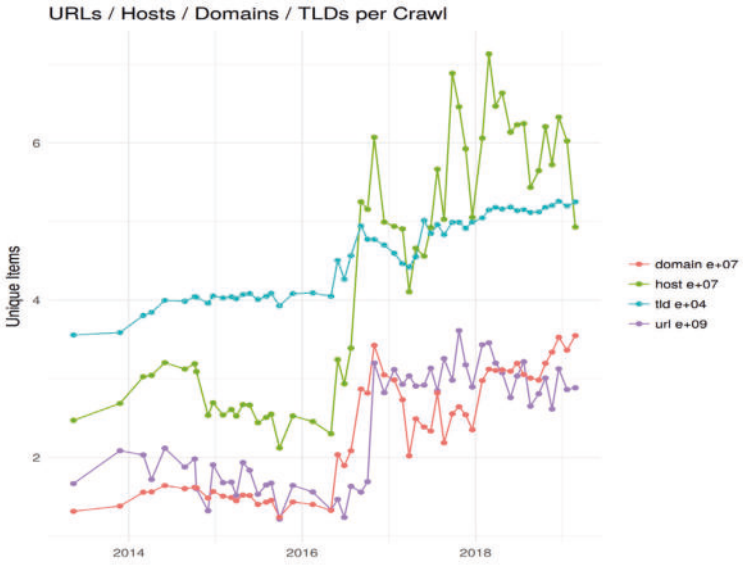


Fig. 1. **Common Crawl**

From the point of view of data collecting, *Common Crawl* is an ideal source to be gradually scaled without significant changes in its structure and proportion of segments. On the other hand, this resource should be extremely painstakingly cleaned of noise, spam, and duplicates, considering duplicates of all resulting corpus segments.

7. The Omnia Russica Project

During the first phase of the Project, three components that are readily available will be merged. The Project, however, is envisaged as "open", i.e. more data is expected to be included in the near future. The name of the

joint corpus was chosen to be *Omnia*[9] *Russica*. The situation with the respective components is shown in Table 1.

Table 1. ***Omnia Russica*** **components**

	Format	Morphology	Syntax	Size
Wikipedia	vertical	TreeTagger	None	0.5 G
Taiga	CoNLL-U	UDpipe	UDpipe	4.5 G
Araneum Russicum	vertical	TreeTagger	None	25 G
Common Crawl	Plain text	None	None	3 G

For merging the data, the following principles will be applied:

- Metadata contained in the respective parts will be preserved;
- The primary format will be (*No*)*SkE*-compatible vertical, the *CoNLL-U* format can be obtained by conversion if necessary;
- All data will be tagged both by *UDPipe* and *TreeTagger* at first, with possible more annotations added in the future;
- Both logical (e.g. paragraphs) and linguistic (e.g. sentences) structures contained in the respective corpus parts will be preserved;
- Both original and deduplicated versions will be available;
- For online access, the corpus will be processed to be accessible by *NoSketch Engine*;
- The users of the merged source data will have to respect the license conditions applied the individual corpus parts.

At the time of writing this paper (May 2019), all respective data has been collected and preprocessed, except the *Common Crawl* part, that is still being scaled to the limit. Now, the *Wikipedia* and *Araneum Russicum* corpora need to be retagged by *UDPipe*, and *Taiga* by *TreeTagger* to get a uniform format suitable for subsequent merging and optional deduplication; *Common Crawl* part should be also deduplicated and processed. The resulting vertical format will consist of 15 columns, i.e., five resulting from the *Aranea* pipeline annotation[10] (*word*, *lemma*, *atag*, *tag* and *ztag*), followed by ten *CoNLL-U*[11] columns provided by UDPipe (*ID*, *FORM*, *LEMMA*, *UPOS*,

[9] *omnis*, pl. *omnia* is the Latin expression meaning "everything".
[10] http://unesco.uniba.sk/aranea_about/index.html
[11] https://universaldependencies.org/format.html

XPOS, *FEATS*, *HEAD*, *DEPREL*, *DEPS*, *MISC*) and XML markup containing metadata.

We believe to be able to demonstrate the results of the first phase of the Project online during the Conference. As a "teaser", Appendix shows the statistics of the Taiga processed by NoSketch Engine.

8. Conclusion

In this article, we described the process of combining modern corpus resources into a single data bank, that may potentially become the largest and most representative corpus of the Russian language ever existed in open source.

Modern language technologies require an increasing amount of textual resources to get better summarization of the contexts and better representation the low-frequency part of the lexicon. We invite researchers to join and offer their resources for open access (both offline and online), as well as to take advantage of the proposed data and develop open source technologies and explore rare language phenomena on *Omnia Russica*.

References

1. *Benko V.* (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014, pp. 257–264.
2. *Benko V., Zakharov V.* (2016) Very Large Russian Corpora: New Opportunities and New Challenges. In Kompjuternaja lingvistika i intellektualʹnyje technologii: Po materialam meždunarodnoj kon-ferencii «Dialog» (2016), vypusk 15 (22). Moskva: Rossijskij gosu-darstvennyj gumanitarnyj universitet, 2016, pp. 79–93.
3. *Bocharov V. V., Alexeeva S. V., Granovsky D. V., Protopopova E. V., Stepanova M. E., Surikov A. V.* (2013) Crowdsourcing morpho-logical annotation. In Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue 2013, Vol. 12(19), Moscow.
4. *Devlin J., Chang M.-W., Lee K., Toutanova, K.* (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Under-standing. ArXiv e-prints.
5. *Kilgarriff A.* (2001). Comparing corpora. International Journal of Corpus Linguistics 6(1): 97–133.
6. *Kilgarriff, A.* (2012). Getting to know your corpus. In Text, Speech and Dialogue, 3–15. Berlin Heidelberg: Springer.
7. *Lample G., Denoyer L, Ranzato M.* (2018). Unsupervised ma-chine translation using monolingual corpora only. ICLR, 2018.

8. *Lyashevskaya O. N., Plungian V. A., Sichinava, D. V.* (2005). O morfologicheskom standarte Korpusa sovremennogo russkogo jazyka [Morphological standard of the Corpus of contemporary Russian]. In Nacional'nyj korpus russkogo jazyka: 2003–2005 [Russian National Corpus: 2003–2005], pp. 111–135, Moscow.
9. *Lyashevskaya O., Sharoff S.* (2008) Frequency Dictionary of the Russian National Corpus: Principles and Technology. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2008" Moscow
10. *Lyashevskaya O., Bocharov V., Sorokin A., Shavrina T., Gra-novsky D., Alexeeva S.* (2017). Text collections for evaluation of Russian morphological taggers. J. Linguist./ Jazykovedný časopis 68(2), 258–267 (2017).
11. *Mikolov T., Chen K., Corrado G, Dean J.* (2013). Efficient Est-imation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
12. *Pennington J, Socher R., Manning C. D.* (2014). GloVe: Global Vectors for Word Representation.
13. *Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L.* (2018). Deep contextualized word representations. In NAACL
14. *Radford A., Narasimhan K., Salimans T., Sutskever I.* (2018). Improving language understanding with unsupervised learning. Technical report, OpenAI.
15. *Selegey, D., Shavrina, T., Selegey, V., Sharoff, S.* (2016). Automatic morphological tagging of Russian social media corpora: training and testing. In Computational linguistics and intellectual technologies. Proceedings of International Workshop "Dialogue 2016", Moscow.
16. *Shavrina,* T. O. (2018) Differential Approach to Web-Corpus Construction. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018" Moscow, May 30-June 2, 2018
17. *Shavrina T.* (2019) Word vector models as an object of linguistic research. In Computational linguistics and intellectual technologies. Proceedings of International Workshop "Dialogue 2019", Moscow.
18. Zipf G. K. (1949). Human Behavior and the Principle of Least Effort. Addison-Wesley Press, 1949. pp. 484–490.

Shavrina Tatiana Olegovna
National Research University Higher School of Economics & Sberbank
(Moscow, Russia)
E-mail: rybolos@gmail.com

Vladimír Benko
Slovak Academy of Sciences & Comenius University in Bratislava (Slovakia)
E-mail: vladimir.benko@juls.savba.sk

Appendix

Taiga Russica (19.05) 4.44 G

Taiga (filtered & deduplicated; version 0.0.00; build #r001)

Counts	
Tokens	4,442,524,751
Words	3,453,456,606
Sentences	377,085,307
Paragraphs	247,119,323
Documents	9,728,056

General info	
Corpus description	Document
Language	Russian
Encoding	UTF-8
Compiled	05/21/2019 20:02:00
Tagset	Description

Lexicon sizes	
word	19,830,619
lemma	19,011,902
atag	14
tag	940
ztag	6
lc	16,385,434
lemma_lc	16,207,966

Subcorpora statistics

Subcorpus	Tokens	Words	%
Arzamas	450,712	~ 350,367	0.01
Fontanka	69,095,350	~ 53,712,203	1.55
Interfax	8,050,421	~ 6,258,103	0.18
KP	7,064,517	~ 5,491,697	0.15
Lenta	8,445,425	~ 6,565,165	0.19
Magazines	229,004,225	~ 178,019,526	5.15
NPlus1	2,114,136	~ 1,643,452	0.04
Social	13,693,793	~ 10,645,054	0.30
Subtitles	98,805,291	~ 76,807,627	2.22
proza	2,900,088,379	~ 2,254,422,863	65.28
stikhi	1,105,712,502	~ 859,540,544	24.88

МОРФОЛОГИЯ И СИНТАКСИС

MORPHOLOGY AND SYNTAX

А. М. Галиева, Ю. Н. Елезарова

A. M. Galieva, Yu. N. Elezarova

ГРАММАТИКАЛИЗАЦИЯ РЕЧЕВОГО КОНВЕРБА *ДИП* В ТАТАРСКОМ ЯЗЫКЕ (НА КОРПУСНЫХ ДАННЫХ)

GRAMMATICALIZATION OF THE TATAR *DIP* SPEECH CONVERB (ON CORPUS DATA)

Аннотация. В статье на корпусных данных анализируются особенности грамматикализации речевого конверба *дип* в татарском языке. Тема интересна тем, что синхроническое исследование позволяет увидеть разные аспекты (по существу, этапы) десемантизации глагола речи *дию* 'сказать, говорить'. Комбинации слова *дип* с глаголами разных семантических классов показывают основные направления развития грамматикализации – функционирование в качестве цитатива или подчинительного союза. Анализ корпусного материала показал, что наиболее часто слово *дип* относится к глаголам речи и интеллектуального действия, вводя соответствующие клаузы, кодирующие чужую речь.

Ключевые слова: грамматикализация, татарский язык, семантические классы глагола, корпусные данные.

Abstract. The paper discusses some aspects of grammaticalization of the Tatar *dip* speech converb. The topic is interesting because a synchronic study discloses different features (essentially, stages) of desemantization of the speech verb *diyü* 'to say'. Combinations of *dip* with verbs of different semantic classes show the main directions of development of grammaticalization – functioning as a quotative or a subordinating conjunction. Corpus data evidences that the word *dip* is most often related to speech and mental verbs, which results in entering corresponding sentences encoding direct and indirect speech.

Keywords: grammaticalization, converb, the Tatar language, verb semantic classes, corpus data.

1. Введение

Структурно-типологические особенности агглютинации, свойственные тюркским языкам (например, использование падежных аф-

фиксов и послелогов в качестве средства подчинения зависимых предложений главному), в значительной степени освобождают эти языки от использования как сочинительных, так и подчинительных союзов. Имеющиеся союзы, как правило, являются заимствованными из других языков или «представляют собой исконные формы (например, местоимения), выполняющие союзные функции под влиянием синтаксического строя индоевропейских языков» [Гузев, Бурыкин 2007]. Тем более интересны процессы грамматикализации, приводящие к образованию служебных лексем, наблюдаемые в настоящее время, в частности, на материале конвербов.

Конструкции с конвербами в тюркских языках представляют собой обширный неоднородный пласт образований с разной семантикой, степенью грамматикализации и лексикализации [Гращенков 2015]. В статье рассматриваются особенности грамматикализации конверба от глагола речи *дию* ‘сказать, говорить, называть’ на данных Татарского национального корпуса «Туган тел» (http://tugantel.tatar/). Глаголы речи являются типологически широко распространенным источником грамматикализации, являясь базой для формирования показателей эвиденциальности и подчинительной связи [Толдова, Сердобольская 2014, Lehmann 2015]. Нас в данной статье интересует, как связаны особенности грамматикализации конверба *дип* и семантический класс глагольного предиката, к которому он относится. Текстовые примеры взяты из Татарского национального корпуса «Туган тел» ((http://tugantel.tatar/)), перевод примеров выполнен авторами статьи; омонимия в корпусе не снята.

2. Семантика и употребление речевого глагола *дию*

Татарский глагол *дию* (основа *дij-*, где *j в* конце слога может сливаться с гласной *и*) ‘сказать, говорить’ вводит прямую (1) или косвенную речь (2):

(1) *Алыгыз, —**диде** кыз, коры гын*а (З. Махмуди). ‘Берите, — **сказала** девушка сухо’.

(2) *Ярамый, **диде** бит* (Г. Гильманов). ‘Он же **сказал**: нельзя’.

Еще одно значение глагола *дию* — ‘называть’, напрямую обусловлено переосмыслением прямой речи:

(3) *Андый кешене халыкта* ***көйсез дилэр*** (Ф. Чанышева). ‘Такого человека в народе **называют капризным**’ (буквально: ‘говорят «капризный»’).

Глагол *дию* не может управлять существительными, обозначающими речь и речевые высказывания, то есть такие конструкции, как *сказать слово, произнести гласный звук,* с *дию* всегда будут интерпретироваться как чужая речь (например, как *сказать: «Слово»*). Морфологически глагол *дию* имеет все стандартные финитные и нефинитные формы (см. Табл. 1), при этом его отдельные формы используются для выражения отношения говорящего к высказыванию: *ди* ‘мол, якобы’, *диярлек* ‘можно сказать’ и некоторые другие. В толковых словах татарского языка эти единицы, в том числе и *дип*, не выделены в отдельные словарные статьи, а представлены внутри словарной статьи глагола *дию* [Ганиев 2015], что свидетельствует о том, что эти единицы мыслятся как особые случаи употребления глагола *дию*, но не как самостоятельные лексемы. Как следует из данных, представленных в Табл. 1, количество употреблений конверба *дип* в корпусе составляет 47,6 % от общего количества употреблений глагола *дию* (по лемме).

Таблица 1. **Распределение некоторых форм глагола *дию***

Форма глагола *дию*	**Количество употреблений в корпусе**
Претерит	263248 (17,1 %)
Перфект (включая омонимичные причастия на -ган)	212978 (13,8 %)
Настоящее время	233492 (15,2 %)
Имя действия на -у	6300 (0,41 %)
Инфинитив	6744 (0,44 %)
Конверб *дип*	730839 (47,6 %)
Конверб *диеп*	5375 (0,35 %)
Всего по лемме	**1536926 (100 %)**

В современном татарском языке синхронно существуют два фонетических варианта конверба от глагола *дию*: стандартный *диеп* и редуцированный *дип*, они отличаются частотностью использования (см. Табл. 1), принципиальных различий между их функционированием мы пока не выявили.

Грамматикализация конверба *дип* описана в литературе. Так, в Татарской грамматике (1993) *дип* называется послеложно-союзным словом и отмечается, что *дип* «как послелог, непременно следует за подчиненным словом и, как союз, не управляет формой подчиненного слова и относит придаточное предложение к главному» [Татарская грамматика 1993: 368 — 370]. О. В Ханина на материале мишарского диалекта татарского языка также показывает амбивалетный характер слова *дип*: с одной стороны, оно ведет себя как подчинительный союз (присоединение актантной предикации к матричному глаголу), с другой стороны, поведение слова *дип* позволяет говорить о нем как о деепричастии глагола *дип* с полноценной моделью управления [Ханина 2007].

С. Ю. Толдова и Н. В. Сердобольская выделяют основные стратегии грамматикализации глаголов речи, описанные в типологических исследованиях: грамматикализация глаголов речи в показатели цитации, образование подчинительных союзов, развитие показателей эвиденциальности [Толдова, Сердобольская 2014: 115]. Различные формы от татарского глагола *дию* показывают развитие всех трех этих стратегий: конвербы могут вводить прямую и косвенную речь, а также обстоятельства и придаточные с обстоятельственным значением. Кроме того, производные от глагола *дию* могут маркировать косвенную эвиденциальность (пересказывательность), в частности, для этого используется дериват *ди* (в словарях может отмечаться как вводное слово или частица):

(4) *Ник пылау ашамый ул халык, дип әйтте* ***ди*** *Бохар әмире, халык ачыкканны ишеткәч* (А. Баян). ‘**Якобы** бухарский эмир, услышав о голоде, сказал: «Почему этот народ не ест плов?»’

Грамматикализация глагола речи может происходить следующим образом*:* одна из форм определённого речевого глагола начинает регулярно употребляться при передаче чужой речи совместно с другим полнозначным глаголом речи, утрачивая автономность и превращаясь в показатель цитации [Толдова, Сердобольская 2014], именно это мы наблюдаем в татарском языке при сочетании конверба *дип* с другими глаголами речи. При этом сохраняются признаки, характерные для цитации, например, аффикс принадлежности 1-му лицу, относящийся к субъекту глагола речи (5), и императив (6):

(5) *Әниem* (POSS_1SG) ***дип әйтте*** (Г. Апсалямов). ‘Назвал <моей> мамой’.

(6) *Кулларын чишегез* (IMP_2SG) ***дип әмер бирде*** *лейтенант* (М. Амирханов). ‘Лейтенант приказал развязать («развяжите») ему руки’.

3. Глагольные конструкции со словом *дип* в корпусе

Из Татарского национального корпуса «Туган тел» были извлечены конструкции *дип* + *синтетический глагол* (всего в корпусе обнаружено 567000 контекстов, содержащих такие конструкции). Таб. 2 представляет распределение семантических классов глагольных предикатов в 100 наиболее частотных конструкциях с конвербом *дип* (конструкции были размечены вручную). Как показывают данные, наиболее часто *дип* встречается перед глаголами речи. Второй по частотности класс — глаголы интеллектуальной деятельности.

Таблица 2. **Конструкции *дип* + глаголы разных классов**

Класс глагола	Количество в корпусе	Количество в процентах
Глаголы речи	186006	49,3
Глаголы инт. д-ти	112878	29,9
Глаголы эмоций	22620	6,0
Глаголы движения	21240	5,63
Глаголы бытия	10002	2,65
Другие	24396	6,47

Таблица 3 представляет распределение 10 наиболее частотных глаголов речи и интеллектуальной деятельности с конвербом *дип*.

При глаголах речи и интеллектуальной деятельности *дип* выполняет функцию цитатива, вводя придаточную клаузу, указывающую на содержание речи или мысли (примеры (5–7)).

(7) *Шушы бер-ике көндә хәл ителер* ***дип уйлыйм*** (А. Тимергалин). ‘**Думаю**, все должно решиться за эти дни’.

Среди наиболее частотных глаголов имеются также глаголы эмоций, движения, бытия (см. Табл. 4). При глаголах эмоций *дип* обычно вводит клаузу (подаваемую как внутреннюю речь/мысль), раскрывающую причину эмоции:

(8) *Укытучылар сорар* ***дип курыкты*** (Л. Ихсанова). ‘**Испугался, что** учителя спросят’.

Таблица 3. **Употребление *дип* с глаголами речи и интеллектуальной деятельности**

Глагол речи	Количество в корпусе	Глагол интеллект. деятельности	Количество в корпусе
Әйтү ‘сказать’	41058	*Уйлау* ‘думать’	49361
Атау ‘назвать’	24456	*Санау* ‘считать’	16998
Сорау ‘спросить’	22141	*Белү* ‘знать’	10716
Язу ‘писать’	19187	*Ышану* ‘верить, полагать’	9830
Сөйләү ‘рассказать’	15563	*Исәплә̆ү* ‘считать’	7051
Кычкыру ‘кричать’	13617	*Аңлау* ‘понимать’	6867
Өстәү ‘добавить’	3180	*Карау* ‘рассматривать’	5124
Кую ‘вставить’	2874	*Өметләнү* ‘надеяться’	5068
Пышылдау ‘шепнуть’	2321	*Бәяләү* ‘оценивать’	2988
Кабатлау ‘повторить’	2189	*Тану* ‘признавать’	2514

Таблица 4. **Употребление *дип* с глаголами движения и эмоций**

Глагол движения	Количество в корпусе	Глагол эмоции	Количество в корпус
Йөрү ‘ходить’	9798	*Курку* ‘бояться’	3910
Килү ‘приходить’	3051	*Көлү* ‘смеяться’	2346
Китү ‘уходить’	1172	*Борчылу* ‘беспокоиться’	1831
Бару ‘идти’	960	*Елмаю* ‘улыбаться’	1507
Чыгу ‘выходить’	894	*Елау* ‘плакать’	1098
Уту ‘проходить’	804	*Сөенү* ‘радоваться’	1014
Керу ‘входить’	739	*Шатлану* ‘радоваться’	653

Слово *дип с* глаголами движения обычно вводит клаузу или конструкцию, указывающую на цель или причину действия. Поскольку данная клауза представляет собой внутреннюю речь, разграничение

того, цель это или причина действия, вводимого глагольным предикатом, формально может быть не выражено и происходит на уровне интерпретации:

(9) *Көчтер агайда берәр дару юк микән* ***дип кергән идем*** (Т. Гиззат). ‘Зашел узнать, нет ли у дяди Кучтура какого-нибудь лекарства?’

Наиболее часто при обозначении цели используются следующие типы конструкций:

1. Инфинитив на *-ырга* + *дип:*

(10) *Дусларымны* ***сыйларга дип*** *кайттым* (Р. Батулла). ‘Я вернулся, чтобы угостить друзей’.

2. Послелог ***өчен*** ‘для, чтобы’ + *дип*:

(11) *Без, абый кеше, акча* ***өчен дип*** *килмәдек килүен …* (А. Тимергалин). ‘Дядюшка, мы пришли не ради денег…’

В примерах (10–11) *дип* может быть опущен.

Отрицательный коррелят конверба *дип — димичә* —характеризуется относительно малой частотой употребления (обнаружено всего 113 контекстов в корпусе) и сохраняет основную семантику глагола *дию* — значения ‘сказать, говорить’ и ‘называть’.

Грамматикализация конверба глагола *дию* происходит с фонетическим упрощением: используется редуцированный вариант *дип* (вместо стандартного *диеп*; ср. соответствующие конвербы созвучных глаголов *кию* ‘надеть’, *тию* ‘тронуть’ — *киеп, тиеп*).

4. Заключение

Анализ корпусных контекстов свидетельствует о том, что основные способы грамматикализации конверба *дип* — функционирование в качестве цитатива или подчинительного союза —во многом определяется семантическим классом глагола, к которому он относится. Наиболее часто слово *дип* относится к глаголам речи и интеллектуального действия. «Размывание» исходного значения конверба сопровождается с расширением его лексической сочетаемости (он может сочетаться с очень большим числом глагольных предикатов при выражении цели или причины), неречевые предикаты во многих случаях допускают использование обстоятельственных слов (конструкций) вместо прямой или косвенной речи. Процесс грамматикализации конверба *дип* про-

должается в настоящее время, и четкое разграничение того, цитатив это или подчинительный союз, во многих случаях затруднено, несмотря на то, что могут быть выделены некоторые формальные признаки для такого разграничения (например, переходность глагола, которому относится *дип*, структура зависимой клаузы и т. п.), что требует дальнейшего исследования.

Литература

1. *Ганиев Ф. Ә.* (ред.) (2005), Татар теленең аңлатмалы сүзлеге.— Казан.
2. *Гращенков В. П.* (2015), Тюркские конвербы и сериализация: синтаксис, семантика, грамматикализация. М.
3. *Гузев В. Г., Бурыкин А. А.* (2007), Общие строевые особенности агглютинативных языков, Acta linguistica Petropolitana. Труды ИЛИ РАН, Т. 3, Ч. 1. СПб., с. 109–117.
4. *Закиев М. З.* (ред.) (1993), Татарская грамматика, Т. 2. Морфология. Казань.
5. Татарский национальный корпус «Туган тел» (2018). URL: http://tugantel.tatar/ (дата обращения: 01.05.2019).
6. *Толдова С. Ю., Сердобольская Н. В.* (2014), Глагол речи *manaš* в марийском языке: особенности грамматикализации, Вопросы языкознания, 6, с. 109–134.
7. *Ханина О. В.* (2007), Конструкции с грамматикализованным конвербом глагола речи // Мишарский диалект татарского языка: Очерки по синтаксису и семантике. Казань; с. 126–140.
8. Lehmann, C. (2015). Thoughts on grammaticalization. Berlin: Language Science Press.

References

1. *Ganiev F. A.* (ed.) (2005), Tatar Explanatory Dictionnary. Kazan.
2. *Grashchenkov V. P.* (2015), Turkic converbs and serialization: syntax, semantics, grammaticalization. [Tyurkskie konverby i serializatsiya: sintaksis, semantika, grammatikalizatsiya]. Moskow.
3. *Guzev V. G., Burykin A. A.* (2007), General constructional features of agglutinative languages [Obshchie stroevye osobennosti agglyutinativnykh yazykov, Acta linguistica Petropolitana, V. 3, Part 1 St. Petersburg, pp. 109–117.
4. *Khanina O. V.* (2007), Constructions with the grammaticalizated speech converb [Konstruktsii s grammatikalizovannym konverbom glagola rechi], Mishar dialect of the Tatar language: Essays on syntax and semantics [Misharskij dialekt tatarskogo yazyka Ocherki po sintaksisu i semantike], Kazan, pp. 126–140.
5. Lehmann, C. (2015). Thoughts on grammaticalization. Berlin: Language Science Press.
6. *Toldova S. Yu., Serdobolskaya N. V.* (2014), The verb of speech *manaš* in the Mari language: peculiarities of grammaticalization, [Glagol rechi *manaš* v marijskom yazyke osobennosti grammatikalizatsii], Voprosy yazykoznaniya, 6, pp. 109–134.
7. 7 "Tugan tel" Tatar National Corpus (2018). URL: http://tugantel.tatar/
8. *Zakiev M. Z.* (ed.) (1993), Tatar grammar [Tatarskaya grammatika], V. 2. Kazan.

Галиева Альфия Макаримовна
Академия наук Республики Татарстан (Россия)
Galieva Alfiia
Tatarstan Academy of Sciences (Russia)
E-mail: amgalieva@gmail.com

Елезарова Юлиана Николаевна
Санкт-Петербургский государственный университет
Elezarova Yuliana
St. Petersburg State University
E-mail: yuliana.elezarova@mail.ru

П.И.Ли
P.I.Li

МОДЕЛИРОВАНИЕ ИМЕН СУЩЕСТВИТЕЛЬНЫХ ТУНДРОВОГО НЕНЕЦКОГО ЯЗЫКА ДЛЯ ЗАДАЧ МОРФОЛОГИЧЕСКОГО ПАРСИНГА

MODELING TUNDRA NENETS NOUNS FOR MORPHOLOGICAL PARSING TASKS

Аннотация. Доклад посвящён разработке морфологического парсера для корпуса эпических произведений на тундровом ненецком языке. Рассматривается морфологическая структура имён существительных в ненецком языке и возможность её моделирования на базе программы для документации и анализа текстов, составления корпусов FieldWorks Language Explorer. Тундровый ненецкий язык относится к языкам агглютинирующего типа, однако аффиксы и основы повергаются сильному алломорфированию, что представляет трудность для морфологического парсинга. Пути решения: точное моделирование морфологической структуры каждой части речи и введение фонологических окружений для алломорфов.
Ключевые слова. существительное, тундровый ненецкий язык, морфологическая модель, парсинг, FieldWorks Language Explorer, разметка текста, морфотактика, морфонология.

Abstract. The aim of this report is to show how a morphological parser for the corpus of Tundra Nenets epic folklore could be constructed. Morphological structure of the Tundra Nenets nouns and the possibility of modeling one using software for language documentation and analysis FieldWorks Language Explorer are taken into account. Tundra Nenets is agglutinative, but the processes of allomorph formation are very common and difficult for parsing. The solutions are: precise modeling of the morphological structure and creating a system of phonological environments.
Keywords. noun, Tundra Nenets, morphological model, parsing, FieldWorks Language Explorer, interlinear texts, morphotactics, morphonology.

Введение

Для лингвистических исследований по тундровому ненецкому языку, для уточнения художественного перевода на русский язык был создан корпус текстов на тундровом ненецком языке. Основной источник текстов — «Фольклор ямальских ненцев» [Янгасова, Кошкарёва 2018]. Корпус состоит из письменных текстов эпических фольклорных произведений — *сюдбабц*, по своей структуре является двуязычным, параллельным, глубиной выравнивания является слово. Для создания корпуса используется программа для анализа языков FieldWorks Language Explorer, разработка Летнего института лингвистики (SIL International). Выбор программы FieldWorks Language Explorer обусловлен тем, что программа подходит для анализа и до-

кументирования «малоресурсных» (low-resource languages) и малоизученных языков (less-studied languages) [Lockwood 2015]. В FieldWorks Language Explorer морфологический парсер xAmple работает по принципу «единицы в окружении» (item-and-arrangement approach): программа справа налево обращается к поверхностной структуре слова и предлагает возможные разметки, то есть, в каждой лексической статье необходимо перечислять все возможные алломорфы и условия их дистрибуции. Моделируются следующие морфологические концепты языка: 1) система категорий (частей речи). Часть речи выбирается из каталога, основанного на GOLD Anthology. К каждой части речи могут быть добавлены слоты аффиксов, слоты аффиксов могут быть составлены в шаблоны аффиксов (порядок следования слотов). В FieldWorks Language Explorer существует встроенный список морфосинтаксических свойств, как и части речи, основанный на GOLD Anthology; 2) лексические статьи. Формообразующие аффиксы могут быть введены в слоты аффиксов определённых частей речи, деривационные аффиксы характеризуются способностью изменять часть речи слова; 3) правила сложения; 4) фонемы. Фонемы делятся на естественные классы, например, гласные и согласные. Могут образовывать виды окружения; входить в фонологические свойства; 5) виды окружения, которые могут быть использованы для распределения алломорфов. Для корректной работы морфологического парсера в программе FieldWorks Language Explorer составляется модель тундрового ненецкого языка в категориях GOLD Anthology и на материале текстов эпических фольклорных произведений.

1. Имена существительные тундрового ненецкого языка

Имя существительное — чётко выделяющаяся на основе морфосинтаксических и словообразовательных признаков часть речи в тундровом ненецком языке.

1.1. Морфотактика

В тундровом ненецком языке существует всего два префикса [Буркова, Кошкарёва, Лаптандер, Янгасова 2010: 232]. Основными маркерами формообразующих и деривационных значений являются суффиксы и клитики. К формообразующим именным аффиксам относятся:

1) интраклитики (слот аффиксов CL), присоединяются непосредственно к основе до других слотов аффиксов:

Таблица 1. **Пример интраклитик в корпусе**

Аффикс и его глосса (CL)	Алломорфы аффикса по грамматике и по корпусу	Пример в словоформе корпуса
-хава	-хова, -хав, -хавы, -хэва, -хэв, -гава, -гав, -кава, -кав	мя=**кав**=’’
-ри	-ли	мала=**ри**=н’

2) аффиксы падежа и числа (слот аффиксов CASE:NUM) могут присоединяться напрямую к основе, могут стоять после интраклитик. Для примера приведём фрагмент падежной парадигмы:

Таблица 2. Пример **аффиксов падежа и числа в корпусе**

Аффикс и его глосса	Алломорфы аффикса по грамматике и по корпусу	Пример в словоформе корпуса
ø NOM.SG		нися=
-’ GEN.SG	-м, -ӈ, -ø, -н	тика=**н**=да
м’ ACC.SG	-м	ед=**м’**
-н’ DAT-LAT.SG	-н, -д’, -т’, -нд	по=**н’**
-хана LOC-INSTR.SG	-хана, -хона, -гана, -кана, -’’на, -нгана, -хан, -кан, -хэна, -ӈгана, -хиня	тэ=**хэна**=т

Сложные формы местных падежей двойственного числа, образованные маркерами падежа и послелога, рассматриваются как единый аффикс на основании передачи единого грамматического значения. Формы винительного падежа множественного числа (ACC.PL) образуются с помощью особого чередования в основе. Для моделирования этого грамматического значения были введены имена основ: NOM.SG и ACC.PL. Каждой основе имени существительного в Лексиконе было приписано собственное дополнительное имя основы ACC.PL с соответствующим алломорфом, например: *ненва* — основа NOM/SG, *ненби* — основа ACC/PL. Для образования форм родительного и продольного падежа множественного числа имени основы ACC.PL были заданы свойства «Согласование в падеже: родительный, продольный;

множественное число»: [caseagr:[case:gen lesscommoncasesgroup:prolat num:pl]]. Соответствующие грамматические свойства были приписаны аффиксам родительного и продольного падежа множественного числа.

3) посессивные или лично-притяжательные аффиксы (слот аффиксов POSS), располагается за слотом падежа и числа имени существительного.

Таблица 3. Примеры **лично-притяжательных аффиксов единственного числа обладаемого в корпусе**

Аффикс и его глосса	Алломорфы и варианты аффикса	Пример в словоформе корпуса
-ми NOM.SG.POSS.1SG	-ми’ (диалектный вариант), -в (нераспределённый вариант)	ню=**ми**
-ми ACC.SG.POSS.1SG	-ми’ (диалектный вариант), -в (нераспределённый вариант)	толи=**ми**
-ни GEN.SG.POSS.1SG	-ни’ (диалектный вариант)	ӈавна=**ни**

Снятие омонимии происходит за счёт введения соответствующих грамматических свойств. Лично-притяжательные аффиксы именительного падежа единственного числа обладаемого напрямую присоединяются к основе, лично-притяжательные аффиксы винительного падежа: посредством элемента *–м*, лично-притяжательные аффиксы родительного падежа — к показателю *–н*, местные падежи — к показателю местного падежа+*-н*+лично-притяжательный аффикс родительного падежа. Лично-притяжательным аффиксам приписываются соответствующие свойства: ACC.SG.POSS: [caseagr:[case:acc]]; GEN. SG.POSS: [caseagr:[case:gen]]. В лексическую статью ACC.SG *–м’* был добавлен алломорф *–м* со свойством [caseagr:[case:acc]]; алломорф *–н* [caseagr:[case:gen]] со свойством был добавлен в статью GEN.SG.

4) дестинативные аффиксы (слот аффиксов DEST): *-да*/ *-та*, слот занимает место между слотом интраклитик и слотом падежа и числа, указывают на планируемое обладание предметом [Буркова, Кошкарёва, Лаптандер, Янгасова 2010: 256].

После последовательного рассмотрения всех возможных слотов формообразующих аффиксов в FieldWorks Explorer было составлено

несколько шаблонов, где слоты были расположены в порядке расположения в словоформе, а в скобках указывалось обязательность/необязательность слота:

- имя существительное абсолютного склонения: основа, (интраклитика), падеж и число;
- имя существительное посессивного склонения: основа, (интраклитика), падеж и число, лично-притяжательный аффикс;
- имя существительное дестинативного склонения: основа, (интраклитика), дестинативный аффикс, падеж и число, лично-притяжательный аффикс.

Деривационные аффиксы располагаются после корня и распадаются на две группы: не меняющие частеречной принадлежности имени существительного (внутрикатегориальное словообразование) и аффиксы, образовывающие имена существительное от глагольной основы (межкатегориальное словообразование).

Кроме того, для распределения алломорфов формообразующих аффиксов были моделированы несколько типов формообразующих классов существительных на основе [Бармич 1999]. В дальнейшем алломорфы аффиксов вручную группировались по сочетаемости с тем или иным классом:

Таблица 4. Пример **распределения именных основ и алломорфов аффиксов**

Название	Описание	Основы	Аффиксы
1B	основы оканчиваются на согласный (но не на гортанный смычный)	107	29
1A	основы оканчиваются на гласный	194	28

Следующим шагом в описании дистрибуции алломорфов для парсера стало введение фонем тундрового ненецкого языка, фонологических свойств, естественных классов и видов окружения.

1.2. Морфонология

Фонологическая система тундрового ненецкого языка моделировалась по [Salminen 1997: 31–44], фонемы системы были поделены на естественные классы, принадлежность к которым играет роль при алломорфировании:

Таблица 5. **Некоторые классы фонем**

Класс	Фонемы	Название класса
GS	**h**, **q**	Гортанные смычки
OC	**b**, **by**, **cy**, **d**, **dy**, **h**, **k**, **p**, **py**, **q**, **s**, **sy**, **t**, **ty**, **x**, **c**	Шумные согласные
SC	**l**, **ly**, **m**, **my**, **n**, **ny**, **ŋ**, **r**, **ry**, **w**, **ÿ**	Сонорные согласные

На основе выделенных классов фонем были составлены виды фонологического окружения, влияющих на алломорфирование:

Таблица 6. Примеры **видов окружений, влияющих на алломорфирование**

Вид	Название	Количество
/#_	Начало слова	29
/_#	Конец слова	18
/_[OC]	Перед шумными согласными	84
/_[SC]	Перед сонорными согласными	86
/_[V]	Перед гласными	19

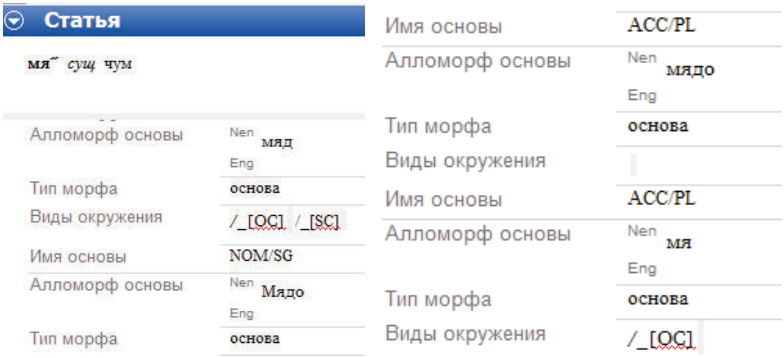


Рис. 1. **Пример лексической статьи с дистрибуцией алломорфов по виду окружения**

Таким образом, существующие лингвистические описания тундрового ненецкого языка позволяют построить подробные модели частей речи в FieldWorks Language Explorer. Они могут быть использованы для морфологической и синтаксической разметки текстов, для уточ-

нения перевода на русский язык, для создания собственной системы перевода. Кроме того, модели позволяют выявить ещё не описанные в существующей литературе диалектологические и стилистические особенности фольклорных текстов на ямальском говоре тундрового ненецкого языка.

Список условных обозначений морфологических показателей

ABL — отложительный падеж, ACC — винительный падеж, CL — клитика, DAT-LAT — дательно-направительный падеж, DEPRIV — именной аффикс со значением необладания, DEST — дестинативный аффикс, DIM — уменьшительный аффикс, EMP — эмфатический аффикс, GEN — родительный падеж, LOC-INSTR — местно-творительный падеж, NOM — именительный падеж, PL — множественное число, POSS — лично-притяжательный аффикс, PROLAT — продольный падеж, Q — вопросительный аффикс, REPOSIT — вместилище, SG — единственное число, VN — отглагольное имя.

Литература

1. *Black H. A.* (2017) A Conceptual Introduction to Morphological Parsing for Stage 1 of the FieldWorks Language Explorer. SIL International.
2. *Black H. A., Simons G. F.* (2006), The SIL FieldWorks Language Explorer Approach to Morphological Parsing. Computational Linguistics for Less-studied Languages: Proceedings of Texas Linguistics Society 10, 3–5 November, Austin.
3. *General Ontology for Linguistic Description* (*GOLD*) (2010), Bloomington, Department of Linguistics (The LINGUIST List), Indiana University, URL: http://linguistics-ontology.org/ (27.02.2019).
4. *Lockwood R. M.* (2011) Machine Parsing of Gilaki Verbs with Fieldworks Language Explorer. SIL International. URL: https://software.sil.org/fieldworks/support/technical-documents/ (27.02.2019)
5. *Lockwood R. M.* (2015), A Linguist-Friendly Machine Translation System for Low-Resource Languages (thesis). University of Washington, Seattle.
6. *Salminen T.* (1997), Tundra Nenets Inflection. Mémoires de la société finno-ougrienne [Reports of Finno-Ugric society], 227, Helsinki.
7. *Бармич М. Я.* (1999) Ненецкий язык в таблицах и схемах. СПб.
8. *Буркова С. И., Кошкарёва Н. Б., Лаптандер Р. И., Янгасова Н. М.* (2010) Диалектологический словарь ненецкого языка. Екатерин бург.
9. *Терещенко Н. М.* (2008) Ненецко-русский словарь. СПб.
10. *Янгасова Н. М.* (*сост.*)., *Кошкарёва Н. Б.* (*ред.*) (2018), Фольклор ямальских ненцев. СПб.

References

1. *Black H. A.* (2017) A Conceptual Introduction to Morphological Parsing for Stage 1 of the FieldWorks Language Explorer. SIL International.
2. *Black H. A., Simons G. F.* (2006), The SIL FieldWorks Language Explorer Approach to Morphological Parsing. Computational Linguistics for Less-studied Languages: Proceedings of Texas Linguistics Society 10, 3–5 November, Austin.
3. *General Ontology for Linguistic Description* (*GOLD*) (2010), Bloomington, Department of Linguistics (The LINGUIST List), Indiana University, URL: http://linguistics-ontology.org/ (27.02.2019).
4. *Lockwood R. M.* (2011) Machine Parsing of Gilaki Verbs with Fieldworks Language Explorer. SIL International. URL: https://software.sil.org/fieldworks/support/technical-documents/ (27.02.2019)
5. *Lockwood R. M.* (2015), A Linguist-Friendly Machine Translation System for Low-Resource Languages (thesis). University of Washington, Seattle.
6. *Salminen T.* (1997), Tundra Nenets Inflection. Mémoires de la société finno-ougrienne [Reports of Finno-Ugric society], 227, Helsinki.
7. *Barmich M. Ja.* (1999), Nenetskij yazyk v tablitsah i shemah [Nenets Language in Templates and Patterns]. Saint-Petersburg.
8. *Burkova S. I., Koshkarjova N. B., Laptander R. I., Jangasova N. M.* (2010), Dialektologicheskij slovar' nenetskogo jazyka [The Dictionary of Nenets Dialects]. Ekaterinburg.
9. *Tereschenko N. M.* (2008), Nenetsko-russkij slovar'. Saint-Petersburg.
10. *Jangasova N. M* (*red.*). *Koshkarjova N. B,* (*ed.*) (2018), Fol'kl'or jamal'skih nentsev [Yamal Nenets Folklore]. Saint-Petersburg.

Ли Полина Игоревна
Новосибирский государственный университет (Новосибирск, Россия)
Институт филологии Сибирского отделения РАН (Новосибирск, Россия)
Li Polina
E-mail: polina.li.14@mail.ru

М. В. Кобозева, Д. Б. Писаревская, А. А. Тугутова, С. Ю. Толдова
M. V. Kobozeva, D. B. Pisarevskaya, A. A. Tugutova, S. Yu. Toldova

ПРОБЛЕМЫ РАЗМЕТКИ КОРПУСА ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ В ТЕРМИНАХ ТЕОРИИ РИТОРИЧЕСКИХ СТРУКТУР: ИЗ ОПЫТА СОЗДАНИЯ RU-RSTREEBANK[1]

ISSUES ON ANNOTATION OF CORPUS WITHIN RHETORICAL STRUCTURE THEORY FRAMEWORK FOR RUSSIAN LANGUAGE: THE CASE OF RU-RSTREEBANK DEVELOPMENT[1]

Аннотация. В статье предложено описание создания дискурсивного корпуса русского языка, размеченного в рамках Теории риторической структуры. Описаны особенности применения данной теории для русского языка, а также проблемы, возникавшие в процессе разметки.
Ключевые слова. Дискурсивная структура, разметка корпуса, теория риторических структур.

Abstract. The article is devoted to the description of the discourse corpus of the Russian language development. The corpus is annotated within the Rhetorical structure theory. The features of the application of this theory for the Russian language are described, as well as the problems encountered in during the annotation.
Keywords. Discourse structure, corpus annotation, Rhetorical structure theory.

1. Введение

Анализ текста на уровне дискурса отражает такое основополагающее свойство текста, как связность, так как в ходе данного анализа выявляются связи между фрагментами текста, их порядок, контекст и окружение. Одним из наиболее популярных подходов к анализу дискурсивной структуры текста является Теория риторической структуры (ТРС, Rhetorical Structure Theory) [Mann, Thompson 1988], которая была адаптирована для самых разных языков, что говорит о её универсальности. Разработка корпусов с риторической разметкой является актуальной задачей как в теоретической лингвистике, так и в области автоматической обработки текста. В данной статье предложено описание корпуса Ru-RSTreebank — первого открытого русскоязычного корпуса письменных текстов с дискурсивной разметкой в рамках ТРС.

ТРС предлагает описание организации текста в виде иерархического дерева дискурсивных единиц (сегментов текста), которые соединяются риторическими отношениями (РО). Минимальной дискурсив-

[1] При частичной поддержке РФФИ, проект № 17-29-07033

ной единицей (ДЕ) в базовом варианте является клауза. Объединяясь одним РО, две единицы становятся новой единицей следующего уровня в иерархии, которая в свою очередь также вступает в РО. Таким образом максимальной ДЕ является весь текст. Каждая единица дискурса входит в РО по крайне мере с одной другой единицей и является узлом дискурсивного дерева. Внутри отношения единицы дискурса могут быть «ядрами» (сообщать более важную информацию) или «сателлитами» (сообщать дополнительную информацию). В зависимости от этого отношения между ними могут быть одноядерными (Детализация, Причина-Следствие и др.) и мультиядерными (Последовательность, Контраст, Конъюнкция и др.). Авторы РРС оставляют возможность варьировать список отношений — в разных работах он может существенно различаться (количество отношений может достигать 80-ти), однако базовый набор включает 23 отношения.

На основе опыта создания корпусов для других языков (для английского, немецкого, японского и др.) в рамках теории РРС [Carlson et al. 2003; Stede 2004; Da Cunha 2011] был создан корпус с дискурсивной разметкой для русского языка. Доклад посвящён описанию корпуса, а также обсуждению различных этапов его создания, проблем разметки.

2. Корпус Ru-RSTreebank

Корпус включает в себя тексты разных жанров. Это обусловлено гипотезой о том, что тексты разных типов будут различаться между собой в своей структуре. Первая часть — 79 текстов в жанрах новостных статей и новостной аналитики, научно-популярных статей. Вторая часть — 100 научных текстов из научной электронной библиотеки «Киберленинка» (https://cyberleninka.ru/): 50 текстов по филологии и лингвистике и 50 текстов по техническим и компьютерным наукам. Корпус включает в себя 203 287 словоупотреблений. Разметка корпуса проводилась несколькими аннотаторами. Использовался инструмент разметки — rstWeb [https://corpling.uis.georgetown.edu/rstweb/info/]. Он позволяет редактировать сегментацию текста, создавать свой список РО, вносить дополнения в инструментарий. В рамках проекта, в rstWeb были добавлены новые функции: вычисление частотного распределения типов отношений, введение иерархии отношений, специальное выделение абзацев для удобства сегментации по элементарным ДЕ, подсветка выбранных типов связей для удобства разметки структуры.

3. Работа над разметкой корпуса

Опыт разработки RST-корпусов показывает, что из-за различий в грамматической системе языков необходима адаптация теории TPC для конкретного языка, поэтому вначале мы уточнили понятие ДЕ и состава РО для русского языка. В результате этой работы была создана инструкция для разметки русскоязычных текстов.

Работа над инструкцией проходила в два этапа. На первом этапе была составлена версия для разметки новостных и научно-популярных текстов.

В первой части инструкции представлены определение и принципы выделения элементарной дискурсивной единицы (ЭДЕ). В определении ЭДЕ мы опираемся на синтаксический принцип: в качестве ЭДЕ используется клауза (предикация), в соответствии с классическим подходом В. Манна и С. Томпсон. Однако мы формулируем ряд важных дополнений с учётом специфики русского языка. Например, придаточные определительные и причастные обороты выделяются в отдельные ЭДЕ в зависимости от их семантики. Описательные (аппозитивные) определения, которые вносят дополнительную информацию, образуют отдельную ЭДЕ, в то время как ограничительные (рестриктивные) определения — нет.

Кроме того, мы выделяем несколько типов ЭДЕ, которые не обладают глагольной составляющей:

- предложные группы со значением причины, следствия, уступки и контраста, содержащие номинализацию (например, предлоги *из-за, для, с учетом, несмотря на*): [*Несмотря на свойственную возрасту впечатлительность,*] [*Прокофьев не злоупотребляет инициалами.*].
- конструкции с номинализациями, в которых риторическое отношение выражено эксплицитно, типа «*X является / стал / был причиной / следствием / свидетельством / и т. д. Y*»: [*Недавняя перестрелка в мечети Газы*] [*стала свидетельством серьёзной напряжённости.*].

Эти и другие принципы разметки, снабжённые рядом примеров, вошли в инструкцию. Во вторую часть инструкции вошли определения и примеры для типов РО, выбранных для разметки. По итогам пробной разметки и экспертных обсуждений, оригинальный набор отношений был несколько изменён. Например, мы объединяем отношения Причина и Следствие в единый тип Причина-Следствие [Pisarevskaya,

Ananyeva, Kobozeva et al. 2017]. Итого было отобрано 17 типов дискурсивных отношений. Это одноядерные отношения Background (Фон), Cause-Effect (Причина), Evidence (Обоснование), Condition (Условие), Purpose (Цель), Concession (Уступка), Preparation (Подготовка), Elaboration (Детализация), Solutionhood (Решение), Attribution (Источник), Interpretation-Evaluation (Интерпретация/Оценка) и мультиядерные отношения Contrast (Контраст), Restatement (Переформулировка), Sequence (Последовательность), Comparison (Сравнение), Same-unit (Прерывающаяся единица), Joint (Соединение).

В рамках второго этапа инструкция была адаптирована для разметки научных текстов. По итогам трёх раундов пробной разметки (трёх, пяти и десяти научных текстов) четырьмя лингвистами-экспертами, в инструкцию были внесены новые дополнения в соответствии со спецификой научных текстов. Во время каждого раунда разметки аннотаторы опирались на последнюю на тот момент версию инструкции. Приведём примеры дополнений, внесённых в первую часть инструкции, посвящённую выделению ЭДЕ.

Для научных текстов характерны сложные предложения с обилием примеров, перечисления, списки. Было уточнено, что иные однородные члены предложения, помимо сказуемых, не могут являться обоснованием для выделения отдельных ЭДЕ. Текстовые примеры на иностранных языках в филологических текстах и формулы/рисунки в технических, независимо от их размера, выделяются как одна ЭДЕ и соединяются с ЭДЕ, которую они иллюстрируют, отношением Детализация.

Что касается текстовых примеров на русском языке, мы используем формальный принцип: элементы нумерованного или маркированного списка выделяются в отдельные ЭДЕ, только если среди них присутствуют номинализации. При отсутствии маркированного списка, примеры объединяются в одну ЭДЕ.

Также ЭДЕ дополнительно выделяются в предложениях характеризации, в которых обычно субъект — это конкретный предмет или класс предметов, а предикат выражает признак, свойство, действие, состояние, отношение к другим предметам и т. п.: [*Для одной из испытуемых была изготовлена маска из термопласта* -] [*особо прочного материала, полностью исключающего возможность «подглядывания»*.].

Во вторую часть инструкции были добавлены принципы, позволяющие выбрать отношения в спорных ситуациях. Так, дискурсив-

ный маркер «*например*» употребляется в отношениях Детализация и Обоснование. Если он отсылает к именной группе внутри ядра (1), то это отношение Детализация, т. е. детализируется некоторый объект/явление (диагностический вопрос: «Какой X?»). Если ко всей клаузе в целом (2) — это отношение Обоснование, т. е. сателлит является подтверждением всего утверждения в ядре (диагностический вопрос: «Чем это подтвердить?»):

(1) [*Некоторые позитивные признаки уже можно заметить.*] [*Например, в июле Международный валютный фонд похвально объявил о существенном увеличении объемов льготного кредитования наименее развитым странам.*]

(2) [*Впрочем, столь высокая цена не отпугнула рекламодателей:*] [*например, Nissan уже заплатил 10 млн.*]

После третьего раунда пробной разметки, окончательная версия инструкции по дискурсивной разметке русскоязычных текстов с учётом специфики научных текстов была утверждена, и четыре аннотаторами осуществили разметку 100 основных текстов в соответствии с ней. Каждый текст размечался двумя аннотаторами. Регулярно проводилась автоматическая проверка согласия аннотаторов. Для её оценки применялась мера Krippendorff's unitized alpha [Krippendorff 2011], позволяющая проводить измерения при любом количестве аннотаторов и в случаях, если разметчики по каким-либо причинам выделили разное количество сегментов текста. В рамках последнего измерения мера составила 81 %, что на данный момент является максимальным значением для данного показателя. На заключительном этапе для каждого текста осуществлялась проверка согласованности разметки, и выбиралась оптимальная разметка, которая была включена в корпус.

Первые версии корпуса (179 текстов) и инструкции для разметки находятся в открытом доступе [http://rstreebank.ru/]. Финальная версия на настоящий момент доступна по запросу.

4. Разбор трудностей при разметке

Можно выделить два типа возникающих трудностей: 1) возможность приписать разные типы отношений одному и тому же фрагменту; 2) последовательность присоединения отношений. Деление на

ЭДЕ, в целом, не вызывало больших разногласий между аннотаторами благодаря четкой инструкции.

Меньше всего спорных случаев возникало с отношениями Условие, Цель, Уступка, Источник и Последовательность. Больше всего разногласий было с одноядерными отношениями Детализация, Обоснование, Интерпретация/Оценка и мультиядерными отношениями Контраст, Сравнение и Соединение. Например, не всегда получается отличить Фон от Подготовки, так как оба отношение несут в себе контекст для следующего предложения: [*В балканских языках существует до 150 сходных элементов. Многими исследователями ощущалось сходство балканских языков не только в грамматике, но и на уровне устойчивых фраз и оборотов речи.*] [*Фразеологические примеры неизменно включаются в фундаментальные балканистические исследования.*].

Вариативность в последовательности присоединения дискурсивных единиц можно проиллюстрировать следующим примером: [*Любой грек <…> может обидеться на человека, который его сделал посмешищем,*] /Конъюнкция/ [*и достаточно сложно будет убедить его в обратном,*] /Причина/ [*потому что греки в большинстве своем чрезвычайно впечатлительны и вспыльчивы*]. Варианты разметки: Причина связывает одновременно две предыдущих ЭДЕ, либо только предшествующую ей.

Частой проблемой становится очерёдность приписывания отношений в самом начале или в конце абзаца. В некоторых случаях отношение Подготовка может быть рассмотрено как относящееся к одному фрагменту из абзаца, а иногда ко всему полностью. При присоединении предложения-вывода в конце абзаца можно сделать это как относящееся к определённому фрагменту, так и предшествующей части целиком.

5. Заключение

Опыт дискурсивной аннотации корпуса на русском языке в терминах TPC показал, что необходима адаптация к особенностям конкретного языка не только правил выделения тех или иных отношений, но и самого списка отношений, а также правил сегментации текста на ЭДЕ. В дальнейшем предложенные принципы разметки будут применяться и к другим жанрам текста (например, к блогам), чтобы в корпусе были представлены тексты различных жанров.

Литература

1. *Carlson L., Marcu D., & Okurowski M. E.* (2003), Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue, pp. 85–112, Springer, Dordrecht.
2. *Da Cunha I., Torres-Moreno J. M., and Sierra G.* (2011), On the development of the RST Spanish treebank. In Proc. 5th Linguistic Annotation Workshop, pp. 1–10, Portland OR
3. *Krippendorff K.* (2011), Computing Krippendorff's Alpha-Reliability. Working Paper, available at http://web.asc.upenn.edu/usr/krippendorff/mwebreliability5.pdf.
4. *Mann W. C., Thompson S. A.* (1988), Rhetorical structure theory: Toward a functional theory of text organization, Text-Interdisciplinary Journal for the Study of Discourse, T. 8, no. 3, pp. 243–281.
5. *Pisarevskaya D., Ananyeva M., Kobozeva M. et al.* (2017), Towards building a discourse-annotated corpus of Russian //Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017), vol. 1, pp. 194–204.
6. *Stede M.* (2004), The Potsdam commentary corpus. In Proceedings of the Workshop on Discourse Annotation, Barcelona, Spain.

Писаревская Дина Борисовна
ФИЦ ИУ РАН (Россия)
Pisarevskaya Dina
FRC 'Computer Science and Control' of RAS (Russia)
E-mail: dinabpr@gmail.com

Кобозева Мария Вадимовна
ФИЦ ИУ РАН (Россия)
Kobozeva Maria
FRC 'Computer Science and Control' of RAS (Russia)
E-mail: marya.kobozeva@gmail.com

Тугутова Ая Андреевна
Московский государственный университет (Россия)
Tugutova Aya
Moscow State University (Russia)
E-mail: atugutova@gmail.com

Толдова Светлана Юрьевна
Высшая школа экономики (Россия)
Toldova Svetlana
Higher School of Economics (Russia)
E-mail: toldova@yandex.ru

Е. Г. Соколова, С. Ю. Толдова
E. Sokolova, S. Toldova

ОСОБЫЕ СВОЙСТВА РИТОРИЧЕСКИХ ОТНОШЕНИЙ «КОНТРАСТ» И «СРАВНЕНИЕ» НА МАТЕРИАЛЕ РАЗМЕТКИ В КОРПУСЕ RU-RSTREEBANK[1]

RELATIONS OF "CONTRAST" AND "COMPARISON" IN TERMS OF T HE RHETORIC STRUCTURE THEORY: ISSUES ON CORPUS ANNOTATION

Аннотация. В докладе обсуждаются проблемы выделения риторических отношений 'Сравнение' и 'Контраст' в терминах Теории риторических структур Манн-Томпсон. Предлагается анализ логических и прагматических оснований данных отношений, а также языковых средств их маркирования. Предложенный анализ позволяет ввести в инструкции аннотирования корпуса уточнения: предложить операциональные критерии, а также очертить круг специфичных для данных отношений маркеров.
Ключевые слова. Теория риторических структур, контраст, сравнение, дискурсивная аннотация.

Abstract. The work is devoted to the detection of the Contrast vs. Comparison relations within the framework of the Rhetoric structure theory Mann-Thomson. The analysis of annotated data in terms of logical or pragmatic constraints is suggested. This analysis makes it possible to suggest some operational criteria for the relations under discussion. These criteria together with the detailed analysis of special markers associated with certain relations can improve the instruction for discourse relations annotation.
Keywords. RST, corpus annotation, relations 'Contrast', 'Comparison'.

1. Введение

1.1. Понятие дискурсивной разметки

В рамках создания корпусов текстов с различными типами разметки все более актуальной становится так называемое «глубокое» аннотирование текстов. На данный момент существует большое количество корпусов с самой разной разметкой различных лингвистических параметров, связанных с характеристиками языковых единиц в пределах одного предложения (ср. проекты по разработке универсальных принципов аннотации грамматических категорий и синтаксических отношений). Более сложной и более актуальной задачей становится аннотация связей между предложениями, т. е. аннотация текстов на дискурсивном уровне.

[1] При частичной поддержке РФФИ, проект № 17-29-07033

Дискурсивные отношения (ДО) устанавливаются между фрагментами текста. В отличие от синтаксических, ДО в принципе не предполагают лексического или грамматического репрезентанта. В каждом конкретном случае ДО выделяются на основе содержания фрагментов текста. Но при этом в тексте, как правило, имеются интуитивно употреблённые говорящим средства, в определённых условиях указывающие на определённый тип ДО. Семантика самого ДО чётко улавливается говорящими и может рассматриваться как значение, создаваемое текстом.

В статье обсуждается проблема выделения двух ДО — контраста (Contrast) и сравнения (Comparison), предлагаются уточнения определений и некоторые практические рекомендации для разграничения этих отношений. Материалом послужили примеры из корпуса Ru-RSTreebank. В статье [Кобозева и др. 2019] подробно рассмотрены определения этих отношений и корпусная статистика, касающаяся средств их выражения по корпусу Ru-RSTreebank. Тем не менее, вопросы выделения этих отношений остаются. В данной статье мы рассматриваем механизмы, лежащие за возникновением этих отношений в тексте, относящиеся, собственно к организации текста. Двухъядерное отношение Contrast определяется в теории на сайте RST http://www.sfu.ca/rst/01intro/definitions.html. Многоядерное отношение Comparison определяется в Инструкции по разметке текстов корпуса Ru-RSTreebank.

2. Comparison

На сайте теории RST отношение Comparison отсутствует. Введение его из других источников, определено в инструкции по разметке корпуса в терминах Теории риторических структур (ТРС) [Mann, Thompson 1988]), которая также подтверждает частотность и важность этого отношения [Carlson, Marcu 2001]. Comparison играет важную практическую роль, например, автоматические системы создают тексты типа Сравнение, организующие сравнение свойств однотипных коммерческих продуктов для потребителя. В этом случае введённое в состав ДО теории RST отношение Comparison можно рассматривать как узкую разновидность отношения Joint, т. е. Joint(Comparison), определяемое целью текста и содержанием смежных фрагментов. Есть и другой вид ДО Comparison, имеющий дополнительные внешние свойства и маркеры, частично рассмотренные нами в статье [Кобозева и др. 2019].

2.1. Верификация Comparison по цели текста

Главный признак текста типа Сравнение — содержательный. Из цели текста и содержания смежных фрагментов должно следовать, что содержание этого текста — сравнение объектов или ситуаций по некоторому параметру. При этом параметр явствует из текста, например:

(1) *Первый заключается в последовательной реализации отдельных приложений, автоматизирующих отдельные процессы управления* (Я1), *второй — во внедрении платформы для реализации комплексной системы автоматизации управления процессами и создании на ее базе приложений, интегрированных в единый комплекс* (Я2) (сравниваемые объекты — *первый* и *второй* (принципы построения системы управления), параметр — организация процессов управления).

Т. е. фрагмент текста типа сравнение может быть представлен перечислением сущностей с указанием особенностей каждой из них по сравниваемому(ым) параметру(ам). Нередко такое перечисление содержит более двух фрагментов, т. е. многоядерно, и выделяется форматированием, например:

(2) • *WebSQL представляет из себя полномерную SQL — базу данных внутри браузера, которая может хранить копии данных веб-приложения для автономной работы, позволяя пользователям продолжить работу с данными даже при потере соединения с сетью. Данные синхронизируются с сервером при последующем подключении к сети* [*9*]. (*Я1*)

• *ApplicationCache дает возможность хранить элементы веб-приложения* (*HTML, CSS и т. д.*) *для их последующего использования в моменты, когда сеть будет недоступна*; (*Я2*)

• *WebStorage основан на механизмах хранения, аналогичных cookies, но при этом представляет собой более гибкую и более мощную их реализацию*; (*Я3*)

Здесь сравниваются компьютерные системы по параметру способа хранения информации.

2.2. Верификация Comparison по лексическому маркеру

Лексические маркеры, в отличие от цели текста, не являются надёжными показателями наличия ДО Comparison. Например:

(3) *Свидания заключенных с родственниками проходят в большом зале со столиками, похожем на кафе* (Я). *По одну сторону столика сидит*

заключенный, по другую семья (С1). *От кафе помещение отличается обилием видеокамер* (С2). В *соседней маленькой комнатке сотрудник постоянно следит за мониторами.*

В примере (3) *кафе* не сравнивается с комнатой для свиданий заключённых, у адресата отсутствует представление о последней. *Кафе* приводится как исходный образ, на основании которого адресат может её представить. Слова *столик* и *помещение* — прямые отсылки к этому исходному образу. Тип связи между выделенными дискурсивными единицами (Я → (С1 и С2)) можно отнести к Elaboration.

По лексическим маркерам можно выделить два типа ДО Comparison: «сравнение-сопоставление» и «сравнение по степени проявления признака».

Во первом случае ДО обычно двухъядерное, например:

(4) *В отличие от английского предложения* (Я1), *во французском предложении он включён вместе с действием в единый концепт, вербализуемый глаголом* (Я2). → (Сравниваемые объекты — структура англ. и структура фр. предложения, параметр — размещение указателя направления движения внутри глагола или в виде отдельного слова.

Сравнение-сопоставление оформляется также маркерами: |*в то время как X (Я1), Y(Я2)*| , |*Y(Я1), в свою очередь X(Я2)*| , |*X* (*Я1*), *тогда как Y*(*Я2*)|, |*X* (*Я1*), *а Y*(*Я2*)|, которые вводят предикации относительно однородных, но разных объектов, например:

(5) *В то время как CO2 остается в атмосфере столетиями,* (Я1) *другие загрязняющие вещества, включая сажу и озон, остаются в ней только на некоторый период — дни, недели, месяцы или годы.* (Я2) (Объекты: *CO2 vs. другие загрязняющие вещества, включая сажу и озон;* параметр — долговечность).

В случае сравнения по степени, лексическими маркерами являются предикаты, указывающие на изменение степени проявления признака, например:

(6) *Фунт и евро немного укрепились* (Я1), а *иена слегка упала* (Я2).

Здесь можно предложить, условно говоря, «перифразирование», выражающее результирующую ситуацию в виде грамматической сравнительной конструкции, в которой эксплицирован параметр сравнения: *Курс фунта и евро* (*стал*) *немного выше, чем иены*. Сравнительная конструкция однозначно свидетельствует о сравнении, а не

о контрасте. Бессмысленность содержания предлагаемой перифразы, (курс, ведь, у каждой валюты индивидуален), не мешает, так как дискурсивный анализ абстрагируется от знаний действительности. Как видно из примеров (4)–(6), ДО сравнение действительно может опираться на лексические маркеры и процессы, с ними связанные: лексический повтор (*английское — французское предложение*), антонимия (*укрепилась — упала*), условная перифраза (*выше*).

3. Contrast

К лексическим маркерам контраста в инструкции отнесены *но, однако, несмотря на то, что*. Слова но и *однако* подробно изучались в качестве грамматических союзов, в последние годы также в качестве дискурсивных слов.

3.1. Верификация Contrast

Семантика союза но подробно рассматривается в [Урысон 2006]. Автор замечает, что союзы по семантике имеют сходства с другими языковыми единицами, в частности, первообразные сочинительные союзы *но, и, а* семантически сближаются с первообразными междометиями — эмоциональными типа *ой!, ай!, ах!* и когнитивными типа *э!, эге!,* Но при этом указывается, что, в отличие от союзов, первообразные междометия употребляются автономно, т. е. образуют высказывание. Однако в некоторых случаях но и *однако* также могут образовывать высказывание, в частности, «*Никаких но!*», «*И…?*», а также знаменитое «*Однако!*» Кисы Воробьянинова. Замечание о семантическом сходстве можно распространить и на дискурсивные слова.

Е. В. Урысон утверждает, что союзу но нельзя дать традиционного толкования, потому что он указывает на некую операцию сознания. Различия между *но*, а и *и* резюмируются следующим образом: «Союз но маркирует, прежде всего, смену ситуации. Союз а маркирует смену объекта, помещаемого в фокус. Кроме того, и союз *а*, и союз но маркируют изменение темы внутри фрагмента повествования. Союз и маркирует отсутствие «переключения сознания». В [Урысон 2006] также показано, что семантика но имеет трёхчастную структуру и предполагает «обманутое ожидание», например;

(7) *День был дождливый*, но *Коля не вымок,*

где P = *День был дождливый;* Q = ‘*в дождливую погоду люди вымокают*’ (ожидание, следует из P); S = *Коля не вымок* (отрицание ожидания Q).

Совершенно аналогичную риторическую фигуру мы видим при установлении ДО Contrast. Оно может быть маркировано дискурсивными словами но и *а*, например:

(8) *В бизнесе сформировались свои, российские, принципы управления, автоматизирующие отдельные процессы управления* (Q'). *Но время идёт* (Я1), а (или *но*) *способы решения проблем остаются на уровне середины 50-ых годов* (Я2).

Отличие дискурсивной структуры от синтаксической состоит в присутствии в тексте фрагмента Q', который и является базой для формулировки ожидания: Q = '*принципы управления формируются в соответствии с имеющимися условиями; условия со временем меняются, соответственно. меняются и принципы управления*'; P = *Но* (это «*но*» можно опустить) *время идёт;* S = а (или *но*) *способы решения проблем остаются на уровне середины 50-ых годов* (отрицание Q).

Риторическая фигура обманутого ожидания в Contrast может строиться и с помощью других дискурсивных слов:

(9) *Такие данные можно добывать штатными силами и средствами, либо получать от компаний, которые работают в этом направлении* (Q'). *Главное, чтобы сведения были достоверными* (Я1). *Получить же достоверную информацию о работе предприятий из финансовой отчётности невозможно, и статистика Госкомстата, таможенного комитета подходит лишь для выявления изменений в динамике рынков* (Я2).

В примере (9) из (Q') следует ожидание Q = '*добываемые у компаний данные достоверны*'; P = *Главное, чтобы сведения были достоверными;* S = *Получить же достоверную информацию о работе предприятий из финансовой отчётности невозможно…* (отрицание Q). Дискурсивное слово, маркирующее контраст, здесь — частица *же*.

Из сказанного следует, что значение ДО Contrast в рассмотренных примерах так же, как и в случае союза но в [Урысон, 2006] — «операция сознания», другими словами — риторическая фигура, имеющая трёхчастную структуру. Условия и лексические средства её выражения требуют дополнительного исследования.

Заключение

В статье показано, что возможно движение от интуитивной разметки ДО к дальнейшему уточнению и описанию более конкретных

контекстных и прагматических ситуаций и языковых средств, организующих данное ДО. Это исследование начато в инструкции по разметке корпуса Ru-RSTreebank, которая уточнялась после каждого этапа создания корпуса [Кобозева и др. 2019]. Основания для такой работы даёт материал самого корпуса, как это показано и в данной статье.

Литература

1. *М. В. Кобозева, Д. Б. Писаревская, А. Тугутова, С. Ю. Толдова.* (2019), Проблемы разметки корпуса текстов на русском языке в терминах теории риторических структур: из опыта создания Ru-RSTreebank (в печати).
2. *Carlson L., Marcu D.* (2001), Discourse tagging reference manual. Technicalreport, InformationScience Institute, Marina del Rey, CA.ISI-TR-545.
3. *Урысон Е. В.* (2006) Семантика союза НО: данные языка о деятельности сознания, ВЯ, №5, 2006.
4. *Mann W. C., Thompson S. A.* (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8, 3, pp. 243–281.

References

1. *Kobozeva M., Pisarevskaya D., Tugutova A., Toldova S.* (2019), Issues on annotation of corpus within Rhetorical Structure Theory for Russian language: the case of Ru-RSTreebank (in print).
2. *Carlson L., Marcu D.* (2001), Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA.ISI-TR-545.
3. *Uryson E. V.* (2006) SemantikasoyuzaNO: dannyjejazykaodeyatel'nostisoznaniya, VJa, [Semantics of the conjunction NO 'but': the data of the language concerning the functioning of the consciousness] no. 5, 2006.
4. *Mann W. C., Thompson S. A.* (1988), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text 8, 3, pp. 243–281.

Соколова Елена Григорьевна
Свободный исследователь (Россия)
Sokolova Elena
Freeresearcher (Russia)
E-mail: minegot@rambler.ru

Толдова Светлана Юрьевна
Доцент Национального исследовательского университета
«Высшая школа экономики»
Toldova Svetlana
Associate professor of National research University "Higher School of Economics"
E-mail: toldova@yandex.ru

Б. Ньеки
B. Nyéki

ИЗВЛЕЧЕНИЕ ПРОИЗВОДНЫХ СЛОВ ИЗ КОРПУСОВ: СБОР ДАННЫХ ДЛЯ ИССЛЕДОВАНИЯ НЕМЕЦКИХ ДИМИНУТИВОВ

EXTRACTION OF DERIVED WORDS FROM CORPORA: DATA COLLECTION FOR THE RESEARCH OF GERMAN DIMINUTIVES

Аннотация. Корпусные исследования могут способствовать изучению продуктивных словообразовательных правил. В докладе излагается подход к решению проблем, связанных с такими исследованиями. Рассматривается фильтрация результатов корпусного поиска с помощью обратного словаря и частеречного анализатора, а также восстановление леммы мотивирующих слов дериватов, основанное на правилах. Данные методы были использованы для извлечения диминутивов из корпуса немецкого языка DWDS.
Ключевые слова: словообразование, продуктивность, фильтрация, мотивирующее слово, морфемный разбор.

Abstract. Corpus research can contribute to the study of productive word formation rules. In the present paper, an approach to the solution of problems connected with such research is explained. The filtration of results of corpus query with the help of a reverse dictionary and a part-of-speech tagger and also the rule-based reconstruction of the base of the derivatives are described. These methods were used to extract diminutives from the DWDS German corpus.
Keywords: word formation, productivity, filtration, base of word formation, morphemic analysis.

1. Сбор корпусных данных для исследования словообразовательной системы языка

1.1. Основная идея

Использование современных электронных корпусов может способствовать исследованиям, посвященным словообразованию. Операторы поиска позволяют быстро находить токены по аффиксам. После фильтрации результатов пользователь может сохранить ценный материал — совокупность предложений, содержащих производные слова выбранного типа. К ним можно добавить лингвистическую разметку, отсутствующую в корпусе.

В настоящем докладе рассматриваются преимущества сбора производных слов из корпусов (1.2.–1.3.), а также излагаются методы извлечения, фильтрации и дополнительной разметки диминутивов на *–chen* (отыменные существительные среднего рода, как например: *Töchterchen* 'доченька', *Mützchen* 'шапочка' и т. п.) из корпуса немецкого

языка DWDS (2.1.–2.3.). Основные результаты эксперимента, основанного на этих методах, указаны в разделе 3.

1.2. Изучение продуктивности

Продуктивность словообразовательного правила рассматривается с точки зрения корпусной лингвистики Х. Байеном [Baayen 2009]. Он выделяет три показателя продуктивности, основанных на статистической характеристике. Первый показатель — реализованная продуктивность, которая характеризуется количеством изучаемых производных слов в корпусе известного объема. Показатель продуктивности по мере распространения — это доля исследуемых дериватов среди всех неологизмов в корпусе. Количество токенов, встречающихся один раз в корпусе (hapax legomena), считается Х. Байеном приближенным значением частоты неологизмов. Похожим образом оценивается потенциальная продуктивность, вычисляемая делением количества hapax legomena, образованных данным правилом, на частоту всех производных того же типа.

Совершенно ясно, что только потенциальную продуктивность можно оценить с опорой исключительно на извлеченные из корпуса предложения, содержащие дериваты. Однако их частотная характеристика является важной информацией для приблизительного вычисления остальных показателей продуктивности.

Значение продуктивности для компьютерной лингвистики заключается в том, что с помощью продуктивного правила образуются неологизмы, отсутствующие в словаре морфологического анализатора. Поэтому, как рекомендует В. П. Захаров [2016], целесообразно включить продуктивный аффикс в словарь анализатора.

1.3. Автоматическое выделение морфем

В настоящий момент существуют алгоритмы, обеспечивающие процесс автоматического разбиения слова на морфемы. Morpho project (http://morpho.aalto.fi/projects/morpho/) предлагает алгоритмы, основанные на машинном обучении без учителя. Например, в [Creutz, Lagus 2002] излагается метод создания лексикона морфем на основе неразмеченного обучающего текста (Morfessor Baseline), а в [Creutz, Lagus 2005] — более развитый вариант алгоритма, учитывающий частоты морфем в обучающем тексте. Одна из реализаций алгоритмов проекта — пакет Polyglot (https://polyglot.readthedocs.io/en/latest/#). Для обучения программы, разбивающей текст на морфемы, были ис-

пользованы списки 50 тыс. самых частотных слов разных языков. Нетрудно убедиться в том, что Polyglot разбирает некоторые дериваты заметно хуже, чем другие слова. Например, немецкие диминутивы на *–lein* обычно анализируются неправильно, сам суффикс разбивается на *–lei* и *–n*. Это, наверное, объясняется тем, что в словарях или в списках самых частотных слов редко встречаются диминутивы на *–lein*. Возможно, использование множества извлеченных из корпуса разных дериватов с контекстом в качестве обучающего текста улучшило бы разборы.

2. Опыт извлечения диминутивов на –chen из корпуса немецкого языка

2.1. Выбор корпуса

Эксперимент извлечения диминутивов на *–chen* (или *–chens в* родительном падеже единственного числа) из корпуса немецкого языка носит иллюстративный характер: предполагается, что изложенные ниже методы применимы к сбору данных обо всех словообразовательных правилах, реализующихся суффиксацией.

Выбор корпуса — нетривиальный шаг в ходе эксперимента. Если целью исследования является обобщенное представление некоего словообразовательного типа, то корпус, из которого извлекаются результаты, должен быть репрезентативен и сбалансирован.

Самый большой электронный корпус немецкого языка — это DeReKo (Deutsches Referenzkorpus, http://www1.ids-mannheim.de/kl/projekte/korpora/). В 2018 году корпус содержал приблизительно 42 млрд токенов. Главная проблема данного корпуса — неравномерная репрезентация разных жанров. Ядро корпуса — публицистические тексты, хотя к корпусу постоянно добавляются тексты, что повышает сбалансированность распределения жанров.

Для того чтобы провести эксперимент, удобнее было пользоваться другим корпусом немецкого языка — DWDS (Digitales Wörterbuch der deutschen Sprache, https://www.dwds.de/). Объем основного корпуса (Kernkorpus), содержащего тексты 20-го века, составляет 120 млн токенов. Преимущество основного корпуса в том, что в нем сбалансированно представлены 4 жанра — художественные, публицистические, научные и другие нехудожественные тексты (см. доли: https://www.dwds.de/d/k-referenz#kern).

Для поиска диминутивов были выбраны десятилетие 1990-1999 гг. и полный корпус 21-го века (Kernkorpus 21), в состав которого входят тексты, появившиеся не позднее 2006 года. Сбалансированность источника диминутивов могло нарушить то, в последнем корпусе (Kernkorpus 21) жанры представлены неравномерно. Диминутивы из устных корпусов еще не были включены в эксперимент.

2.2. Фильтрация результатов

На сайте DWDS нельзя задать такой поисковой запрос, который полностью устранил бы шум в выдаче. Можно заранее исключить из результатов лишь некоторые частотные существительные, оканчивающиеся на *–chen*, но не являющиеся диминутивами.

Дальнейшая фильтрация осуществлялась с помощью обратного словаря [Mater 1970]. Из словаря были выбраны те существительные, которые имеют хотя бы одну словоформу, оканчивающуюся на *–chen*, но сами не являются диминутивами, например: *Tisch* 'стол' → *Tischen* (д. п. м. ч.), *Flasche* 'бутылка' → *Flaschen* (любой падеж м. ч.). Эти словоформы и также диминутивы, потерявшие композициональность значения (*Mädchen* 'девочка', *Stiefmütterchen* 'анютины глазки', *Tastkörperchen* 'тактильное нервное тельце' и т. п.), были добавлены в общий список стоп-слов. Если ключевое слово результата корпусного поиска, экспортированного в формате csv, совпало с любым элементом списка, то данный результат (целое предложение) был удален. Учитывались и частичные совпадения по последним символам ключевого слова ввиду возможных сложных слов. После этого частеречный анализатор TreeTagger присвоил некоторый тег каждому ключевому слову с переведенными в нижний регистр символами. Если оно получило не тег существительного, содержащее его предложение было удалено из результатов. Таким образом, удалось отфильтровать такие слова, как *Suchen* 'поиски' *Glücklichen* 'счастливый человек/счастливые люди' (в косвенном падеже или в и. п. м. ч.). Это субстантивированные формы, поэтому частеречная разметка корпуса не могла их исключить. Некоторые инфинитивные формы были включены в список стоп-слов, поскольку субстантивированные инфинитивы иногда выступают в качестве компонента сложного существительного, например, *Sportmachen* 'занятие спортом'.

После фильтрации из 30 593 результата осталось 4174.

2.3. Восстановление леммы мотивирующего слова

Лингвисту, изучающему производные слова, часто приходится прибегать к мотивирующим словам: встречаются ли они без исследуемого аффикса, и если да, сколько раз, в каких контекстах и т. д. Поэтому целесообразно добавить в лингвистическую разметку информацию о лемме мотивирующего слова (т. е. практически об основе, к которой применяется деривационное правило), которая сделала бы возможным быстрый доступ к его токенам.

Даже безупречная морфемная сегментация не может решить задачу в силу возможности появления алломорфов. Однако для немецких дериватов можно написать правила восстановления мотивирующего слова. Эти правила требуют лингвистических знаний. Например, свойства образования диминутивов с суффиксом *–chen* описаны в [Fleischer, Barz 2012]. Алгоритм восстановления мотивирующего слова должен опираться на сведения об алломорфах (расширенных формах) суффикса, выпадении букв с конца основы при суффиксации и других изменений основы.

Итак, программа сначала строит список гипотетических основ без внутренних изменений. Выделяются три алломорфа суффикса: *–chen*, *–elchen*, *–erchen* (последние встречаются в таких дериватах, как *Blümelchen* 'цветочек' от *Blume*, *Prösterchen* 'чоканье' от *Prost*). Если диминутив *d*, состоящий из *n* символов, оканчивается на *–elchen* или *–erchen*, то порождаются следующие гипотетические формы: *d* до *n*-6 символа, *d* до *n*-6 символа + *e*, *d* до *n*-6 символа + *en*, *d* до *n*-4 символа. Вторая и третья формы нужны ввиду того, что при суффиксации безударные *–e* и *–en* выпадают в конце основы. Однако нет необходимости добавить в список еще вариант «*d* до *n*-4 символа + *e*/*en*», поскольку буквосочетания *–ele*/ *–ere* в конце слова нетипичны для немецких существительных. Если диминутив оканчивается на *–chen, но* не на *–elchen* или *–erchen*, то создаются гипотезы *d* до *n*-4 символа, *d* до *n*-4 символа + *e*, *d* до *n*-4 символа + *en*. Гипотетические формы упорядочены: например, «*d* до *n*-4 символа» чаще приводит к правильной основе, чем «*d* до *n*-4 символа + *en*»; эта форма применима к большему количеству основ, поэтому она получает более высокий ранг. Потом символы *ä*, *ö*, *ü* внутри порожденных форм заменяются на регулярные выражения (*ä*|*a*), (*ö*|*o*), (*ü*|*u*), таким образом устраняются проблемы, связанные с явлением «umlaut».

Проверка наличия гипотетических форм в словаре часто приводила бы к отрицательному результату ввиду весьма продуктивного сло-

восложения существительных в немецком языке. Поэтому программа на каждом шагу цикла добавляет к последнему символу гипотетической формы один символ слева, пока полностью не будет восстановлена данная форма. Наличие строки среди заглавных слов словаря постоянно проверяется. Если соответствующее заглавное слово найдено в словаре, программа сохраняет его, потом продолжается конкатенация. После этих операций выбирается сохраненное слово с максимальным рангом. Если таких слов несколько, то из них выбирается самое длинное. Наконец, оно добавляется к левой, несовпадающей части гипотетической формы (в идеальном случае эта часть представляет собой пустую строку). Получившаяся последовательность символов считается леммой мотивирующего слова. Если ни одно заглавное слово словаря в ходе выполнения алгоритма не было идентифицировано, то мотивирующим словом считается гипотеза «*d* до *n*-4 символа».

Итак, например, имея диминутив *Pferdefigürchen* 'фигурка лошади', можно восстановить мотивирующее слово *Pferdefigur*, даже если в словаре нет заглавного слова *Pferdefigur*, только *Figur*.

В качестве словаря в эксперименте был использован частотный список словоформ корпуса DeReKo 2014 г. (http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html, дата обращения: 13.03.2019). Для каждой словоформы указаны лемма и часть речи. После объединения словоформ, относящихся к одной и той же лемме, список содержал более чем 37 тыс. существительных.

Мотивирующие слова были разбиты на морфемы с помощью пакета Polyglot.

3. Результаты эксперимента

В рамках эксперимента было получено 4174 ключевых слова в контексте одного предложения. Токены дериватов относятся к 1330 разным леммам. 473 из них встречаются более чем один раз. 5 самых частотных лемм — *Päckchen* 'небольшой пакет/пачка' (122 раза), *Städtchen* 'городок' (73 раза), *Kästchen* 'ящичек' (71 раз), *Stückchen* 'кусочек/небольшая часть' (66 раз), *Fläschen* 'бутылочка' (63 раза). Каждое ключевое слово размечено следующими сведениями: лемма, мотивирующее слово и его морфемный разбор. Для оценки качества методов, изложенных в предыдущих разделах, были взяты две выборки объемом 100 и 200 результатов. Более чем 80 % ключевых слов действительно оказались диминутивами. Приблизительно в 75 % случа-

ев лемма мотивирующего слова была восстановлена правильно (если ключевое слово не было диминутивом, результат выполнения данного алгоритма считался ошибочным). Полнота извлечения диминутивов из экспортированных материалов DWDS (30 593 результата) составляет примерно 90 %. Естественно, некоторые диминутивы были искючены вследствие случайного совпадения их последних символов с каким-либо стоп-словом, поэтому нельзя было ожидать стопроцентной полноты. Значение F_1-меры, вычисленное по показателям точности и полноты, — 0,847.

4. Заключение

В перспективе планируются совершенствование методов фильтрации и восстановления леммы мотивирующего слова и вовлечение в исследование DeReKo и устных корпусов. Добавление семантических признаков к разметке результатов также представляется целесообразным. Созданная таким образом совокупность дериватов могла бы служить базой для морфологического и семантического исследования словообразовательных правил.

Литература

1. *Захаров В. П.* (2016), Словообразовательные неологизмы в русском языке (корпусное исследование). Буцева Т. Н. (ред.), Неология и неография: современное состояние и перспективы (к 50-летию научного направления): Сборник научных статей. СПБ, с. 57–63, режим доступа: https://iling.spb.ru/pdf/neologia.pdf (26.02.2019).
2. *Baayen H.* (2009), Corpus linguistics in morphology: Morphological productivity. Lüdeling A., Kyto M. (eds.), Corpus Linguistics. An international handbook. Berlin, pp. 900–919, available at: http://www.sfs.uni-tuebingen.de/~hbaayen/publications/BaayenHSK2009.pdf (26.02.2019).
3. *Creutz, M., Lagus, K.* (2002), Unsupervised discovery of morphemes. *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*. Philadelphia, Pennsylvania, pp. 21–30, available at: https://arxiv.org/pdf/cs/0205057.pdf (28.02.2019).
4. *Creutz, M., Lagus, K.* (2005), Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (*AKRR'05*), Espoo, pp. 106–113, available at: http://research.ics.aalto.fi/events/AKRR05/papers/akrr-05creutz.pdf (28.02.2019).
5. *Fleischer W., Barz I.* (2012), Wortbildung der deutschen Gegenwartssprache. Vierte, neu bearbeitete Auflage. Berlin.
6. *Mater E.* (1970), Rückläufiges Wörterbuch der deutschen Gegenwartssprache. 3. Auflage. Leipzig.

References

1. *Zakharov V. P.* (2016), Slovoobrazovatel'nye neologismy v russkom jazyke (korpusnoe issledovanie). [Word Formation Neologisms in Russian (a Corpus Research)]. Buceva (ed.), Neologija, neografija: sovremennoe sostojanie i perspektivy (k 50-letiju nauchnogo napravlenija): Sbornik nauchnykh statjej. [Neologism, Neography: Recent Conditions and Perspectives (for the 50-year Anniversary of the Scientific Direction): Collection of Scientific Papers]. Saint Petersburg, pp. 57–63, available at: https://iling.spb.ru/pdf/neologia.pdf (26.02.2019).
2. *Baayen H.* (2009), Corpus linguistics in morphology: Morphological productivity. Lüdeling A., Kyto M. (eds.), Corpus Linguistics. An international handbook. Berlin, pp. 900–919, available at: http://www.sfs.uni-tuebingen.de/~hbaayen/publications/BaayenHSK2009.pdf (26.02.2019).
3. *Creutz, M., Lagus, K.* (2002), Unsupervised discovery of morphemes. Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02. Philadelphia, Pennsylvania, pp. 21–30, available at: https://arxiv.org/pdf/cs/0205057.pdf (28.02.2019).
4. *Creutz, M., Lagus, K.* (2005), Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, pp. 106–113, available at: http://research.ics.aalto.fi/events/AKRR05/papers/akrr-05creutz.pdf (28.02.2019).
5. *Fleischer W., Barz I.* (2012), Wortbildung der deutschen Gegenwartssprache. [Word Formation of Contemporary German]. Fourth, revised edition. Berlin.
6. *Mater E.* (1970), Rückläufiges Wörterbuch der deutschen Gegenwartssprache. [Reverse Dictionary of Contemporary German]. Third edition. Leipzig.

Бенце Ньеки
Санкт-Петербургский государственной университет (Россия)
Bence Nyéki
Saint Petersburg State University (Russia)
E-mail: nyeki.bence96@gmail.com

Ю. Тао

Yu. Tao

АНАЛИЗ УНИВЕРСАЛИИ ОТЧУЖДЕНИЯ ПРИ ПЕРЕВОДЕ С РУССКОГО НА КИТАЙСКИЙ НА ПРИМЕРЕ КОНСТРУКЦИЙ СО СЛОВОМ 对 (DUI)

ANALYSIS OF DEFAMILIARIZATION UNIVERSAL IN RUSSIAN-CHINEESE TRANSLATION BY THE EXAMPLE OF 对 (DUI) CONSTRUCTIONS

Аннотация: Статья посвящена проявлению универсалии отчуждения при переводе с русского языка на китайский. В качестве материала исследования использовался параллельный корпус русского и китайского языков. Для проверки гипотезы об увеличении размера конструкций в корпусе переводных текстов по сравнению со сопоставимым корпусом исследовался размер конструкций в переводном языке и в оригинальных китайских текстах. Показано, что гипотеза подтвердилась. Одной из причин универсалии отчуждения является влияние исходного русского языка.

Ключевые слова: предложения с 对; размер конструкций; универмалия отчуждения; европеизация.

Abstract: The article is about the universal of defamiliarization in translation from Russian into Chinese. A parallel corpus of Russian and Chinese was used as the research material. The size of constructions in the translated language and in the original Chinese texts was investigated. To test the hypothesis of increasing the size of structures in the translated corpus was compared with the comparable corpus. The lengths of structures was compared with 对 (dui) in two corpus. The study shows that the hypothesis was confirmed. One of the reasons for the defamiliarization is the influence of the original Russian language.

Keywords: sentences with 对 (dui), size of structures, the universal of defamiliarization, Europeanization.

1. Вводные замечания

Настоящая работа представляет собой корпусное исследование, посвященное изучению универсалий перевода, а именно, универсалии отчуждения применительно к конструкциям со словом 对 в переводном языке.

Универсалии перевода (translation universals) — это закономерности, которые глобально наблюдаются при переводе с одного языка на другой [Baker 1993]. Эти языки принято называть исходный и целевой. При этом утверждается, что целевой переводной язык отличается от целевого непереводного. Универсалии перевода включают в себя такие пары как «упрощение — осложнение», «экспликация — импликация», «доместификация — отчуждение» и т. д.

В переводоведении под универсалиями перевода понимают отдаление переводного языка от целевого непереводного языка, когда переводной язык приобретает особенности, отличные от оригинального целевого языка. И эти особенности могут формироваться под влиянием исходного языка.

Предложения со словом 对 (duì) в текстах научного стиля встречаются достаточно часто. В современном китайском языке иероглиф 对 (и близкое к нему слово 对于 (duìyú)) имеет три значения и три варианта употребления.

对$_1$ — это счетное слово. Используется для обозначения классификации двух понятий, противопоставленных друг другу (по принципу полового различия, по размещению слева и права, по обозначению положительного и отрицательного).

对$_2$ — это имя прилагательное. Имеет значение «подходящий», «нормальный», «правильный».

对$_3$ — глагол или глагол-предлог. Имеет значение «относиться к». Мы исследуем конструкции с 对 именно в этом значении.

2. Отчуждение vs европеизация

Вопрос особенностей перевода на китайский язык вновь приобретает свою актуальность в связи с возникшим в последние годы интересом к европеизации китайского языка [Song Wenhui 2016; Жукаускене, Холдаенко 2015]. В китаистике принято изучать влияние перевода на изменения в целевом языке в рамках европеизации [Wang Li 1985: 478].

Нам хочется обратить внимание на сходство и различие терминов «европеизация» и «отчуждение». Европеизация — это принятие идей и ценностей европейской культуры как истинных, прогрессивных, этически оправданных и этически совершенных [Баженова 2008]. Европеизация китайского языка — это внедрение в китайский язык грамматических правил или лексики европейских языков. Мы считаем, что в этом случае лучше пользоваться терминами «отчуждение» (defamiliarization) или иностранизация (foreignization). Они хорошо укладываются в дихотомию универсалий перевода, являясь противопоставлением термину «доместикация» (или «нормализация»). Доместикацию и отчуждение принято относить к двум стилям перевода, следуя которым автор в языке перевода подлаживается либо под читателя, либо под автора переводимого текста. Доместикация приближа-

ет иностранный текст к духу и культурным ценностям целевого языка, а отчуждение приближает перевод к особенностям иностранного языка и иностранной культуры.

И если понятие «европеизация» относится к общественно-социальной сфере, а также к сфере контактов китайского и иностранных языков, то отчуждение входит в парадигму переводоведения и относится к дескриптивному анализу особенностей переводного языка, который описывает особенности переводного языка, в том числе сравнивая их с характеристиками текстов на исходном языке.

Существует несколько исследований, посвященных универсалии отчуждения. Ху Сяньяо изучал особенности использования местоимения 他 (ta) в переводном китайском языке. Он установил, что: 1) среднее расстояние между двумя местоимениями 他 в переводном тексте меньше, чем в непереводном, что означает частое повторение 他 в переводном тексте; 2) в переводном тексте могут встретиться несколько вхождений местоимения 他, а в непереводном языке оно обычно одно; 3) в переводном тексте встречаются такие предложения, в которых местоимение относится к разным людям, а в непереводном такого не наблюдается. Исходя из этого автор делает вывод о проявлении универсалии «отчуждение».

Тао [2016] на основе корпуса русского и китайского языков показала, что количество сложных предложения в переводном тексте при переводе с русского на китайский больше, чем в непереводных текстах. Это коррелирует с распространенностью сложных предложений в русском языке, в то время как китайский язык тяготеет к простым предложениям.

3. Гипотеза исследования

Универсалии перевода свойственны всем переводным языкам, но различные типы языковых явлений проявляют себя по-разному. Предложения со словом 对 формируют в китайском языке относительно закрытые конструкции. Размещение нужных слов, словосочетаний или минор-предложений[2] допускается лишь между 对 и главным словом конструкции (статические слова или предикаты). Китайский язык очень восприимчив к «размеру» закрытых конструкций, который выражается в количестве лексических единиц, включаемых в конструкцию.

[2] Минор-предложения — придаточные предложения в китайском языке.

При переводе с русского на китайский язык обороты с 对 могут быть переводом отглагольных существительных, предложных или причастных оборотов русского языка. Характерной особенностью данных конструкций в русском языке является их открытость, допускается использование постпозитивных определений с нанизыванием ряда слов в родительном падеже или другие способы для органичного расширения конструкции вправо. Мы сделали предположение, что в переводном языке имеет место тенденция расширения размера конструкций под влиянием исходного русского языка.

Для проверки вышеизложенной гипотезы и было проведено данное исследование. Оно проводилось на материале параллельного корпуса русского и китайского языков (ПК) и сопоставимого корпуса (СК) примерно такого же объема, содержащего оригинальные китайские тексты той же тематики (научные тексты гуманитарной и социальной направленности) [Тао, Захаров 2015].

4. Эксперимент. Длины конструкций

Длина конструкции — это количество слов в рамках одной конструкции. В оборотах с 对 — это количество слов, которые входят в «окно» от 对 (или 对于) до главного слова, размещающегося справа. При этом сам иероглиф 的 не включается в расчет размера конструкции. В указанном корпусе переводных текстов было найдено 614 оборотов с 对 (ipm=392,04), в сопоставимом корпусе примерно такого же размера имеется всего 485 оборотов с 对 (ipm=333,73). Измерение длин конструкций в двух корпусах дало следующие результаты (Табл. 1).

Таблица 1. **Длины конструкций с** 对

	Длина конструкций в словах									**Средняя длина (Σ длин/частота)**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
	Количество конструкций данной длины									
ПК	51	44	56	98	138	125	73	26	3	4,70
СК	53	47	99	109	73	68	26	10	0	3,95

Из информации, приведенной в таблице, следует, что: 1) общее число конструкций с 对 в корпусе переводных текстов превосходит количество конструкций данного типа в сопоставимом корпусе; 2) средний

размер конструкции с 对 в корпусе переводных текстов больше среднего размера конструкции в сопоставимом корпусе; 3) максимальную частотность в переводных текстах имеют конструкции с длиной 4–6 слов, а в исходных — с длиной 3–4. В китайском языке при длине конструкции с 对 больше 4 слов предложение начинает казаться растянутым и усложненным.

5. Анализ конструкций

Проанализируем конструкции с 对 и соответсвующие им исходные предложения.

(1) Кроме того, *понимание феномена холодной войны и ее составляющих* позволяет контрастнее выявить *новизну современного состояния международной безопасности*.

此外，对**冷**战/这一/**特殊**/现象/**及其/内容的**/认识，导致对**当前/国**际/**安全/状况的/新**现象产生不同的理解

(слово 对 и главное слов конструкции выделены **полужирным**)

Пословный перевод: Кроме того, *для* (*dui*) холодной войны/, это особое /явление/ и познание/ ее содержаний/ позволяет/ *для* (*dui*) / современное /новое состояние /международной/ безопасности/ появиться /разные/ понимания.

В примере 1 присутствуют 2 оборота с 对, которым в русском тексте соответствуют постпозитивные определения с использованием ряда слов в родительном падеже. Слова в родительном падеже являются типичным явлением для русского языка. Так как имеется такой наглядный маркер формы как «падеж», то множество определений после имени существительного, как правило, не влияют на степень понимания данного предложения читателем. Переводчики при переводе с русского на китайский, стараясь сохранить набор лексических единиц и логику связей, увеличивают размеры конструкций с 对. В оригинальном китайском языке, как это видно из данных сопоставимого корпуса, процент таких конструкций меньше, чем в переводном языке. Там, где в переводном языке используются длинные конструкции с 对, в оригинальном китайском языке вместо них часто используется постановка определений за пределами конструкции с 对, использование внутри конструкций местоимений и др.

5. Заключение

Предложения с 对 являются типичными для научных текстов формами предложений. Однако в переводных текстах они приобретают черты, не характерные для оригинального китайского языка, а именно, они удлиняются, наследуя лексическое наполнение и логические связи из исходного текста на русском языке. Нами были изучены особенности конструкций с 对 в переводных текстах с точки зрения их частоты и длины конструкции. Было показано, что в среднем они длиннее аналогичных конструкций в сопоставимом корпусе, что подтверждает нашу гипотезу. Широкое использование этих конструкций при переводе с русского на китайский и их характер обусловлены строем русских предложений и тем самым несут черты универсалии отчуждения, которые ранее были обнаружены в рамках исследований по переводу с английского языка на китайский.

Литература

1. Баженова Т. П., Семина В. С. Сущность европеизации русской культуры. Аналитика культурологии. URL: http://cyberleninka.ru/article/n/suschnost-evropeizatsii-russkoy-kultury (09-06-2018)
2. Жукаускене Т. С., Холдаенко И. С. Заимствования в китайском языке: влияние английского языка в условиях глобализации. М.: Языкознание, 2015.
3. Тао Ю., Захаров В. П. Разработка и использование параллельного корпуса русского и китайского языков. Научно-техническая информация. Серия 2. Информационные процессы и системы. 2015. № 4. С. 18–29.
4. Baker M. Francis, & Tognini-Bonellis (eds.). *Text and Technology. In Honor of John Sinclair*. Amsterdam: John Benjamins, 1993, pp. 233–250.
5. Song Wenhui. Europeanization of Chinese Nominal Parallel Structure. Language teaching and research, 2016, No. 2, pp. 80–91.
6. Tao Yu, Jiang Z. Research on Operating Norms Governing Translation on the Basis of Russian-Chinese Corpora with a Case of "чтобы" Clauses in Russian // Вестник Санкт-Петербургского государственного университета. Серия 9, 2016. No. 1, pp. 107–119.
7. Wang Li. Collected works (Volume 2). Modern Chinese Grammar. Shandong Education Press, 1985.

References

1. Baker, M. (1993). Text and Technology. In: Honor of John Sinclair. Amsterdam: John Benjamins, pp. 233–250.
2. Bazhenova, T. P., Semina, V. S. (2008). The essence of the Europeanization of Russian culture [Bazhenova T. P., Semina V. S. Sushchnost' yevropeizatsii russkoy kul'tury]. In: Cultural Analytics [Analitika kul'turologii], (09-06-2018).

3. Song Wenhui. Europeanization of Chinese Nominal Parallel Structure. Language teaching and research, 2016, no. 2, pp. 80–91.
4. Tao Yu, Jiang Z. Research on Operating Norms Governing Translation on the Basis of Russian-Chinese Corpora with a Case of "чтобы" Clauses in Russian // Bulletin of St. Petersburg State University. Series 9, no. 1, pp. 107–119.
5. Tao, Yu., Zakharov, V. (2015). Development and use of a parallel corpus of Russian and Chinese. Scientific and technical information. Series 2. Information processes and systems. No. 4, pp. 18–29.
6. Wang Li. Collected works(Volume 2). Modern Chinese Grammar. Shandong Education Press, 1985.
7. Zhukauskene, T. S., Kholdaenko, I. S. (2015). Borrowing in Chinese: the influence of English in the context of globalization. M.: Linguistics.

Тао Юань
Шэньсийский педагогический университет (Китай)
Tao Yuan
Shaanxi Normal University (China)
E-mail:tao1973@mail.ru

М. В. Хохлова, В. И. Рубинер
M. V. Khokhlova, V. I. Rubiner

К ВОПРОСУ О КОЛИЧЕСТВЕННОМ АНАЛИЗЕ ПРЕДЛОЖНО-ПАДЕЖНЫХ СОЧЕТАНИЙ В РУССКОМ ЯЗЫКЕ НА ПРИМЕРЕ ЗАКОНОДАТЕЛЬНЫХ ТЕКСТОВ[1]

ON QUANTITATIVE ANALYSIS OF RUSSIAN PREPOSITIONAL CONSTRUCTIONS BASED ON LEGISLATIVE TEXTS

Аннотация. В статье рассматриваются предложно-падежные сочетания в русском языке на материале корпуса законодательных текстов. Дается обзор модели «предлог + именная группа» и ее частотность, анализируется длина найденных цепочек, а также их наполнение.
Ключевые слова. Предлоги, предложные конструкции, количественная грамматика, корпусы текстов, русский язык.

Abstract. The paper deals with Russian prepositional constructions based on a corpus of legislative texts. The analysis focuses on the model "preposition + noun phrase" and its frequency, and also pays attention to the length of the constructions.
Keywords. Prepositions, prepositional constructions, quantitative grammar, text corpora, Russian language.

Введение

В современной русистике давно существует потребность в количественном описании лингвистических единиц разных уровней. Частотные словари предоставляют статистическую информацию о лексическом разнообразии того или языка, но при этом также необходимы справочники, которые бы описывали более длинные цепочки (например, типичные словосочетания, синтаксические конструкции или комбинации грамматических признаков). Вопросы создания подобных инструментов ставились в работах ряда авторов (см., например, [Мустайоки 1973; Копотев 2011; Janda, Lyashevskaya 2011]). Так, в статьях М. В. Копотева [Копотев 2008; Копотев 2011] высказывается мысль о создании частотной грамматики русского языка, а также проводится количественный анализ падежной системы русского языка на материале современных корпусов текстов. Рассмотрение конструкций, включающих предлоги, может служить частным случаем грамматики такого рода.

[1] Статья подготовлена при поддержке гранта РФФИ № 17-29-09159 «Квантитативная грамматика русских предложных конструкций».

Материал

В рамках нашего исследования рассматривались предложные конструкции разной длины, под которыми мы понимаем сочетания предлогов с управляемыми ими именными группами в определенных падежах. В качестве материала для исследования были отобраны тексты юридической тематики, отличающиеся своей структурированностью. Общий объем составил 1,5 млн слов.

Нами были проанализированы следующие предложные конструкции: 1) предлог + существительное (к *выполнению*); 2) предлог + местоим. / местоим. прил. + сущ. (*со своими полномочиями*); 3) предлог + местоим. / местоим. прил. + прил. + сущ. (*на каждое физическое лицо*); 4) предлог + прил. + сущ. (*на благотворительные цели*). Таким образом, модель «предлог + именная группа» может включать в себя разные зависимые от существительного лексические единицы.

Анализ предложных сочетаний

Подавляющее большинство примеров приходится на конструкции с первообразными предлогами. Из перечисленных в работе [Сичинава] единиц наиболее частотными оказались следующие простые предлоги (в порядке убывания частоты): *в, на, с* (*со*), *по, о* (*об*), *от, за, для до, к, при, из, после* и *без*. Распределение частот коррелирует с данными словаря [Ляшевская, Шаров 2008].

Сложные предлоги не имеют соответствующей разметки, в корпусе они представлены при помощи комбинации морфологических тегов (например, «предлог + существительное» в *течение* или «наречие + предлог» *одновременно с*), поэтому для их поиска запрос должен был быть изменен. Примерами могут выступать следующие словосочетания: *от имени некоммерческой организации*, *независимо от вида жилищного фонда*.

Конструкция «предлог + существительное»

Наибольшее количество примеров приходится на конструкции с предлогами, управляющими предложным падежом (табл. 1). Явное предпочтение данной падежной форме может также указывать на отличительные особенности законодательных текстов, в которых предлоги активно используются, например, в названиях законов («*О внесении изменений и дополнений*»), документов («*приказ о взыскании алиментов*», «*заявление о выдаче*», «*отчет о деятельности*», «*свидетельство о праве собственности*») и др.

Таблица 1. **Распределение форм существительного по падежам**

Падежная форма	Частота (в ipm)
Gen	8526
Dat	5925
Acc	9948
Abl	6057
Loc	28735

Существительное может зависеть от глагола (22,1 %), существительного (76,0 %) или прилагательного (1,9 %). Количественные отличия между управляющими лексемами разных частей речи являются значимыми (p-value<<0,001).

При анализе данной конструкции не учитывались входящие в состав именной группы местоимения и прилагательные. Примеры подобных предложных сочетаний будут рассмотрены ниже.

Конструкция «предлог + местоимение / местоименное прилагательное + существительное»

Длина данной предложной конструкции варьируется от трех (*со своими полномочиями*) до четырех слов (*ото всех других источников, ко всем своим документам*). В случае цепочки из двух местоименных прилагательных, первое из них или оба относятся к определительным (кванторным): *каждый, любой, другой, иной*. Аналогично приведенной

Таблица 2. **Распределение конструкции «предлог + местоимение / местоименное прилагательное + существительное» по падежам**

Падежная форма	Длина конструкции	
	3 слова	4 слова
Gen	678	3
Dat	309	2
Acc	497	3
Abl	153	0
Loc	2040	6
Общее количество:		

выше модели наиболее частотными оказываются сочетания с формами предложного падежа.

Конструкция «предлог + местоимение / местоименное прилагательное + прилагательное + существительное»

Необходимо отметить, что в использованной нами разметке морфоанализатора TreeTagger разделение между прилагательными и причастиями является весьма условным.

Данная конструкция может включать в себя 4 или 5 слов (табл. 3).

Таблица 3. **Распределение конструкции «предлог + местоимение / местоименное прилагательное + прилагательное + существительное» по падежам**

Падежная форма	Длина конструкции	
	4 слова	5 слов
Gen	144	11
Dat	50	3
Acc	170	17
Abl	79	15
Loc	138	9
Общее количество:		55

Предложные сочетания, состоящие из пяти слов, в большинстве случаев подразумевают вхождение двух прилагательных в состав именной группы (*на одного указанного медицинского работника*, из *иных государственных информационных систем*). Также было зафиксировано 5 случаев с двумя местоимениями (к *одному иному общественному объединению*, *за каждый такой непредставленный расчет*). Были также найдены цепочки иные словоформ, включающие два местоимения, но они находятся вне сферы нашего рассмотрения (в *соответствии с ним иными правовыми актами*, *выделения из него другого юридического лица*).

Конструкция «предлог + прилагательное + существительное»

Данная конструкция является более распространенной в корпусе по сравнению с двумя вышеперечисленными.

Таблица 4. **Распределение конструкции «предлог + прилагательное + существительное» по падежам**

Падежная форма	Длина конструкции			
	3 слова	4 слова	5 слов	6 слов
Gen	2280	249	4	0
Dat	1568	209	9	0
Acc	4748	807	29	3
Abl	2308	637	23	0
Loc	9828	878	94	0
Общее количество:				

Наиболее длинная цепочка данной конструкции включает в себя четыре согласованных прилагательных в винительном падеже (табл. 4), входящих в состав именной группы (в *единую федеральную автоматизированную информационную систему*, *в указанные государственные социальные внебюджетные фонды*). Предложный падеж встречается чаще всего за исключением конструкций максимальной длины.

Подавляющее число примеров приходится на простые первообразные предлоги. Простые производные предлоги встречаются в следующих сочетаниях: *включая обязательную государственную дактилоскопическую регистрацию*, *вследствие сбросов радиоактивных отходов*, *вследствие обстоятельств непреодолимой силы*, *путем аккумулирования бюджетных средств*, *посредством составления специального акта, ниже суммы рублевого эквивалента*, *посредством проведения психофизиологического исследования, относительно перемещенных культурных ценностей*.

Заключение

Настоящее экспериментальное исследование является первым шагом по исследованию предложных сочетаний в текстах разных функциональных стилей и структурных особенностей. В качестве следующего этапа планируется сравнение результатов с другими источниками (например, НКРЯ и корпусами большего объема).

Полученные данные, репрезентирующие реальное языковое употребление, могут найти применение, например, при преподавании русского языка.

Литература

1. *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.
2. *Ляшевская О. Н., Шаров С. А.* (2009), Частотный словарь современного русского языка (на материале Национального корпуса русского языка), М.: Азбуковник.
3. *Мустайоки А.* (1973), Опыт составления частотной грамматики русских существительных, Хельсинки, (рукопись).
4. *Сичинава Д. В.* Предлог. // Русская корпусная грамматика. [Электронный ресурс] URL: http://rusgram.ru (дата обращения 15.05.2019)
5. *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.

References

1. *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.
2. *Kopotev M.* (2008), K postroeniju chastotnoj grammatiki russkogo jazyka: padezhnaja sistema po korpusnym dannym [On Building Russian Frequency Grammar: Case System based on Corpus Data] // Mustajoki A., Kopotev M. V., Birjulin L. A., Protasova E. Ju. (eds.), Instrumentarij rusistiki: korpusnye podhody [Tools for Russian Studies: Corpus Methods], Hel'sinki.
3. *Ljashevskaja O. N., Sharov S. A.* (2009), Chastotnyj slovar' sovremennogo russkogo jazyka (na materiale Nacional'nogo korpusa russkogo jazyka) [Frequency Dictionary of Contemporary Russian based on the Russian National Corpus data]. Moscow: Azbukovnik.
4. *Mustajoki A.* (1973), Opyt sostavlenija chastotnoj grammatiki russkih sushhestvitel'nyh [On Compiling Frequency Grammar of Russian Nouns]. Hel'sinki, (manuscript).
5. *Sichinava D. V.* Predlog [Preposition]. In Russkaja korpusnaja grammatika [Russian Corpus Grammar]. [Online] Available at: http://rusgram.ru (accessed on 15.05.2019).

Хохлова Мария Владимировна
Санкт-Петербургский государственный университет (Россия)
Khokhlova Maria
St. Petersburg State University (Russia)
E-mail: m.khokhlova@spbu.ru

Рубинер Виктория Игоревна
Санкт-Петербургский государственный университет (Россия)
Rubiner Victoria
St. Petersburg State University (Russia)
E-mail: andantino.v@gmail.com

СЕМАНТИКА И ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ КОРПУСОВ

SEMANTICS AND INFORMATION EXTRACTION FROM CORPORA

I. V. Azarova, V. P. Zakharov

TOWARDS A COMPUTATIONAL ONTOLOGY OF RUSSIAN PREPOSITIONS[1]

Abstract. Our aim is to create a corpus-based semantic and grammatical description of prepositional constructions using empiric corpus data. The methodology for processing corpus data and calculating frequency characteristics of prepositional constructions in modern Russian texts is presented. The prepositional values are described in terms of semantic rubrics based on the notion of syntaxemes introduced by G. Zolotova. We investigate prepositional constructions within the specific semantic rubrics: mediative and transitive.

Keywords: Russian prepositional constructions, preposition meaning, corpus statistics, semantic rubrics.

1. Introduction

The paper represents a part of a research project which is aimed at the development of corpus-driven semantic-grammatical description of Russian prepositional constructions. To achieve this goal we carry out four interdependent tasks: 1) the integral description of the Russian prepositional system as an interconnected structure in terms of the sense metalanguage specified according to prepositional meanings; 2) the collection of corpus statistics for pairs "preposition-its meaning" from this structure; 3) the sense representation of prepositional constructions as a function expressed by prepositions from this structure over the unity of their governors and governees; 4) the

[1] The research has been supported by the Russian Foundation for Basic Research under the project No. 17-29-09159 and partly (sect. 2, 3) under the project 17-04-00552-ОГН-А.

exposition of prepositional semantics as a part of syntactic links between classes of content words as a type of the prepositional ontology.

These tasks are challenging due to the prepositional ambiguity which is manifested in selectional preferences of particular prepositions expressing synonymic relations between similar content words. We couldn't rely on abstract scholastic presentation of prepositional meanings because they are the part of the grammar which may be illogical and poorly structured. Therefore, all elements of our description are based on the corpus data: the enumeration of prepositional constructions, variance of their grammatical features, synonymy and near synonymy between them, and so on. This corpus-based semantic and grammatical description of Russian prepositional constructions uses empiric data from various contemporary Russian corpora in order to identify and then formalize the basic ontologic semantic patterns of "prepositional grammar". We foresee the results of our research to be a useful part of NLP resources because prepositions have not been getting much attention by the specialists in this sphere. For instance, they were included in the so-called "stop-word" lists, which drop them from vector models in information retrieval procedures. It is true that frequent prepositions have low "specificity" index (tf-idf) but they do convey a clear semantic-syntactic relationship between the content words *переводить с английского* ('to translate from English') *vs переводить на английский* ('to translate into English'), с *21 января* ('from January, 21 ') vs *до 21 января* ('until January 21 '). These examples indicate the importance of prepositions for the generation and understanding of Russian texts. The other reason for additional attention to prepositions in NLP is the problem of secondary prepositions (see 2), that is, phrases which may be used as equals for primary prepositions.

The semantics of Russian prepositions was a matter of a large number of works which are mentioned in [Solonitsky 2003] and [Filipenko 2000] though they follow more traditional structural methodology and investigate several specific aspects of prepositional constructions. Anyway the linguistic prerequisites of prepositional construction analysis are the vital part of our method (see below).

2. Prepositions inventory

Prepositions as a part of speech form a rather obscure subset of the Russian vocabulary, though they are extraordinary frequent. In the Russian National Corpus (RNC) prepositions hold more than 10 % of tokens. Three

prepositions (в 'in', на 'on', с 'with') are in the list of ten most frequent words and 18 prepositions are in that of hundred ones.

The prepositions in Russian texts are heterogeneous and diverse: there is a small group of primary prepositions and a large number of secondary ones, the latter being motivated by the content parts of speech (nouns, adverbs and verbal forms), which may be combined with the primary prepositions forming multiword expressions. This fact shows that corpus frequencies of the primary prepositions are regularly overrated because they may be used as parts of secondary prepositions. Even the strict division between secondary multiword prepositions and prepositional noun phrases is not specified. There are some ideas about what is specific for the status of a secondary preposition. Firstly, the noun in the construction should have an abstract meaning (for prepositional derivatives on the basis of adverbs and gerunds it is always true). Secondly, the impenetrability of a linear sequence of such multiword prepositional expression is obligatory, for instance, в *течение* ('during'), *несмотря на* ('despite'), etc. The difference in spelling of such constructions may is an indirect and inconsistent indicator of their new status: в *течение года* ('during a year'); *несмотря на непогоду* ('despite the bad weather').

The most significant evidence of the "prepositional" nature of the multiword expression is its full or partial synonymy to the meaning of some primary preposition, bearing in mind that secondary prepositions which are highly frequent or have modified spelling, gained the confirmed status before. In order to compile the full list of secondary prepositions it is possible to look through existing inventories, however, they are only partially overlap and we can find strong arguments against one or another and sometimes against their "secondary" status in some papers.

In our project we will consider prepositions to be a stereotypical way for clarifying case meanings of nouns expressing valencies of content words (fist of all, verbs and verbal derivatives) and/or different circumstantial qualifiers in a sentence. The stereotype in our method should be proven by high corpus frequency of prepositional expressions with a particular meaning. The synonymy or partial synonymy between primary and secondary prepositions will be based on the semantic class specification for governing words and dependent nouns. This point of view allows us to outline the corpus strategy for picking up secondary prepositions. We may use some frequent noun that is used recurrently in a construction with some primary preposition as a "prepositional pattern", then we may specify roughly or finely its meaning (see below). After that we may find repeated multiword

expressions in the corpus sample expressing a similar sense. They are candidates for further consideration: e.g. their corpus frequency should not be less than the threshold value (hypothetically 1 or 3 IPM) and so on. For example: в *войне* („during the war"), *во время войны* („during the war"), в *годы войны* („during the war"), в период войны („during the war"), etc.

3. The metalanguage for prepositional meaning description

The meaning of prepositions in explanatory dictionaries are usually expressed by primary or secondary synonyms, but it is not uncommon that they are not interchangeable with the prepositions in question. For example in Russian Wiktionary (https://ru.wiktionary.org/) the first meaning of preposition *через* ('through') is described by secondaries "сквозь, поперек" and a gloss *Он помог женщине перейти через дорогу* ('He helped the woman cross the road'). However, it is impossible to insert equivalents from the definition into the gloss text. Prepositions from the dictionary interpretation may be defined by the primary one forming "vicious circle". The translation dictionaries demonstrate a number of phrasal examples which may be extrapolated by a user for other cases, but which one to use for a particular noun is not clear.

That is the reason why it is very important to provide a special semantic metalanguage for description of prepositional meanings. Its notions may be very coarse as in Russian grammar [Russian grammar 1980] description (objective, subjective, attributive, adverbial), more detailed as semantic arguments (objective, addressee, instrumental, spatial, temporal, etc.) [Mustajoki 2006] or specially invented for description of prepositional constructions such as syntaxemes proposed by G. A. Zolotova [Zolotova 2011].

The syntaxeme is characterized by a morphological arrangement (a preposition plus a case form) which has a unity of a form and a sense that functions as a constructive and significant component of a phrase or a sentence. The syntaxeme is a minimal grammatical construction which couldn't be split further into meaningful elements. The syntaxeme types are direction, destination, correlation, quantification, qualification, location, mediation, temporative, etc. The designation of syntaxemes in original Zolotova's version is formed according to a of semantic role pattern: directive, destinative, correlative, quantitative, qualitative, locative, mediative, temporative [Zolotova 2011: 383]. This nomenclature forms a starting point for our metalanguage which may be extended or shortened due to the needs of its expressive capability.

For the sake of quantitative grammar of prepositional constructions, we use the term "semantic rubric" as a generic name of the group of meanings of prepositions. It is often the case that these rubrics correspond to some Zolotova's syntaxeme. It is hardly possible to use this nomenclature downward, thus we start our description upward, from the most frequent primary prepositions. The minor variants of prepositional meanings associated usually with secondary prepositions create chains of synonymous or quasi-synonymous constructions.

To identify the semantic rubric we find out the basic meaning, that is represented by the frequent use of the primary preposition, sometimes it is a meaning of the case form. It is a prototypic representation of the rubric. Synonymous constructions are selected according to analyses of corpus samples.

4. Methodology for selection pairs "preposition — its meaning"

To solve this task we need appropriate corpora: representative, balanced, annotated and highly functional. A pilot research shows that Deeply Annotated Corpus (treebank) from the Russian National Corpus with the semantic annotation does not afford us to select the pairs of prepositions and their meanings. Thus, we carry out our research on the basis of morphologically annotated corpora.

We've chosen 2 corpora of Russian texts: Russian National Corpus (RNC) (http://www.ruscorpora.ru/en/index.html), Russian corpora of the Aranea corpus family (http://unesco.uniba.sk). They are different in size and in balance of textual genres. Russian National Corpus includes the balanced Main corpus and particularized subcorpora: the Spoken Russian corpus, the Newspaper corpus, the Dialectal corpus, the Poetry corpus and others.

The Aranea family consists of web corpora created by the wacky technology [Benko 2014]. For Russian there are three region-specific variants with a size from 120 to 1200 mln. tokens. We use mainly Araneum Russicum Minus (120 mln. tokens, 91 mln. lexical words). Araneum Russicum Externum corpus permits to create domain-specific subcorpora such as .ua, .by, .il, etc. that gives an opportunity to study Russian prepositional constructions in their regional variations.

A preliminary analysis of data shows that different corpora should be used to receive reliable data. Values of separate meanings vary noticeably from one corpus to another concerning both: primary prepositions and secondary ones.

Our technique involves corpus tools, other software instruments, and manual procedures. The semantic analyses of relations between lexical items expressed in prepositional constructions cannot be performed entirely automatically thus the manual linguistic annotation is a part of our methodology. Due to this stage we select the pairs "preposition-its meaning" from the corpus data for further ontology building and compile a number of annotated corpus samples which may be used later as a gold standard for prepositional meanings.

The crucial point of our methodology is a compilation of a random sample of contexts with prepositional constructions from the chosen corpora. The contexts are annotated loosely by a linguist according to the set of meanings from explanatory dictionaries, some meanings being united into one group if they are aligned with some semantic rubric or Zolotova's syntaxeme. Propositional meanings are ranked according to the percentage of a particular meaning of a certain preposition. The upmost ranks demonstrate the regular use of prepositional constructions, The bottom ranks show the irregular use. The meanings from the top ranks are extrapolated according to the total frequency of a preposition in the corpus and normalized to a number of million token presented in the corpus processed. These are IPM frequencies of prepositional meanings. They may be used for aligning of "preposition-its meaning" pairs with a similar meaning, that is a rubric or a syntaxeme.

5. Examples of corpus-based analyses of prepositional constructions

We demonstrate results of our methodology on two rubrics.

The *Mediative* as a semantic rubric has a narrow and a wide interpretation. Generally it is considered as a particular semantic role in a predicate structure of a verb. In the narrow sense the mediative is understood as a means, a substance or an object being used during the performance of an action or a process. In a broader sense the mediative is a tool (instrumentative) and includes its material and abstract implementations [Mustajoki 2006]. In the Russian language both mediative and instrumentative are regularly expressed by the instrumental case form: *красить стены <u>валиком</u>* (to roller the walls = to paint the walls with a roller), *рисовать картину <u>красками</u>* (to paint a picture = to draw a picture with paints), though they may be combined: *Писец <…> рисовал картинки тончайшей кистью красками на яичном желтке* ('The scribe <…> drew pictures with the finest brush by paints on egg yolk').

The *Transitive* is one of the possible ways of the proposition localization. Unlike the characteristics of location, which are applicable for diverse set of actions, states and processes, this specification is often associated with a "framework" structure of a prefix *пере-* for the verbs of motion and their derivatives: *перейти через дорогу* ('cross the road'), *перевозки нефти через Атлантику* ('an oil transportation across the Atlantic ocean'), etc.

The experiments and results received demonstrate the ability of corpus tools to obtain data showing individual prepositional meanings in Russian texts which were not mentioned in scientific articles.

5.1. Mediative prepositional constructions

The preposition «через» ('through, via') is used in the mediative rubric with a high corpus frequency of 173 IPM, the nuances of its meanings being very diverse: *гладить брюки через влажную ткань* (with an implicit tool *утюг*); *настроить протокол TCP/IP через вкладку Конфигурация; ультразвук воздействует на организм через воздух*; *ставить горчичники через газету*; *отмывать деньги через другие фирмы; протереть творог через дуршлаг; получить кредит через знакомых* etc. In many cases clear circumstantial characteristics are ambivalent. For instance, the mediative prepositional construction *пропустить мясо через мясорубку* ('to mince meat' = to skip the meat though a meat grinder) adds to "means" notion an idea of the real movement of the stuff (see Transitive below). Moreover, the first meaning of this preposition in the Russian Wiktionary is defined as transitive with synonyms «сквозь, поперек». It may be easily seen that mentioned synonyms are not interchangeable in this context, though in other constructions the preposition «через» can be substituted with «сквозь» ('through, across'): *протереть творог через/сквозь дуршлаг* ('to rub cottage cheese through a colander'), or *видеть через/сквозь стекло* ('to see through the glass').

The preposition «сквозь» is considered to be secondary (from an adverb), though there is no usage of «сквозь» as an adverb in the corpus. This preposition occurs less frequently (20 IPM) in the corpus than «через», this fact is normally interpreted as a cause of its more strict selectional preferences, however, we may see several contexts *видеть сквозь дымку/ туман/ крону деревьев* ('to see through the haze/ the fog/ the tree crown'), in which it is hardly possible to insert the frequent preposition «через» with a mediative-transitive meaning. In the given contexts constructions with the preposition «сквозь» designate that some thing|stuff can not prevent the implementation of some action, We will refer to such usage as mediative-in-

active as a new variant inside the mediative rubric. I seems that this modification results from the interference of mediative-transitive meaning with a less frequent use (7 IPM) of the preposition «сквозь» referring to auditory perception despite some kind of interference: *слышать смех/крик сквозь шум/ сон* ('hear laughter/ the cry through the noise/a dream').

Frequent secondary prepositions in this rubric are motivated by the noun «помощь» ('help') which may combined with different primary prepositions: с *помощью* ('by dint of') (98 IPM), *при помощи* ('by means of") (27 IPM). The most frequent semantic classes of governee nouns in this construction are actions с *помощью игры/ приема/ публикации* ('using the game/ reception/ publication'), tools с *помощью зеркала/ микроскопа/ проволоки* ('using a mirror/ a microscope/ the wire'), and human assistants с *помощью друзей, секундантов* ('with the help of friends/ seconds').

The secondary preposition «путем» ('by means of') has high frequency (51 IPM). It is used in texts of official genre with nouns denoting actions: *путем угроз/ обмана / обещаний* ('by threats/ deception/ promises).

We interpret the difference of statistical parameters of two mediative prepositions «через» and «с помощью» in different corpora shown in Table 1 as an argument that linkage of prepositional constructions to texts of the particular functional style are to become a subject of a separate investigation.

Table 1. **Frequencies of mediative prepositional constructions**

preposition	NRC balanced (200 examples)	NRC the Newspapers (200 examples)	Araneum Russicum Minus (200 examples)
через	IPM 173,53 20,5 % of all usages	IPM 221,30 33,2 % of all usages	IPM 185,21 32,5 % of all usages.
с помощью	IPM 76,63	IPM 113,21	IPM 228,64

Another noun motivating a group of secondary prepositions is «посредство», the frequent one is *посредством* ('by means of') (19 IPM) and infrequent *при посредстве, через посредство, благодаря посредству*, any one having the same interpretation: *посредством/ при посредстве, через посредство, благодаря посредству переговоров* ('through negotiation'). The interesting fact mentioned above is that the word «посредство» used as a noun occurs with extremely low frequency, not comparable with a frequency of the secondary preposition. The semantic classes of nouns used with

«посредством» are the same as was cited for secondaries motivated by a noun *помощь*: actions *посредством нажатия правой кнопки мыши* ('by right-clicking'), tools *посредством специальных приборов* ('using special devices') and substances *посредством анилиновой краски* ('by aniline dye').

Some uses of prepositional constructions are very peculiar, they are typically described as a metaphorical shift. It is highly common for primary prepositions, though some secondaries follow the same pattern. For example, the phrase *бить* ***через край*** ('to beat over the edge') means 'to be too agile"; *делать* ***через задницу/ жопу*** ('to do through the ass') => 'to do something awkwardly/ improperly'; *обратиться* ***через голову X*** ('to turn through the head of X') => "to skip X in a social hierarchy"; *оценивать* ***через/сквозь призму X*** ('to evaluate through a prism of X') => 'to evaluate from the point of view X'; *переступить* ***через себя*** ('to step over yourself') => 'to force one's feeling"; *пропустить X* ***через себя*** ('to pass X through yourself') => 'to make sense of X"; *хочет провалиться* ***сквозь землю*** ('wants to fall through the earth') => 'wants to disappear anywhere'; *смотреть* ***сквозь пальцы на X*** ('to look through the fingers at X') => 'to pretend not seeing X'*; цедить/ говорить* ***сквозь зубы*** ('to strain/ speak through his teeth') => "to speak in an inarticulate manner expressing scorn to the *interlocutor*".

5.2. Transitive prepositional constructions

The Dictionary of Russian [Dictionary 1985] provides three meanings for the preposition *через* that can be associated with the transitive semantic rubric: (1) "through, across, from one side to the other" (*перейти через улицу* 'to go across the street'); (2) "through space, medium, device" (*идти через толпу* 'to go through the crowd'*, смотреть через стекло* 'to look through the glass'); (3) "over something" (*перелезть через забор* 'climb over the fence').

The preposition «сквозь» is partly synonymous: the idea of crossing is not totally significant, what is essential here is the spatial nature of what is being crossed: *пройти/ пролететь сквозь атмосферу/ толпу/ пространство* ('to pass /fly through the atmosphere/ the crowd/ space'). Wikipedia mentions *насквозь* as a synonym to *сквозь*, however, there are no examples of a usage of *насквозь* as a preposition in real corpora. Some features of the action are stressed which are supported by the adverb co-occurence with the preposition *через*: *винт проходит* <u>*насквозь через*</u> *гриф* ('the screw passes through the neck'), *мы прошли* <u>*насквозь через*</u> *весь холл* ('we went through

the hall'). When *сквозь* and *через* combined with the motion verbs that refers to overcoming of an obstacle *росток пробился сквозь асфальт* ('the sprout made its way through the asphalt').

The distribution of frequencies for transitive propositional usage in different corpora is shown in Table 2.

Table 2. **Frequencies of transitive prepositional constructions**

preposition	NRC(balanced) (200 examples)	NRC newspapers subcorpus (200 examples)	Araneum Russicum Minus (200 examples)
через	IPM 245,39 25,5 % of all usages	IPM 118,65 17,8 % of all usages	IPM 157,71 27,5 % of all usages.
сквозь	IPM 119,83	IPM 19,25	IPM 23,90

5. Conclusion and further work

Semantic rubrics presented in our approach help to organize rather vague prepositional meanings. Their resemblance and difference may be explained by the overlap of semantic classes of governing and subordinate words. The whole structure of prepositional frequencies and distributions of neighboring semantic groups are resources for the compilation of the Russian quantitative prepositional grammar.

We plan the following stages:

- to compile sets of prepositional constructions occurring with reliable frequencies in corpus texts of different functional styles;
- to calculate a seria of statistical characteristics for each syntaxeme type with respect to corpus stylistic and thematic pecularities:
  - the IPM ranks in various corpus types;
  - the distribution of percentage ratios for frequent as well infrequent prepositional meanings;
  - the vector representation for typical semantic classes and/or lexemes acting as a "governor" for each prepositional meaning;
  - syntaxeme and/or lexemes acting as a "governee" for each prepositional meaning.

References

1. *Benko, V*. (2014) Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue.

17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pages 257–264.
2. Dictionary of the Russian language (1985). Vol. 1–4. 3rd edition. Moscow.
3. *Filipenko, M. V.* (2000). Problems of the description of prepositions in modern linguistic theories [Problemy opisaniya predlogov v sovremennykh lingvisticheskikh teoriyakh] In Preposition semantics research [Issledovaniya po semantike predlogov] . Moscow.
4. *Mustajoki, A.* (2006). Theory of Functional Syntax [Teorija funtsionalnogo sintaksisa]. Russia, Moscow.
5. *Slolonitskiy, A. V.* (2003). Problems of semantics of Russian primitive prepositions [Problemy semantiki russkikh pervoobraznykh predlogov]. Vladivostok.
6. *Zolotova, G. A.* (2011). Syntactical Dictionary: a Set of Elementary Units of Russian Syntax [Sintaksicheskiy slovar': repertuar elementarnykh edinits russkogo sintaksisa], 4th edition, Russia, Moscow.
7. Russian grammar (1980). Vol. 1–2. Moscow.

Irina Azarova
St. Petersburg State University (Russia)
E-mail: i.azarova@spbu.ru
Victor Zakharov
St. Petersburg State University (Russia)
E-mail: v.zakharov@spbu.ru

О. В. Блинова, С. А. Белов
O. V. Blinova, S. A. Belov

РУССКИЕ ОФИЦИАЛЬНЫЕ ТЕКСТЫ ДОМЕНА «ЗДРАВООХРАНЕНИЕ» И ОЦЕНКА ИХ ЛЕКСИЧЕСКОЙ СЛОЖНОСТИ С ИСПОЛЬЗОВАНИЕМ КЛЮЧЕВЫХ СЛОВ[1]

RUSSIAN OFFICIAL TEXTS OF THE «HEALTH CARE» DOMAIN AND ASSESSMENT OF THEIR LEXICAL COMPLEXITY USING KEYWORDS

Аннотация. Представленное исследование выполняется в русле изучения доступности для восприятия и понимания русских официальных документов из социальных доменов здравоохранения, культуры и образования. Материал – Корпус русских локальных документов и актов CorRIDA, подкорпус документов здравоохранения (617 107 токенов). Исследование направлено на выявление лексической специфики официальных документов домена с помощью метода извлечения ключевых словоформ, а также на оценку полученных ключевых словоформ с точки зрения их общеязыковой частотности. Анализируя ключевые словоформы в контексте общеязыковой частотности, мы исходили из идеи, что частотные единицы проще для восприятия и понимания носителями языка. Эта идея традиционно используется при оценке лексической сложности текстов.
Ключевые слова. Языковая сложность, лексическая сложность, официальные документы, корпус русских документов, клиентоориентированные тексты, ключевые словоформы, целевой корпус, референтный корпус, нормализованная частота, общеязыковая частотность.

Abstract. This article describes first findings of the study of Russian official documents comprehensibility. The research material is the Corpus of Russian local documents and acts «CorRIDA» (subcorpus of healthcare domain, consisting of 617107 tokens). The study aims to identify lexical peculiarities of official documents using the method of extracting keywords, as well as to evaluate the obtained keywords using their language frequency. Analyzing the key word forms in the context of linguistic frequency, we proceeded from the idea that frequent items are easier to understand. This idea is traditionally used for lexical complexity assessing.
Keywords. Language complexity, lexical complexity, official documents, corpus of Russian documents, client-oriented texts, key word forms, target corpus, reference corpus, frequency.

1. Цель и материал исследования

В компьютерной лингвистике активно развиваются новые междисциплинарные направления — «Computer Assisted Legal Linguistics» (CAL) и «Law and corpus linguistics» (LCL), подробнее см. [Hamann et al. 2016, Mouritsen 2017]. Создаются корпусы, которые позволяют изучать юридический язык в реальном употреблении на широком материале.

[1] Исследование выполняется при поддержке гранта РНФ № 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика».

Настоящая статья находится в русле исследования, посвящённого функционированию русских официальных документов в социальных доменах здравоохранения, культуры и образования, подробнее см. [Белов и др. 2018]. Исследование имеет два направления: «перцептивное» и «дескриптивное». В рамках «перцептивного» направления мы опрашиваем носителей русского языка с целью выяснить, как они воспринимают тексты документов. В рамках «дескриптивного» направления мы начали выполнять описание языка документов методами корпусной лингвистики. Это описание направлено на выявление особенностей языка документов, которые могут быть названы сложными.

Текстовым ресурсом для «дескриптивного» направления стал создаваемый Корпус русских локальных документов и актов CorRIDA (Corpus of Russian Internal Documents and Acts). В корпус входят т. наз. «локальные документы» (Internal Documents). Именно с такими документами часто сталкиваются носители русского языка, профессиональная деятельность которых не связана с делопроизводством, документооборотом и т. д.

В корпус вошли только «клиентоориентированные» тексты, адресованные широкой аудитории, выпущенные исключительно государственными учреждениями и находящиеся в открытом доступе на сайтах поликлиник, школ, театров, библиотек и т. п.

Настоящая статья направлена на выявление лексической специфики официальных документов домена «здравоохранение» с помощью метода извлечения ключевых слов, а также на оценку полученных ключевых слов с точки зрения их общеязыковой частотности.

2. Метод исследования

Списки ключевых слов формируются путём сравнения частот слов в двух корпусах. Первый корпус принято называть «целевым», или «объектным»; второй, фоновый, — «референтным», подробнее см., например, [Culpeper, Demmen 2015]. Слова, которые встречаются в целевом корпусе статистически значимо чаще, чем в референтном — это т. наз. «положительные ключевые слова» (positive keywords).

Анализ списков ключевых слов позволяет выявить существенные особенности текстов, связанные с их жанровым своеобразием [Xiao, McEnery 2005].

В рамках описания лексической специфики документов домена «здравоохранение» на настоящем этапе мы: 1. получили списки клю-

чевых словоформ (КС) с помощью AntConc [Anthony], 2. ранжировали КС по убыванию коэффициента «keyness coefficient», 3. снабдили каждую КС значением общеязыковой частотности, 4. проанализировали список КС с учётом общеязыковой частотности входящих в него единиц.

В качестве референтного корпуса мы использовали оффлайновую версию подкорпуса Национального корпуса русского языка (НКРЯ, ruscorpora.ru) со снятой омонимией. Он включает художественную и учебно-научную прозу, газетную публицистику, тексты электронной коммуникации, материалы устной речи (в общей сложности 1062625 словоформ), то есть характеризуется сбалансированным составом текстов.

Целевым корпусом является токенизированная коллекция текстов домена «Здравоохранение» из корпуса CorRIDA, включающая в общей сложности 617107 токенов, см. Таблицу 1.

Таблица 1. **Состав домена «Здравоохранение»**

	Тип документа	**N токенов**
1	Договор на оказание платных медицинских услуг	117088
2	Порядок оказания платных медицинских услуг	118863
3	Информированное добровольное согласие на медицинское вмешательство	61902
4	Согласие пациента на обработку персональных данных	62074
5	Правила поведения пациента, Правила поведения посетителя медицинского учреждения	126604
6	Правила госпитализации, записи на приём, консультацию, обследование и др.	130576
	Всего	**617107**

Каждой ключевой словоформе была соположена частотность соответствующей леммы из «Нового частотного словаря русской лексики» [Ляшевская, Шаров 2009]. Словоформы, относящиеся к одной лемме, получали одинаковые показатели частотности в ipm. Единицы, которых нет в частотном словаре, получали показатель «0».

Одним из традиционных параметров оценки сложности текстов является частотность входящих в него слов, см. об этом, например

[Chen, Meurers 2016], [Solovyev et al. 2018]. Этот подход основан на допущении, что слова с большей частотностью имеют более высокий уровень активации в сознании реципиента, и, соответственно, требуют меньших дополнительных усилий при их извлечении из ментального лексикона. Соответственно, частотные слова в среднем читаются быстрее и понимаются лучше, см. [Haberlandt, Graesser 1985], [Just, Carpenter 1980].

Согласно [Ляшевская 2017], утверждение, что «читатель замедляется или «спотыкается», встречая низкочастотные и/или незнакомые ему слова», является одной из презумпций, используемых при оценке удобочитаемости текстов. Так, в формуле удобочитаемости Дэйла-Чейла [Chall, Dale 1995] учитывается количество знакомых читающему слов текста. Между тем, именно частотность единицы обычно используется как показатель её «знакомости» для читателя, см., например, [Shanahan 2000, 26], а также [Ягунова 2010].

3. Результаты

При обработке целевого корпуса из 617107 токенов мы получили 15032 КС и рассмотрели 500 словоформ с максимальными значениями «keyness coefficient». Высокочастотные служебные и другие слова, которые принято включать в т. наз, стоп-листы, мы из рассмотрения сознательно не исключали.

Среди пятисот «самых ключевых» КС, имеющих соответствия в «Новом частотном словаре русской лексики», преобладают существительные (319 позиций списка, 324 вместе с аббревиатурами) и прилагательные (90 позиций списка).

Среди анализируемых пятисот КС всего 9 единиц (без учёта имён собственных) отсутствуют в «Новом частотном словаре русской лексики», это:

1) аббревиатуры *ГБУЗ* ‘государственное бюджетное учреждение здравоохранения’, *ОМС* ‘обязательное медицинское страхование’, *ЦРБ* ‘центральная районная больница’, *РБ* ‘районная больница’, *ДМС* ‘дополнительное медицинское страхование’, *ЛПУ* ‘лечебно-профилактическое учреждение’, *СНИЛС* ‘страховой номер индивидуального лицевого счёта’,
2) токен *ая*, возникающий в составе вхождений типа *именуемый*(*-ая*), специфичных именно для документов,
3) существительное *обезличивание*.

Среди КС с наименьшей общеязыковой частотой соответствующих лемм присутствуют, например, следующие (см. Таблицу 2).

Таблица 2. **КС с общеязыковой частотностью ниже 5 ipm**

PoS	ключевая словоформа	keyness coeff.	ipm	PoS	ключевая словоформа	keyness coeff.	ipm
s	нетрудоспособности	357.163	1,00	v	ознакомлен	336.424	3,60
s	регистратуру	357.163	1,00	v	удостоверяющий	357.163	4,20
s	регистратуре	304.164	1,00	v	удостоверяющего	304.164	4,20
s	посредством	232.851	1,60	a	амбулаторных	327.207	1,00
s	противопоказаниях	297.252	2,20	a	амбулаторного	246.558	1,00
s	госпитализации	1354.915	2,40	a	возмездной	345.641	1,50
s	госпитализация	483.898	2,40	a	предусмотренных	827.235	1,60
s	госпитализацию	474.681	2,40	a	предусмотренные	288.573	1,60
s	несоблюдение	290.339	3,80	a	предусмотренным	232.732	1,60
s	стационара	615.242	4,30	a	информированное	744.281	1,80
s	стационар	394.031	4,30	a	информированного	368.684	1,80
s	стационаре	271.905	4,30	a	лечащего	1007.365	1,90
s	неисполнение	417.095	4,60	a	лечащим	513.854	1,90
s,prop	УЗИ	632.684	3,40	adv	амбулаторно	276.513	0,40
v	обязуется	471.924	2,30	adv	натощак	269.600	1,10

Здесь есть единицы с относительно высоким keyness coefficient и низкими показателями общеязыковой частотности (это, например, ключевые словоформы *госпитализации*, *лечащего*). Это единицы, в наибольшей степени специфичные для документов домена «здравоохранение» и характеризующие тематику текстов.

Если же обратить внимание на КС с наиболее высокими показателями общеязыковой частотности, то после исключения служебных слов и слов, входящих в закрытые списки (различных местоимений), мы получим перечень, в который входят в том числе КС типа *лица*, *сторон*, *настоящего* и т. д. Значения таких слов в документах не эквивалентны общеупотребительным, и их анализ должен производиться отдельно.

4. Некоторые выводы и перспективы исследования

Таким образом, в настоящей статье был предложен ещё один возможный подход к выявлению лексической сложности, в рамках которого список ключевых слов исследуемой коллекции текстов анализируется с привлечением информации об общеязыковой частотности элементов списка.

Мы получили результаты, согласно которым из 500 ключевых словоформ целевого корпуса с максимальными показателями keyness coefficient только 9 отсутствуют в «Новом частотном словаре русской лексики» (НЧСРЛ).

Направлением дальнейшей работы будет, во-первых, получение результатов анализа КС с точки зрения общеязыковой частотности для всего списка КС (15 тыс.), во-вторых и в главных, более содержательная интерпретация списка КС с привлечением информации о распределении (дисперсии) соответствующих лемм по данным НЧСРЛ и с применением кластерного анализа, причём параметром кластеризации станет общеязыковая частота единиц (в общем о кластерном анализе для выявления ключевых слов см. [Gabrielatos 2018]).

Если же говорить о составлении рекомендаций для авторов официальных документов, адресованных широким аудиториям носителей языка, то представляется, что общих рекомендаций может быть две: не включать в документы низкочастотные единицы, а если без них не обойтись, то непременно пояснять их значение для читателя. Для получения сведений о частотности можно использовать, например, «Новый частотный словарь русской лексики» в он-лайн версии.

Литература

1. *Белов С. А., Блинова О. В., Гулида В. Б., Зубов В. И., Ларионова Е. Ю., Толстикова П. С.* (2018), Корпус русских локальных документов и актов CorRIDA: цели формирования, состав, структура // Компьютерная лингвистика и вычислительные онтологии, Вып. 2 (Труды XXI Международной объединенной конференции «Интернет и современное общество, IMS-2018), Университет ИТМО, СПб, с. 112–120.
2. *Ляшевская О. Н.* (2017), К определению сложности русских текстов // XVII Апрельская международная научная конференция по проблемам развития экономики и общества. Кн. 4, с. 408–418.
3. *Ляшевская О. Н., Шаров С. А.* (2009), Новый частотный словарь русской лексики, CSV-версия словаря. URL: http://dict.ruslang.ru/freq.php.
4. *Ягунова Е. В.* (2010), Исследование избыточности русского звучащего текста

// Acta Linguistica Petropolitana. Труды института лингвистических исследований, Т. 6, № 2, Избыточность в языке и речи, с. 90–115.

References

1. *Anthony L.* AntConc (Version 3.5.6), Waseda University, Tokyo, Japan [Computer Sof tware], available at: http://www.laurenceanthony.net/software.
2. *Belov S. A., Blinova O. V., Gulida V. B., Zubov V. I., Larionova E. Ju., Tolstikova P. S.* (2018), Korpus russkih lokal'nyh dokumentov i aktov CorRIDA: celi formirovanija, sostav, struktura [Corpus of Russian Internal Documents and Acts CorRIDA: Goals, Composition and Structure] // Komp'juternaja lingvistika i vychislitel'nye ontologii (Trudy XXI Mezhdunarodnoj objedinennoj konferencii «Internet i sovremennoe obshhestvo, IMS-2018) [Computational linguistics and computational ontologies (Proceedings of the XXI International Joint Conference «Internet and modern society», IMS-2018)]. St. Petersburg, University ITMO, Issue 2, pp. 131–147.
3. *Chall J. S., Dale E.* (1995), Readability Revisited: The New Dale-Chall Readability Formula, Brookline Book, Cambridge, MA.
4. *Chen X., Meurers D.* (2016), Characterizing text difficulty with word frequencies // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84–94.
5. *Culpeper J., Demmen J.* (2015), Keywords // The Cambridge handbook of English corpus linguistics / ed. by D. Biber and R. Reppen, Cambridge University Press, 2015, pp. 90–105.
6. *Gabrielatos C.* (2018), Keyness analysis: nature, metrics and techniques // Taylor, C. & Marchi, A. (eds.) Corpus Approaches to Discourse: A critical review, Routledge, London, pp. 225–258.
7. *Haberlandt K. F., Graesser A. C.* (1985), Component processes in text comprehension and some of their interactions // Journal of Experimental Psychology: General, 114(3), pp. 357–374.
8. *Hamann H. et al.* (2016), Computer Assisted Legal Linguistics (CAL²) // F. Bex and S. Villata (Eds.), Legal Knowledge and Information Systems, pp. 195–198.
9. *Jagunova E. V.* (2010), Issledovanie izbytochnosti russkogo zvuchashhego teksta [The study of Russian sounding text redundancy] // Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskih issledovanij. Izbytochnost' v jazyke i rechi [Proceedings of the Institute of Linguistic Studies, Vol. 6, № 2. Redundancy in language and speech], pp. 90–115.
10. *Just M. A., Carpenter P. A.* (1980), A Theory of Reading: From Eye Fixations to Comprehension // Psychological Review, Vol. 87, N 4, pp. 329–354.
11. *Lyashevskaya O. N.* (2017), K opredeleniju slozhnosti russkih tekstov [Towards the definition of the complexity of Russian texts] // XVII Aprel'skaja mezhdunarodnaja nauchnaja konferencija po problemam razvitija ekonomiki i obshhestva [XVII April International Academic Conference on Economic and Social Development], Book 4, pp. 408–418.
12. *Lyashevskaya O., Sharov S.* (2009), Chastotnyj slovar' sovremennogo russkogo jazyka

na materialach Nacional′nogo korpusa russkogo jazyka [The frequency dictionary of modern Russian language], Moscow, available at: http://dict.ruslang.ru/freq.php.
13. *Mouritsen S. H.* (2017), Corpus Linguistics in Legal Interpretation — An Evolving Interpretative Framework // International Journal of Language & Law, N 6, pp. 67–89.
14. *Shanahan T.* (2000), The National Reading Panel Report: Practical Advice for Teachers, North Central Regional Educational Laboratory.
15. *Solovyev V., Ivanov V., Solnyshkina M.* (2018), Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics // Journal of Intelligent & Fuzzy Systems, 34(5), pp. 3049–3058.
16. *Xiao R., McEnery T.* (2005), Two approaches to genre analysis: Three genres in modern American English // Journal of English Linguistics, 33, pp. 62–82.

Блинова Ольга Владимировна
Белов Сергей Александрович
Санкт-Петербургский государственный университет (Россия)
Blinova Olga, Belov Sergei
Saint Petersburg State University (Russia)
E-mail: {o.blinova, s.a.belov}@spbu.ru

V. Bobicev, Y. Hlavcheva,
O. Kanishcheva, V. Lazu

AUTHORSHIP ATTRIBUTION IN SCIENTIFIC PUBLICATIONS

Abstract. The goal of our study was to compare several machine learning methods and various sets of features in the task of authorship attribution. We used specific data for our experiments, namely scientific articles from different domains. While the domain specific words could affect the classification, we selected sets of texts from the same domain and performed authorship attribution within the domain. The results demonstrated the ability of machine learning methods to identify the author of a text with relatively high confidence. We obtained F scores in the interval of 0.85-0.95.
Keywords. Text classification, author recognition, machine learning methods, character based learning, word based learning.

1. Introduction

Authorship attribution is an important problem and it has practical applications in such areas such as law, journalism, education where knowing the author of a document may solve disputed texts.

It is a comparatively old task that was studied in various domains and settings. Our paper presents authorship attribution experiments for scientific papers published in Ukraine. Our work is a part of the complex task of plagiarism prevention. Using software and technological tools to detect possible plagiarism is a necessary action to promote academic integrity.

2. Related Work

The problem of authorship attribution is an old one being analysed in [Mendenhall 1887] for Shakespeare plays and [Mosteller, Wallace 1964] for disputed Federalist Papers.

During the last decades, this field of research has been developed substantially due to computational methods such as machine learning and natural language processing. Two general methods were developed in this area: the classical one, called 'stylometry' and purely statistical-based, mostly by using machine learning techniques. An indicative paper in this regard is [Kukushkina et al. 2002]. Two methods were compared in the paper: one method used grammatical information, namely, part of speech tags and their sequences from text and the other used just bi-grams of letters from the same text. The second method performed better and the authors wrote in their conclusion: "It is amazing that the use of such simple units as pairs of letters from text gives more precise results than the use of grammatical codes and their sequences".

The subject has been addressed in several surveys. [Stamatatos 2009] in his survey analysed in details and classified all stylometric features used to define the author of a text. He classified them in character, lexical, syntactic, semantic, and application specific and discussed when each type of features was more suitable than others. Finally, various issues of different methods evaluation were addressed and open questions were listed as a conclusion of the survey.

Series of online challenges dedicated authorship attribution, author profiling and several adjacent tasks have been organised by PAN network of experts on digital text forensics[1]. It has been organising shared task, computer science events that invite researchers and practitioners to work on specific problems of authorship attribution.

[Koppel 2009] analysed the previous works in details and addressed more difficult scenarios of real life authorship attribution as the author profiling problem when there is no candidate set and the task is to provide such information about the author of the text as gender, age, provenance or some psychological information or detection of the most probable author having thousands of candidates with a limited writing samples or verification problem, when there is one suspect and the task is to determine whether the suspect is the author of the given text. Machine learning methods adaptations were discussed for each of these difficult tasks.

Most of the modern studies in this domain are concentrated on user-generated online texts such as blog, forum posts, comments or reviews. [Bobicev et al. 2013] is an example of such paper; 100 forum posts of each of 30 authors were considered and the task was to define the author of a single post comparing to all given texts. The results were quite impressive: 97.9 % on messages containing at least 300 words.

We, however, are working with scientific publications and the shortest texts in our collection are abstracts of approximately 150 word length. To the best of our knowledge there were no such studies for these types of text in Ukrainian.

3. Experiments

3.1. Data Description

We worked with research publications written in Russian and Ukrainian languages. Statistics for these documents is presented in Table 1. Each file is a research publication of exactly one author.

[1] https://pan.webis.de/

Table 1. **Statistics for Russian and Ukrainian parts of the corpus**

	Ukrainian	Russian
Num of authors	32	8
Num of documents	271	77
Total volume (words)	546 293	138 509
Average num words/author	17 072	17 313
Max num words/author	63 950	34 705
Min num words/author	4 577	7 134
Average num words/file	2 016	1 799
Max num words/ file	11 086	9 753
Min num words/file	153	165

There were 9 files per author in average. However, as it s seen in the table 1, the distribution of text volumes among the authors and the files is highly unbalanced. The smallest files contained only 150-160 words. For the author the difference was even greater: the author with the smallest volume of texts had approximately 15 times less words than the author with the biggest volume.

3.2. Preprocessing

All publications we worked with were in text format. Most of the research articles contain formulas, figures and tables. While the articles were transformed from pdf file to text, most figures disappeared, tables and formulas were corrupted. Thus, we manually checked all files and removed what remained from broken tables and formulas. Also, various specific characters used in formulas and text to denote used values were impossible to erase from text. We encoded all text in utf-8 to preserve the text and other elements.

The texts had initial metadata annotation using xml tags. Table 2 contains the used tags and their description.

It was obvious that we had to remove the parts of the documents that are in tags <author>, <references>, <inf>

Finally, we decided to run experiments only on <text> parts of the articles. The texts from the <text> parts were pre-processed. First of all, the texts contained multiple unnecessary line breaks introduced in the process

of transformation. To solve this problem, we removed the line breaks and joined all rows, keeping the breaks only after full stops.

Table 2. **Description of the xml tags used for corpus mark-up**

Tag	Description
<title>	title of the paper
<author>	author's name and affiliation
<description>	abstract, key-words
<text>	main text of the paper
<references>	references
<inf>	additional information about the author, her/his affiliation or the project

In order to solve the problem of different file sizes we divided each file in fragments with the length of the smallest files. The minimum length of 150 words was considered as acceptable. Table 3 contains data about the texts used in the experiments.

Table 3. **Statistics for the sets of data used in the experiments**

	Ukrainian	Russian
Num of fragments of length 150-200 words	2 634	928
Num of word features (with frequency >4)	9 972	5 175
Num of selected words features	600	1 654

3.3. Method Description

Algorithms. We tested two methods of text classification: character based and word based. The character based method we used was based on PPM compression algorithm and was described in [Bobicev et al. 2013]. Word based methods were classical ones and included Bayes based algorithms (NB), Support Vector Machine (SVM), Decision Trees and K-Nearest Neighbours (KNN). We used Weka[2] implementations of these algorithms.

Features. For the character based methods the features were all sequences of characters from text. All character sequences of length 5, 4, 3, 2 and 1 character from texts were used in the algorithm. We tested two variations

[2] https://www.cs.waikato.ac.nz/ml/weka/

of this method. The first one worked with absolutely all characters used in texts including upper and lower case letters, numbers, spaces, all kind of punctuation marks and other specific characters which appear in scientific publications. The other used only words and spaces between them. All other characters were ignored and all letters were transformed in lowercase.

For word based methods we created text's vocabulary and selected all words with frequency more 4 (5 and more). Then, to make the feature set smaller, we selected words using information gain ratio of the feature calculated as:

$$IG(class, feature) = (H(class)\text{-}H(class/feature))/H(feature) \quad (1)$$

Then we selected all words with positive information gain ratio.

We see the task of author recognition as a classification task [Stamatatos 2009] with a predefined set of authors as classes and a number of text fragments as instances. The task is to attribute each fragment/instance to one of the authors/classes from the set of candidate authors/classes. We run several experiments using algorithms described above. The classification evaluation measure was F-score which is a harmonic average of Precision and Recall [Sasaki, 2007]. To test which algorithms are better for our case we shuffle all obtained files and run ten-fold cross-validation when 9/10 of all files were used for training and 1/9 part was the test set. These settings are repeated ten times changing the test part every time. Then the average of the obtained results was calculated.

Table 4. **Results of the experiments for Russian and Ukrainian**

	Ukrainian	**Russian**
character based PPM	0.854	0.972
letter based PPM	0.865	0.979
On the base of words as features		
Naive Bayes	0.855	0.897
Naive Bayes Multinomial	0.896	0.940
Support Vector Machine	0.875	0.938
On the base of selected features		
Naive Bayes	0.798	0.945
Naive Bayes Multinomial	0.832	0.978
Support Vector Machine	0.827	0.968

3.4. The Obtained Results

We run experiments on Russian and Ukrainian texts apart. The results of the first round of the experiments are presented in table 4. We presented only algorithms with the better results; Decision trees and K-nearest neighbours had worse results and we have not included them in our final table.

As it is seen in the table 4, the obtained results are quite good. Even among 32 authors in Ukrainian corpus F-measure is 0.85-0.89.

We supposed that topics of the texts are influencing the classification and selected sets of documents grouped by the domains of research. All papers in the corpus were collected from five scientific domains: Engineering (46 papers), Humanities (38 papers), IT (65 papers), Economics (175 papers) and Chemistry (24 papers). In order to test our idea we selected two domains with the largest number of papers: Economics and IT and performed the experiments on these sets of documents. The statistics for these subsets is summarized in Table 5.

Table 5. **Statistics for topics: IT and Economics**

	IT	Economics
Num of authors	10	18
Num of fragments of length 150-200 words	1 334	1 328
Num of word features (with frequency >4)	6 015	6 115
Num of selected words features	1 106	585

The same experiments were run for the texts from these two domains. The results are presented in table 6.

Table 6. **Results of the experiments for topics: IT and Economics.**

	IT	Economics
character based PPM	0.958	0.852
letter based PPM	0. 943	0.843
On the base of words as features		
Naive Bayes	0.892	0.838
Bayes Multinomial	0.958	0.900
Support Vector Machine	0.955	0.860

On the base of selected features		
Naive Bayes	0. 915	0.828
Bayes Multinomial	0.945	0.851
Support Vector Machine	0.961	0.838

4. Discussion and Future Work

The obtained results are comparatively good and demonstrate that we are able to define the author of the text even for short texts, less than 200 words. The best result was obtained for Russian papers: F-measure of 0.979 using letter based method. In this case we think topic related features helped to distinguish the texts. Also, there was the smallest number of authors, only 8 in this set. It is interesting to compare the results of character and letter based methods. While working with the whole corpus, letter based method gained slightly better results but on the topic subsets the results were the opposite. Probably, in the whole set, topic specific words were more influential in classification. However, while all texts were from the same topic, word based features were less helpful but some specific characters helped distinguish texts.

While comparing character based and word based methods we cannot univocally say that one of the methods is better than other. In all experiments their results are similar. For Ukrainian corpus, the best result F=0.896 was obtained by Naive Bayes Multinomial on the whole set of words and letter based method gave F=0.865. For Russian the best result F=0.979 was obtained by letter based method and F=0.978 was obtained by Naive Bayes Multinomial on the selected set of features.

IT and Economics domains sub-corpora were comparable in sizes but the number of authors was different; 10 and 18 authors respectively. The results reflect this difference: for Economics (with 18 authors), the best result F=0.900 (Naive Bayes Multinomial on the whole set of word features) and for IT the best F=0.961 (Support Vector Machine on the selected word features). The most difficult was the whole Ukrainian sub-corpus with totally 32 authors and it should be noticed that even in such difficult conditions machine learning methods showed good F=0.896. This allows us to claim that we are able to define the author of the text with high precision.

However, we cannot claim that one of the used methods is definitely better than others on the basis of this study. The same situation is with features; we cannot say that word or character based features are definitely better.

Our future plans include following directions of investigations:

- We plan to experiment with different sizes of text fragments as it was done in [Bobicev et al., 2013] and see how the accuracy of author detection changes;
- We are going to explore the publications authored by two authors who are already in our collection of text as single authors of the papers. This allows us to see which part of the text was written by which co-author.

References

1. *Bobicev V., Sokolova M., Khaled El Emam, Jafer Y., Dewar B., Jonker E., Matwin S.* (2013) Can Anonymous Posters on Medical Forums be Reidentified? Journal of Medical Internet Research 2013 (Oct 03); 15(10):e215.
2. *Koppel M., Schler J., Argamon S.* (2009) Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology, Vol. 60, Issue 1, pag. 9–26.
3. *Kukushkina O. V., Polikarpov A. A., Khmelev D. V.* (2002) Using Literal and Grammatical Statistics for Authorship Attribution. Problems of Information Transmission 37(2).4. *Mendenhall T. C.* (1887). The characteristic curves of composition. Science, IX, 237–49.
4. *Mosteller, F. & Wallace, D. L.* (1964). Inference and disputed authorship: The Federalist. Addison-Wesley.
5. Sasaki, Y. (2007). The truth of the F-measure. Teach Tutor mater 1 (5), pp. 1–5.
6. *Stamatatos E.* (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), pp. 538–556.

Bobicev Victoria
Technical University of Moldova
E-mail: victoria.bobicev@ia.utm.md

Hlavcheva Yulia
National Technical University "Kharkiv Polytechnic Institute"
E-mail: glavjul@gmail.com

Kanishcheva Olga
National Technical University "Kharkiv Polytechnic Institute"
E-mail: kanichshevaolga@gmail.com

Lazu Victoria
Technical University of Moldova
E-mail: victoria.lazu@ia.utm.md

V. Broz

A CORPUS-BASED CRITICAL DISCOURSE ANALYSIS OF *BREXIT* IN THE ENGLISH LANGUAGE PRESS

Abstract. This paper analyses the language of the representation in the English language press of «Brexit», the threatened withdrawal of the United Kingdom from the European Union. In the referendum of 23 June 2106, the people of the UK voted to leave the EU by a majority of 51,89 % to 48,11 %. It is claimed that the British media had influenced the public opinion contributing to a Eurosceptic mood. We will analyse the recent development of the collocational range of the term «Brexit».

Keywords. Brexit, Corpus Linguistics, Critical Discourse Analysis, nationalism, idealism, Euroscepticism, xenophobia.

1. Introduction

The word *Brexit* has become ubiquitous, but in fact it is a neologism only seven years old. This lexical blend was officially recognised in December 2016 by the Oxford English Dictionary, defining its meaning as «the (proposed) withdrawal of the United Kingdom from the European Union, and the political process associated with it». Before the word entered the dictionary, most newspapers used to write *Brexit* in quotation marks.

In June 2016 the people of the UK were asked a historic question — «Should the United Kingdom remain a member of the European Union or leave the European Union?». By a majority of 51,89 % to 48,11 % they voted to leave the EU and the consequences of this momentous event in modern European history are still ongoing. It is often claimed, e.g. [Anderson & Weymouth 2014] that Britain's media had influenced the public opinion contributing to a Eurosceptic mood, so it is worth taking a look at the linguistic features used by the media to gain a clearer understanding of the effect and control that language might exert.

There are many news media to choose from in the UK. According to Buckledee [2018: 4], «the various papers were divided almost 50-50 during the referendum campaign: predictably, the right-wing *Daily Express, The Telegraph* and *Daily Star* campaigned for Leave, and *The Sun* maintained its long tradition of insulting Europe in general and France in particular, and it was equally predictable that *The Guardian, The Observer, The Independent* and *Daily Mirror* would take up the Remain cause. There were also surprises, however: *The Times* campaigned to stay in the EU while its sister newspaper, the *Sunday Times,* encouraged its readers to vote to leave, and the *Daily Mail* campaigned aggressively for Leave but *Mail on Sunday* opted for Remain.» In

addition, *The Economist* is pro-globalist. However, it would be too simplistic to suggest a clear left-right divide in anti-pro Brexit sentiment. It would be too strong to say that the Labour party is pro-Brexit. The Labourite attitude towards Brexit is very complex and is getting more complicated from day to day. Labour has frequently railed against a «Tory Brexit» as cover for the ambivalence of its own leadership and in order to make the government fully «own» the process, which they present as being very badly handled. The current government of the UK is Tory (the Conservative Party). While it is currently more left-leaning than it has been in several generations, the majority of its MP's are anti-Brexit. In the referendum, many strong Labour constituencies voted pro-Brexit. There tends to be a working-class anti-immigration alignment in some places, which is how the UK Independence Party (UKIP) managed to take Labour votes.

Some of the Brexit campaign's controversial issues include immigration, sovereignty, autonomy, nationalism, populism, Euroscepticism and European identity. The debate is driven more by fear than idealism. There have been some changes in the Daily Mail's approach over the past 12 months. They changed the editor. The new editor is much less Eurosceptic than his predecessor. BBC's position as unbiased commentator has come under huge stress because Remainers think it is backing Brexit, while Brexiteers, who have always hated it because they perceive it as an instrument of the left-wing state think it is «talking Britain down».

The semantic prosody of keywords in *Brexit* discourse can be discursively contested. Newspapers have considerable power to influence public opinion, as many academic studies indicate, e.g. [van Dijk 1991]. What we can prove is that there was or was not systematic patterning around the term *Brexit* in the media that made the newly coined concept more or less palatable to the audiences. It is possible that the media made the concept unpalatable, but the voters ignored this. This paper can show the degree of palatability.

In order to carry out a linguistic analysis of *Brexit* collocations, it is useful to examine their usage in corpus. However, rather than building a corpus from scratch, which is what we did in our preliminary research, an existing corpus has been used: the NOW (News on the Web). The NOW is part of the Brigham Young University (BYU) collection of corpora and one of the features of this collection is allowing to build «virtual corpora», which are actually a selection from a larger corpus, such as NOW. It contains about 6,64 billion words of data from web-based newspapers and magazines from 2010 to the present time. Virtual corpora can be based on date and source, such as once specific newspaper only.

2. Previous research

There is a growing number of publications on the topic of *Brexit*, often focusing on analysing the political agendas of different newspapers during the build-up to the 2016 EU Referendum.

One of the most noteworthy ones is Fontaine's [2017] article on the early semantics of the neologism *Brexit.* She discusses how the term *Brexit* was coined in 2012 and she explores the development of this nominalised blend word using a corpus of 1 641 903 words including 2 345 instances of *Brexit* from its first use in May 2012 to the UK general election in 2015. The purpose of her article was to show that all nominalizations are instances of grammatical metaphors. In other words, she shows how a neologism that denotes a contested and complex topic effectively gains «common currency». The corpus that she used for her study was *Lexis Nexis Academic,* an online database that contains full text newspaper articles from around the world in a variety of languages.

As the word *Brexit* was only in its beginnings of usage, the n-grams and collocations that she listed as most frequent reflect the times before Brexit became reality.

Fontaine's [2017] most frequent collocations and n-grams roughly express provisionality (*so-called*) with 132 occurrences, hypotheticality (*would be, possible, potential*) with 310, 60 and 56 occurrences each and uncertainty (*risk*) with 60 occurrences. The remaining two groups (*exit and referendum*) resist such categorization.

3. The Analysis Methodology

The language of Brexit is analysed within the theoretical framework of Critical Discourse Analysis (CDA), as postulated by Fairclough [1995], Van Dijk [1991] and Wodak [2006]. CDA is an interdisciplinary approach to the study of discourse that views language as a form of social practice or more specifically how power is exercised through language. Mautner [2009: 32-33] discusses why corpus linguistics and CDA are «occasionally seen as uneasy bedfellows», but she explains that none of the CDA principles are inimical to those of corpus linguistics. In fact, recent developments have shown much benefit from the application of corpus linguistics methods to CDA, as researchers used to be criticised for cherry-picking individual texts to suit their own political agendas. Cherry-picking in fact presents a big problem in corpus analysis research, as any collection of texts is easily criticized when they are selected to represent a broad and complex debate such as *Brexit* in this case,

especially in a world that has largely gone digital. Baker & Levon [2015] have approached these questions in their article named "Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity".

Hunston [2002: 12] slightly adapts Firths' traditional definition of collocation saying that "Collocation is the tendency of words to be **biased** in the way they co-occur". The key word is **biased.** Collocation is closely related to discourse prosody that is suggestive of attitudes since it studies the relations of words' meaning to the speaker and hearer.

Frequency can reveal some facts about discourse and attitudes. Frequency can also be an indicator of markedness, which is a way to understand something based on its relation with other things, sometimes by its opposite or a binary distinction. Analysing frequencies is beneficial for discourse analysis, as this may help us uncover evidence of bias, ideology and political stances in texts, especially when combined with other background knowledge.

Collocations can be motivated so in order to observe them, it is most reliable to measure them statistically. Statistics can be accounted for with logical explanations why some patterns occur more often than others and why certain words attract the company of others. Therefore, it is useful for critical discourse analysis as it provides not only the semantic definition of the word but also other implicit aspects of that word within a particular discourse.

4. The Results of Analysing Adjectival Collocates of *Brexit*

Overall, the collocation *hard Brexit* is leading the frequency chart, occurring four times more than the second place *soft Brexit*. Shortly after the referendum, *hard Brexit* and *soft Brexit* have become conventionalised collocations. They were not so transparent at first for many speakers of English, which is why some politicians think the collocations *clean Brexit* or *full Brexit* should replace *hard Brexit.* However, when something becomes conventionalised in language, it is difficult to change it.

Hard Brexit is favoured by ardent Brexiteers and would see the UK refusing to compromise on the free movement of people even if it meant leaving the single market or having to give up hopes of free trade arrangements. It would prioritise giving the UK full control over its borders, making new trade deals and applying laws within its own territory. Thus the UK will leave the EU single market and trading with the EU, as if it were any other country

Table 1. **The frequency list of the collocates of *Brexit***

1	*hard*	3065	11	*extreme*	127
2	*soft*	698	12	*key*	119
3	*chief*	686	13	*formal*	117
4	*final*	430	14	*successful*	114
5	*Tory*	216	15	*surrounding*	110
6	*softer*	164	16	*so-called*	106
7	*good*	154	17	*leading*	105
8	*possible*	150	18	*continuing*	103
9	*no-deal*	146	19	*best*	100
10	*new*	135	20	*potential*	96

outside the EU, based on the rules of the World Trade Organization. In other words, the UK and the EU would probably apply tariffs and other trade restrictions on each other.

Soft Brexit could involve keeping close ties with the EU, possibly through some kind of membership form of the EU single market, in return for a degree of free movement. This approach would leave the UK's relationship with the EU as close as possible to the existing arrangements, as is preferred by many Remainers (also known as «Bremainers»).

The metaphors *hard Brexit* and *soft Brexit* compare the negotiations between the EU and UK to the firmness of objects like rocks or pillows. A *no-deal Brexit* is a scenario in which the UK leaves the EU without formal agreements for the future relationship. So it is a kind of *hard Brexit*.

5. The Results of Analysing the Adjectival Collocates in Terms of Negative and Positive Discourse Prosody

Using a corpus of naturally occurring language, a lexical item can be classed to have negative, positive or neutral discourse prosody if it predominantly co-occurs with unpleasant, pleasant and neutral collocates. The following two tables present some negative and positive *Brexits* with the number of occurrences in the corpus.

May's favoured Brexit will be «smooth and orderly». No one, presumably, is in favour of a *rough Brexit*, like a rough sea; still less would anyone dream of a *disorderly Brexit*. *Blind Brexit* is a situation where the UK leaves the EU and

Table 2. **Negative *Brexits***

hard	3065	*tough*	65	*reckless*	30
disorderly	281	*damaging*	51	*botched*	26
bad	180	*disastrous*	48	*harsh*	17
looming	142	*destructive*	44	*car-crash*	14
chaotic	129	*disruptive*	44	*brutal*	9
cliff-edge	108	*messy*	44	*Kamikaze*	8
harder	87	*blind*	42	*phoney*	3

enters a transition period before agreement is reached about the long-term future relationship.

Botched Brexit starts appearing in January 2017. This adjective is used to describe something, usually a job that is done badly. *Phoney Brexit* refers to the period of nine months after the referendum until Theresa May formally triggered the process under Article 50 of the Lisbon Treaty. *Cliff-edge*, *car-crash* and *Kamikaze* make for particularly vivid image metaphors as the collocates of *Brexit*.

Table 3. **Positive *Brexits***

soft	698	*great*	71
softer	422	*better*	64
good	275	*business-friendly*	28
successful	263	*favourable*	27
orderly	158	*glorious*	21
best	157	*coherent*	14
smooth	147	*beneficial*	13
sensible	80		

When we take a look at the list of positive *Brexit* collocates, it is interesting to note that the opposites, as in *hard* vs. *soft* or *disorderly* vs. *orderly* come in unbalanced numbers. The list of positive *Brexits* is skimpier than the negative ones, although *better* and *best* appear more often than *worse* and *worst*. Some positive collocates are in fact ironic, such as *glorious*.

6. Conclusion

Studying the combination of an adjective preceding the word *Brexit*, many patterns and trends of usage can be observed, analysed and discussed. In this paper, we have shown how dynamic the development of its collocations is in a really short period of time after it was invented. Negative meanings of the collocates far outnumber the positive ones, pointing to the fact that the lexical item *itself* has been absorbing the negative affect or connotation. Being such an unprecedented event, no wonder Brexit has generated an ever-growing array of metaphors to try and make sense of it.

What makes this topic really interesting for research both in corpus linguistics and in critical discourse analysis is the fact that never before in living memory have some newspapers fed the public's hopes, fears and prejudice against Europe to this extent.

References

1. *Anderson P.J., Weymouth T.* (2014), Insulting the Public?: The British Press and the European Union. Routledge.
2. *Baker P., Levon E.* (2015), «Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity» [Discourse & Communication] Vol. 9 (2), pp. 221–236.
3. *Buckledee S.* (2018), The Language of Brexit. How Britain Talked Its Way Out of the European Union. London: Bloomsbury.
4. *Fairclough N.* (1995), Critical Discourse Analyis. Boston: Addison Wesley.
5. *Fontaine L.* (2017), "The Early Semantics of the Neologism BREXIT: a Lexicogrammatical Approach", [Functional Linguistics] Vol. 4 (6).
6. *Hunston S.* (2002), Corpora in Applied Linguistics. Cambridge: Cambridge University Press.
7. *Mautner G.* (2012), "Corpora and critical discourse analysis". *Baker P.* (*ed.*), [Contemporary Corpus Linguistics] New York: Continuum, pp. 32–46.
8. *Van Dijk T.* (1991), Communicating Racism: Ethnic Prejudice in Thought and Talk. London: Sage.
9. *Wodak R.* (2006), "Discourse-Analytic and Socio-Linguistic Approaches to the Study of National(ism)" [The Sage Handbook of Nations and Nationalism], (*ed.*) Delanty G. and Kumar K. London: Sage Publications.

Vlatko Broz
São Paulo University (Brazil)
E-mail: vlatkobroz@gmail.com

Л.Л.Иомдин
Leonid Iomdin

РУССКИЕ МИКРОСИНТАКСИЧЕСКИЕ ЭЛЕМЕНТЫ, МОТИВИРОВАННЫЕ СЛОВОМ *ВИД*: КОРПУСНОЕ ИССЛЕДОВАНИЕ СЕМАНТИКИ[1]

RUSSIAN MICROSYNTACTIC ELEMENTS DERIVED FROM THE NOUN *VID*: A CORPUS STUDY OF SEMANTICS

Аннотация. С точки зрения теории микросинтаксиса рассматриваются адвербиальные элементы русского языка, образованные сочетаниями предлогов с существительным *вид* (в *виде, в виду, из виду* и т.д. Исследование ориентировано на синтаксические и семантические особенности этих единиц и основано на материале НКРЯ и части СинТагРус'а, содержащего микросинтаксическую разметку. Показано, что данные единицы по-разному соотносятся с мотивирующим существительным и принадлежат к разным разрядам адвербиалов. Например, с *виду* обычно выступает как приименной атрибут (*Он с виду сельский учитель,* но не **Он с виду учительствует в селе*), а *на виду* — как приглагольное обстоятельство (*Целовались на виду у всех*).
Ключевые слова. Микросинтаксический словарь, микросинтаксическая разметка корпуса, адвербиалы, семантика конструкций, валентная структура.

Abstract. The paper, written in the framework of microsyntax, discusses Russian adverbials formed by prepositional phrases with the noun *vid* 'view/species' (*v vide* 'in the form of, as', *v vidu* 'in sight', *iz vidu* 'from sight' etc. The study is focused on the syntactic and semantic features of these units and is based on the material of RNC and SynTagRus containing microsyntactic markup). These units correlate differently with the motivating noun and belong to different categories of adverbials. E.g. *s vidu* is a normal attribute of an NP *On s vidu sel'skij učitel'* 'He is a rural teacher by sight', but not **S vidu on učitel'stvuet v sele* 'He teaches in the country by sight' while *na vidu* is an adverbial modifier of a verb (*Tselovalis' na vidu u vsex* 'They kissed in front of everyone').
Keywords. Microsyntactic dictionary, microsyntactic annotation of the corpus, adverbials, semantics of constructions, valency structure.

1. Вводные замечания

Исследования микросинтаксических единиц, проводимые автором в течение ряда лет, и в первую очередь работа над Микросинтаксическим словарем русского языка и микросинтаксической разметкой корпуса СинТагРус (Иомдин 2017a, 2018, Iomdin 2017b) показали, что значительная часть таких единиц — конкретнее говоря, синтаксических фразем — представляет собой предложные группы, т.е. сочетания предлога с полнозначным существительным. Примерами являют-

[1] Работа выполнена при поддержке гранта РНФ 16-18-10422. Автор выражает фонду искреннюю признательность.

ся разнообразные предложно-именные сочетания типа *при помощи*, *на слуху, без спросу, на редкость, в числе и* многие десятки других. Одна из особенностей этих единиц — их семантическая некомпозициональность: их значения могут быть сильно удалены от значений входящих в них существительных. Так, говоря *при помощи молотка*, мы не имеем в виду, что молоток «помог» нам что-то сделать, если какая-то *новость у нас на слуху*, то это в общем не значит, что мы используем способность слышать или же что речь идет о какой-либо сплетне или молве и т. д. Тем самым мы имеем дело с особыми языковыми единицами, требующими индивидуального описания.

Любопытно, что некоторые существительные порождают достаточно большое количество синтаксических фразем. Скажем, слово *раз* фигурирует более чем в тридцати таких единицах (в *n раз*, *на раз-два, через раз, за раз, от раза* к *разу* и др.; см. Иомдин 2015). Много разнообразных предложно-именных сочетаний формирует и слово *вид*. Само это слово имеет достаточно много значений, однако, несколько огрубляя лексикографическую картину и игнорируя терминологические употребления, можно сгруппировать эти значения в два блока: объект видения ВИД 1 (*вид на Москву с Воробьевых гор*) и элемент классификации ВИД 2 (*виды млекопитающих*). Однако микросинтаксические единицы, образованные с участием этого слова, впрямую не соотносятся с этими значениями. В большинстве таких единиц, правда, просвечивает идея видения, но лишь в самых общих чертах.

Наше изложение строится по следующему плану: в разделе 2 будет дан краткий обзор микросинтаксических единиц, сформированных с участием слова *вид, а в* разделе 3 мы подробнее остановимся на одной группе таких единиц: предложно-именных сочетаний в *виде и в X-овом виде*. Материалом исследования послужили в первую очередь основной корпус НКРЯ и микросинтаксически размеченная часть глубока аннотированного корпуса СинТагРус.

2. Микросинтаксические единицы со словом *вид*: общий обзор

Мы будем рассматривать здесь лишь предложно-именные сочетания адвербиального типа, т. е. такие, которые в предложении играют синтаксическую роль наречия, выступая в качестве обстоятельства или приименного атрибута. Часть этих сочетаний требуют при себе зависимого, обычно именной группы в родительном падеже: такие единицы сближаются с предлогами.

За бортом описания остаются фразеологические или полуфразеологические сочетания (впрочем, немногочисленные) типа *видывать <видать> виды, делать вид* (*что*), *не подавать виду, ставить на вид и* некоторые другие. Однако и без этих сочетаний число микросинтаксических единиц, образованных словом *вид,* достаточно внушительно. Перечислим основные из них: (1) в *виде,* (2) в *виду,* (3) *для виду,* (4) из *виду,* (5) *на вид,* (6) *на виду,* (7) *по виду,* (8) *под видом,* (9) *при виде,* (10) с *виду.*

Ниже эти единицы будут кратко прокомментированы и проиллюстрированы примерами.

2.1. В виде

Эта микросинтаксическая единица прототипически выступает в качестве предлога, управляющего генитивом:

(1) *Над ними поднималась в небе луна в* ***виде*** *косвенно обращенного серпа из яркого червонного золота.* (Н. В. Гоголь).

(2) *Фонарь был пышный и старинный, / Но в* ***виде*** *женщины чугунной* (Н. А. Заболоцкий)

Семантические и синтаксические особенности в *виде* будут подробнее рассмотрены ниже, в разделе 3.

2.2. В виду

Единица в *виду*, внешне отличающаяся от предыдущей, казалось бы, незначительно: тем, что словесный элемент *вид* стоит в ней не в предложном, а в местном падеже, характеризуется совсем другими, чем у в *виде*, семантическими и синтаксическими свойствами. В изолированном виде (вне специального контекста типа *иметь*(*ся*) *в виду*) эта единица встречается редко и обозначает факт присутствия наблюдателя, имеющего возможность видеть объект, выражаемый генитивной ИГ при в *виду*, ср. пример из НКРЯ (3) и СинТагРус'а (4):

(3) *Нехлюдов вернулся на тротуар и, велев извозчику ехать за собой, пошел в* ***виду партии*** (Л. Н. Толстой).

(4) *Миновав высоковольтную линию, уже в* ***виду сосняка****, он почувствовал, что ему мучительно идти прежней дорогой, и направился в обход березовым перелеском* (Ю. Нагибин).

Конструкции *иметь в виду и иметься в виду*, на наш взгляд, представляют собой отдельные микросинтаксические единицы, в которых в *виду* содержится как составная часть. В этих конструкциях объект в *виду* переходит от существительного к глаголу (ср. *имеется в виду сосняк*), а идея непосредственного видения объекта выцветает.

Добавим, что еще в начале XX века в текстах регулярно встречалась единица в *виду с* другим значением: 'вследствие', 'из-за', что адекватно отражено в НКРЯ; ср, например,

(5) *В виду истощения котиковых лежбищ на Командорских островах, в текущем году предполагается допустить к убою не более 400 котиков* («Московские ведомости», 1911).

По ныне действующим правилам эта единица пишется в одно слово (*ввиду*), но это не более чем орфографический каприз. Недаром в современных письменных текстах ошибки, когда вместо в *виду* пишется *ввиду* (и наоборот), носят массовый характер. Формально говоря, мы не должны включать в число микросинтаксических единиц (они по определению неоднословные) предлог *ввиду, но* содержательно это именно такая единица[2].

2.3. Для виду

Эта единица, как и ее чуть более современный морфологический вариант *для вида*, где партитивный падеж заменен родительным, выступает всегда как наречие, играющее роль обстоятельства причины; ср.

(6) *После изгнания в 1492 году, после всех конфискаций часть из них* [евреев. — Авт.] *осталась в Испании, крестившись* ***для виду***. (Д. Рубина);

(7) … *Теперь улыбнись хоть для вида, / Хоть слово промолви со зла.* (А. Тарковский).

---

[2] В области микросинтаксиса орфографические курьезы не уникальны. Другим примером является предлог *насчет,* который по современным правилам пишется слитно. Однако число ошибок здесь столь же велико, сколь и в случае с *ввиду.* Например, из 10 первых примеров основного корпуса НКРЯ, выдаваемых по запросу НА+СЧЕТ$_{sg,acc}$+GEN, три представляют фразы, где должно стоять *насчет* (*На счёт болезней он всё прекрасно понимает*), а в текстах начала XX века раздельное написание – просто норма. Кстати, единица *насчет* имеет все основания писаться раздельно и сейчас: выражения *насчет этого* и *на этот счет,* содержат одну и ту же единицу.

У единицы *для виду* отсутствуют активно выражаемые валентности (невозможно сказать **для вида Ивана* или **для вида сочувствия*, хотя семантически она, передавая идею притворства, имеет минимум четыре валентности: (i) кто притворяется, (ii) что хочет продемонстрировать, (iii) каким способом он это демонстрирует и (iv) кому он это демонстрирует. Обычно при *для виду* пассивно реализуется третья из данных валентностей: в (6) и (7) это, соответственно, глаголы *креститься и улыбнуться*.

Добавим для полноты картины, что единица *для виду* относится к тем агломератам, которые способны разрываться в диалогах, где ответная реплика подвергает сомнению уместность использованного собеседником выражения:

(8) *Он сделал это для виду. — Для какого виду? Это было* совершенно серьезно.

2.4. Из виду

Эта микросинтаксическая единица, как и ее вариант с родительным падежом из *вида,* также ведет себя как наречие, не допуская практически никаких зависимых. В целом единица обычно заполняет валентность исходной точки, обозначая удаление из поля зрения наблюдателя (присутствие которого в семантической структуре, соответствующей фраземе, обязательно). Перечень глаголов, допускающих такое заполнение валентности исходной точки, ограничен: это (i) *упускать, выпускать, терять* и (ii) *исчезать, пропадать, выпасть, скрыться*, а также некоторые их синонимы:

(9) *Надо умудриться — упустить в Москве из виду такого человека, как Лучников* (В. Аксенов)*;*

(10) *Маина тронулась и скоро скрылась из виду.* [И. Грекова].

В некоторых случаях из *виду* используется при глаголах движения, ориентированных на исходную точку:

(11) *Вдруг они ушли из виду, ушли вниз* (Г. Владимов).

(12) *Оные казаки, оборотя лошадь свою, поскакали назад, и поворотя в переулок уехали из виду* (А. С. Пушкин, История Пугачева).

Другие же глаголы с бесспорной валентностью исходной точки из *виду* не допускают: **сбежать из виду, *переместиться из виду в укрытие и* т. п. исключены.

В случае переходных глаголов речь может идти только о ненамеренных действиях субъекта: нельзя *рассчитывать упустить из виду. В* случае непереходных глаголов субъект не может совпадать с наблюдателем: если *Иван скрылся из виду*, то он исчез из поля зрения кого-то другого. Само же действие может быть и намеренным: *спрятались из виду, надо скрыться из виду.*

Спорадически из *виду* могут принимать отглагольные существительные: *исчезновение из виду, потеря из виду* и т. д.

Добавим, наконец, что из *виду* порождает особую глагольную конструкцию *упустить <выпустить> из виду*: тут поле зрения отражается метафорически, а сама конструкция обладает специфичным управлением (на союз *что* и инфинитив):

(13) *Я совершенно упустил из виду, что этот дебил может оказаться занят* (А. Геласимов).

(14) *Как ни странно, на «Ладоге» из-за спешки упустили из виду приобрести приемник для кают-компании* (Л. Лагин).

2.5. На вид — по виду — с виду

Эти три квазисинонимичные микросинтаксические единицы выступают как оценочные наречия, указывающие, что объекту приписывается некоторое свойство или состояние на основе визуального впечатления наблюдателя, ср.

(15) *Юная продавщица в белоснежном халате,* ***на вид*** *прохладная и потому приятная, работает молча, мягко, равномерно* (Ф. Искандер);

(16) *В кабинете директора, когда туда опять вошёл Иван, сидела некая милая женщина,* ***по виду*** *врач* (В. Шукшин);

(17) *Ведь это только с* ***виду*** *большие артисты — люди, уверенные в себе* (Сати Спивкова).

Между этими тремя единицами есть тонкие семантические различия, исследовать которые мы сейчас не будем. Осторожно заметим лишь, что последняя фразема — с *виду* — чаще двух других предполагает, что оцениваемого наблюдателем свойства у его субъекта на самом деле нет (ср. пример (17)).

Системное отличие этих фразем друг от друга обусловлено, в частности, наличием разных наборов параллельных выражений у одних (*на вид и по виду*, ср. *на вид — на вкус — на ощупь — на слух — *на*

запах vs. *по виду — по вкусу — по запаху — *по ощупи*) и их отсутствием у с *виду*. Кроме того, в случае *по виду в* текстах часто имеет место контаминация интересующей нас фраземы со свободным сочетанием *по виду* 'в соответствии с' (ср. *Он узнал его только по виду*), которое бывает очень трудно отличить от интересующей нас фраземы *по виду*.

Ни одна из фразем не допускает при себе зависимых: **с его виду, *на вид Маши* и т. п. Исключение составляет вариант *по внешнему виду*, как в примере

(18) *… флегматично покуривает трубку-носогрейку смуглый коренастый капитан, по внешнему виду итальянец или грек* (В. П. Катаев).

2.6. На виду

Эта синтаксическая фразема имеет два отчетливо различных значения (назовем их, соответственно, *на виду*[1] и *на виду*[2]). Обе единицы ведут себя как наречия, однако *на виду*[1] прототипически выступает как обстоятельство образа действия, ср.

(19) *Я играю на гармошке у прохожих на виду* (А. Тимофеевский)

— т. е. 'играю так, что прохожие могут меня видеть', в то время как *на виду*[2] выступает прежде всего как заполнитель локативной валентности глагола, ср.

(20) *Осмелев, он стоял совсем на виду, по колено в воде, и вдруг понял, что не так расселины придают тому берегу вид неприступности, как его нагота* (Г. Владимов).

Как *на виду*[1], так и *на виду*[2] имеют семантическую валентность объекта, факультативно выражаемую предложной группой у + $NP_{род}$ (как в (19)). Этот объект, он же наблюдатель, соответствует субъекту входящего в семантическую структуру предиката 'видеть', а объект этого предиката ('кого видят') совпадает с субъектом вершинного глагола: в (19) это *я*, а в (20) — *он*. *У на виду*[1] этот субъект — почти исключительно человек, *на виду*[2] уместно для любого физического объекта.

Добавим к сказанному, что в семантике предложений с *на виду*[1] часто содержится компонент неодобрения того, что имеет место возможность видеть субъекта действия, которое происходит «на виду», со стороны говорящего или наблюдателя: неодобрение может вызывать публичная демонстрация себя этим субъектом (что должно быть стыдно или чересчур дерзко).

2.7. Под видом

Эта микросинтаксическая единица выступает как предлог, управляющий генитивом, и, как и в предыдущем случае, имеет два разных значения; назовем их *под видом*[1] и *под видом*[2].

Под видом[1] характеризует ситуацию, в которой один ее участник намеренно выдается за другого с намерением ввести в заблуждение наблюдателя. Этот участник может быть как человеком (21), так и любым другим физическим объектом (22):

(21) *Под видом рабочих они по подложным документам проникли в клуб и установили в трех разных местах адские машинки* (Ю. Домбровский);

(22) *Просит она чаще всего «мартель», хотя все знают, что под видом коньяка ей приносят остуженный чай* (В. Овчинников).

Под видом[2] вводит ситуацию, протагонист которой заявляет цель, отличную от его реальной цели (как правило, заявленная цель «благовиднее» реальной). *Под видом*[2] имеет близкий синоним — микросинтаксическую единицу *под предлогом*. Зависимым элементом у *под видом*[2] может быть существительное со значением целесообразного действия:

(23) *Долго он потом, под видом перевязки, хаживал в старый барак иль водил свою зазнобу в лес, по грибы* (В. Астафьев);

Изредка валентность *под видом*[2] может заполняться местоименным прилагательным, выступающим определением при *видом*:

(24) *Любопытно будет знать, под каким видом будут теперешние рапповцы оттирать мою пьесу «Главный инженер»?* (Вс. Иванов).

Такая реализация нашей фраземы порождает еще одну единицу *ни под каким видом*, имеющую синонимы в том же классе микросинтаксических единиц: *ни в коем случае и ни под каким предлогом.*

2.8. При виде

Эта микросинтаксическая единица также ведет себя в целом как составной предлог, управляющий NP в родительном падеже. Эта единица, по-видимому, теснее других, рассмотренных выше, связана с семантикой видения как процесса. Объект видения выражается как раз зависящей от *при виде* именной группой, а субъект видения — это

субъект действия или процесса, выражаемого предикатом, заполняющим пассивную валентность *при виде*. Тем самым во фразе

(25) *При виде вошедшего сидящий за столом побледнел* (М. Булгаков)

говорится, что *сидящий* увидел *вошедшего*. Выражение *при виде* некомпозиционально в том смысле, что содержит семантический компонент начинательности, относящейся к реакции субъекта: в (25) сидящий побледнел, **как только** вошедший попал в его поле зрения. Поэтому семантически небезупречны фразы типа

(26) *#При виде вошедшего сидящий продолжал курить*.

Заметим в заключение, что фразема *при виде* относится к единицам, которые не до конца превратились в неделимый предлог: во-первых, бывают ситуации, когда объект видения выражается притяжательным прилагательным, зависящим от фрагмента *виде,* как в (27), а, во-вторых, внутрь конструкции могут вклиниваться и другие определения (28–29):

(27) *Но и радости при их виде не испытывал Данилов* (В. Орлов).

(28) *При одном лишь виде слабого человека он* (авторитарный характер. — Авт.) *испытывает желание напасть, подавить, унизить* (Э. Фромм, пер. А. Александровой);

(29) *Уж сердце в радости не бьется* / *При милом виде мотылька* (А. С. Пушкин).

3. Конструкция *в виде* и ее разновидности

Выражение в *виде в* целом выступает как составной предлог, выступающий как при имени, так и при глаголе; ср.

(30) *Исследователи обнаружили под водой необычное сооружение в виде дуги*;

(31) *Дом он построил в виде рыцарского замка*.

Приименное употребление группы в *виде чего-л.* является первичным: оно не требует каких-либо предварительных семантических условий от существительного, которому подчиняется, в то время как для глагола такое употребление вторично: оно характеризует внутренний объект такого глагола (в смысле Апресян 2009:492 и сл.) и не может

относиться к внешнему объекту. Так, (31) можно считать перифразой предложения

(31') *Он построил дом в виде рыцарского замка.*

В то же время присоединить в *виде чего-л. к* глаголу, объект которого является внешним (в частности, к деструктивному глаголу), практически невозможно:

(32) **Дом он уничтожил в виде рыцарского замка.*

В пользу трактовки этой единицы как предлога говорит тот факт, что при нем обязательно заполнение валентности, исключительно в родительном падеже. При этом предлоге допускаются формы местоимения третьего лица на *н-* (типа *него*), которые даже предпочтительнее, чем формы типа *его*; ср.

(33) *… натыкаюсь на позабытый фонарик с голой теткой. Ручка в виде нее, которую полагается держать за талию* (Мариам Петросян);

(34) *Даже о жуке имеются сведения, что в виде него почитался сам Зевс* (А. Ф. Лосев).

Как известно, предложно-именные сочетания, порождающие составные предлоги, не всегда полностью утрачивают независимость своих элементов. Например, единицы в *пользу* или *по поводу,* которые обычно относят к составным предлогам, свободно употребляются с адъективным определением к существительным (как правило, такое определение — местоименное или притяжательное прилагательное; ср. в *пользу Пети = в Петину пользу, в пользу меня = в мою пользу, в пользу кого = в чью пользу, по поводу этого = по этому поводу* и т. д. Единица в *виде* продвинулась дальше по пути создания составного предлога, хотя единичные определения при существительном встречаются и здесь:

(35) *В чьем виде, избранном тобой, / Явился б ты передо мной? / — Я был бы ангел твой хранитель!* (И. И. Козлов).

Обратим внимание на следующую семантическую особенность в *виде*: эта единица практически не может вводить конкретно-референтной именной группы (в смысле Е. В. Падучевой, см. например, Падучева 1985). Даже в редких примерах из НКРЯ типа

(36) *Я его представлял в виде Петьки придурка, только с бородой* (А. Приставкин)

речь идет о человеке, похожем на Петьку, а не о нем самом.

Заметим, что однозначно установить, какое из двух основных значений слова *вид* сформировало единицу в *виде*, вряд ли возможно. С одной стороны, в толковании в *виде* необходим смысловой компонент 'выглядеть', т. е. компонент значения ВИД 1. *С* другой стороны, в примерах типа

(37) *В природе нитрат натрия встречается в виде минерала чилийской селитры* (*нитронатрит*) (НКРЯ, газетный подкорпус)

выражение в *виде* явственно передает идею разновидности, т. е. соотносится и с ВИД 2.

Еще одно интересное свойство единицы в *виде* проявляется в своеобразии его валентной структуры. У этой единицы, как и у мотивирующего слова *вид,* имеются две семантических валентности: валентность субъекта (что/кто имеет вид) и валентность содержания (каков вид). В выражении типа *вид креста* родительный падеж активным образом выражает валентность субъекта, а в выражении в *виде креста* слово *крест в* родительном падеже активным образом выражает валентность содержания. Это сближает пару *вид — в виде с* парой *причина — по причине,* которая в числе прочих подробно рассматривалась И. М. Богуславским (см., например, Богуславский 2005).

Отметим ещё, что выражение в *виде* порождает отдельную синтаксическую фразему (назовем ее в *виде*^{2}), которая по семантике не укладывается в рассмотренную выше картину. Мы имеем в виду интерпретационную единицу, в которой валентность, выражаемая генитивом при в *виде*, заполняется существительными типа *исключение, одолжение, поощрение, наказание и* др.. Здесь речь идет не о том, что нечто представлено как что-то другое, а о том, что нечто интерпретируется специальным образом. Сравним примеры (38), где представлена основная единица в *виде,* и (39), где фигурирует в *виде*^{2}:

(38) *Нарушение, допущенное более трех раз, влечет ответственность в виде исключения из состава союза;*

(39) *В виде исключения его не стали выгонять из союза.*

У основной единицы в *виде и у в виде*^{2} — разные синонимы: у первой — в форме, наподобие, а у второй — в порядке. Обратим внимание также на то, что в *виде*^{2} нормально выступает как приглагольное обстоятельство, но не как приименной атрибут.

Единица в *виде* имеет, по нашим наблюдениям, еще и третье значение. Мы уже видели на примере (35), что внутрь основного значения этой единицы может вклиниваться прилагательное. В таких случаях семантика в *виде и* состав ее валентностей не меняется: в *моем виде* означает то же, что и в *виде меня*. Посмотрим теперь на некоторые другие примеры:

(40) *Это была в чистом виде ознакомительная поездка;*

(41) *При этом казалось, что поспешность его, даже заботливая поспешность, вызвана тем, что он хочет донести до кого-то эти лепёшки ещё в тёплом виде*. (Ф. Искандер).

По нашему убеждению, в этих примерах мы имеем дело с особым значением нашей конструкции (обозначим его через в *виде*[3]), у которой посредством прилагательного выражается отсутствующее в ранее рассмотренных значений валентность признака исходного субъекта (того, что имеет вид). Легко заметить, что в (40)–(41) невозможно вместо в виде использовать ни наподобие, ни в порядке. Весьма выпукло различие между в виде и в виде[3] можно обнаружить, сравнив две близких фразы:

(42) *Актер появился на сцене в нетрезвом виде* и

(43) *Актер появился на сцене в виде нетрезвого человека*.

В (42) мы имеем в *виде*[3] (здесь одному персонажу приписывается признак 'нетрезвый'), а в (43) присутствует стандартное в *виде* (здесь *нетрезвый человек* выступает в генерическом качестве).

Таким образом, единица в *виде* оказывается весьма специфичной даже на фоне других нетривиальных синтаксических фразем: в отличие от всех адвербиалов, рассмотренных выше, она имеет три уникально соотносящихся между собой значения.

Литература

1. *Апресян Ю. Д.* (2009), Исследования по семантике и лексикографии. Т. I: Парадигматика. М., Языки слав. культуры.
2. *Богуславский И. М.* (2005), Валентности кванторных слов. // Квантификативный аспект языка. М., 2005, с. 139–165.
3. *Иомдин Л. Л.* (2015), Конструкции микросинтаксиса, образованные русской лексемой *раз*). SLAVIA, časopis pro slovanskou filologii, ročník 84, 2015, sešit 3, s. 291–306.

4. *Иомдин Л. Л.* (2017a), Между синтаксической фраземой и синтаксической конструкцией. Нетривиальные случаи микросинтаксической неоднозначности. SLAVIA, časopis pro slovanskou filologii, ročník 86, sešit 2-3, s. 230–243.
5. *Leonid Iomdin L.* (2017b), Microsyntactic annotation of Corpora and its use in Computational Linguistics Tasks. Jazykovedný časopis, ročník 86, číslo 2, s. 169–178.
6. *Иомдин Л. Л.* (2018), Еще раз о микроконструкциях, сформированных служебными словами: То и дело. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодноǐ Международноǐ конференции «Диалог». М.: Изд-во РГГУ. Вып. 17 (24). С. 267–283.
7. *Падучева Е. В.* (1985), Высказывание и его соотнесенность с действительностью. М.: Наука.

References

1. *Apresjan Ju. D.* (2009), Issledovanija po semantike i leksikografii. T. 1: Paradigmatika. [Studies in semantics and lexicography. Vol. 1: Paradigmatics]. Moscow, Jazyki slavjanskix kul'tur.
2. *Boguslavsky I. M.* (2005), Valentnosti kvantornyx slov [Valencies of quantifier words. // Kvantitativnyj aspect jazyka. Moscow, p. 139–165.
3. *Iomdin L. L.* (2015), Konstruktsii mikrosintaksisa, obrazovannye russkoj leksemoj *raz* [Microsyntactic Constructions Formed by the Russian Word *raz*]. SLAVIA, časopis pro slovanskou filologii, ročník 84, sešit 3, s. 291–306.
4. *Iomdin L. L.* (2017a), Meždu sintaksičeskoj frazemoj i sintaksičeskoj konstruktsiej. Netrivial'nye slučai mikrosintaksičeskoj neodnoznačnosti. [Between the Syntactic Idiom and Syntactic Construction. Complicated Cases of Microsyntactic Ambiguity]. SLAVIA, časopis pro slovanskou filologii, ročník 86, 2017, sešit 2–3, s. 230–243.
5. *Iomdin Leonid.* (2017b), Microsyntactic annotation of Corpora and its use in Computational Linguistics Tasks. Jazykovedný časopis, ročník 86, číslo 2, s. 169–178.
6. *Iomdin L. L.* (2018), Ešče raz o mikrokonstrukcijax, sformirovannyx služebnymi slovami: To i delo. [Once Again On Microsyntactic Constructions Formed With Functional Words: To i delo 'every now and then'. // Komp'juternaja lingvistika i intellektual'nye texnologii: Trudy meždunarodnoj konferentsii «Dialog–2018» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2018»]. Issue 17 (24). P. 267–283.
7. *Paducheva E. V.* (1985), Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju [The utterance *and its* Relation to Reality]. Moscow, Nauka Publishers.

Иомдин Леонид Лейбович
Институт проблем передачи информации им А. А. Харкевича РАН (Москва, Россия)
Leonid Iomdin
Institute for Information Transmission Problems (Moscow, Russia)
E-mail: iomdin@iitp.ru

AUTOMATIC TERM EXTRACTION — EFFICIENCY OF SELECTION AND RELEVANCE OF EXTRACTED TERMS AS APPLIED TO THE SPECIALIZED CORPUS OF LIBRARY AND INFORMATION SCIENCE IN SLOVENE LANGUAGE

Abstract. A specialized synchronous text corpus had been designed and constructed to support the research in the field of library and information science terminology and dictionary construction in Slovene language. The size of the corpus has exceeded 4 million words represented in 725 Slovene technical and scientific texts of the subject field. It supports a variety of specialized search methods and statistic computation, including automatic term extraction. Efficiency of automatic selection and relevance of extracted multi-word terms have been studied. The results, presented in the paper, show a considerably high quality and reliability of the automatic term extraction method.
Keywords: corpus linguistics, text corpus, text mining, automatic term extraction, library science, terminology, Slovene language.

1. Introduction

In recent past different extensive quantitative and semantic ana- lyses have been successfully carried out making use of a smaller specialized corpus, providing precious data to terminologists and lexi- cographers. The construction of two dictionaries, their adaptation and the evaluation of long established terms compared against recent practical use in scientific texts have proven a great help and undis-pensable tool. The inventory of recent, often unsettled terms, remains a crucial, painstaking and time consuming supporting work. In this respect appropriate corpus analyses are advanced enough to provide good and useful results. The main objective of the research is to test the less known features of automatic term extraction in subject speci- fic terminology belonging to a highly inflected language. An evaluati- ve automatic term extraction has been prepared and the assessment of results analysed as for their grammatical correctness, subject specific relevance and general efficiency of the procedure.

2. Specialized corpus in Slovene language

«*Korpus bibliotekarstva*»[1], a synchronous specialized text corpus, represents the technical language in the specific field of library and

[1] Kanič, I. (2011), Slovenski besedilni korpus bibliotekarstva. Library: open space for dialogue and knowledge : proceedings, Maribor, Oct. 20-22, 2011. [Conference paper]. URL: http://eprints.rclis.org/16730/ 1/Kanic-Besedilni%20korpus%20bibliotekarstva.pdf

information science, shared among the community of practitioners, researchers, translators, teachers and students in the present and very recent past in Slovene language. Primarily designed and constructed in 2011 to support the work of the *Commission on Library Terminology* in the frame of the *Slovene Library Association* it assisted termi- nologists and lexicographers to discover and determine the exact inventory and verify the occurrence of words and phrases in technical and scientific texts, enabling researchers to obtain a variety of struc-tured lists of words and phrases, be it in their original form or lemma-tized and tagged with part of speech labelling. Insight into the occurrence of words, their collocations and frequency in technical and scientific texts support immensely the work of terminologists, thus the corpus has proven an indispensable and powerful tool for the prepa-ration of modern dictionaries and updating of the existing mono-lingual explanatory and multilingual translating dictionaries of library terminology.

The corpus was upgraded in 2012 and 2018, exceeding an inven- tory of 4 million words now, excerpted from 725 texts by more than five hundred authors. As a rule, all the works included had been origi- nally published in electronic form, mostly born digital or digitized by publishers. Data capture was focused predominantly on selected texts published in the last three decades, depending on their availability, of course. Original texts in Slovene language were chosen as a rule, rare translations are an exception.

2.1. Particularities of the Slovene language

Slovene or Slovenian (*slovenski jezik* or *slovenščina*) is an Indo-European language with a highly developed inflectional system, it belongs to the group of South Slavic languages, spoken by not more than 2.1 million Slovenian people and is one of the 24 official and working languages of the European Union. The standard Slovenian orthography makes use of the ISO basic Latin alphabet, it has no letters x, y, w, and q but the following three letters representing palato-alveolar sibilants are added: č, š, ž.

Linguistic particularities of the language, represented in the cor- pus, present an important issue regarding the term formation and automatic terminology extraction. Our research focused on multi-word terms only so nouns and adjectives are expected to be the pre- dominant constituent parts of speech. Adverbs, numerals, pronouns and other parts of speech were detected in a few examples only.

Nouns in Slovene are either of masculine, feminine or neuter gen- der and are declined for six cases and three numbers. Each gender has different and specific declension patterns.

Adjectives and most pronouns decline for three genders, three numbers and six cases. The adjective expresses three main ideas: qua- lity, relation and possession, the majority of adjectives are of the first kind this being crucial for the term formation in our case.

Verbs are conjugated for 3 persons and 3 numbers, there are 4 tenses (present, past, pluperfect, and future), 3 moods (indicative, im- perative, and conditional) and 2 voices (active and passive). Except for the present tense the verb shows the gender as well.

Slovene is one of the rare Indo-European languages which apart of singular and plural still uses *dual* — a grammatical number used when referring to precisely two persons, objects or concepts, iden- tified by the noun or pronoun. Grammatically, Slovene retains forms expressing the dual number in nouns, adjectives, pronouns and verbs, in addition to singular and plural. In term formation it is not signi-ficant but this agreement based on grammatical number and gender is an important issue to be taken into account in automatic terminology extraction.

3. Tools and procedures

3.1. Sketch Engine

Sketch Engine is a user-friendly and efficient web-based corpus manager and text analysis software, its purpose is primarily to enable people studying language behaviour to search large text collections according to complex and linguistically motivated queries. Currently, it supports and provides several hundreds of reference and special corpora in almost one hundred languages, allowing individual users to define, construct, maintain and explore their own user-defined cor-pora. Particular features of the tool are corpus building and mana-gement including part-of-speech tagging and lemmatization, word sketches and word sketch differences based on comparative collo-cation analysis, automated thesaurus, concordance search, colloca-tion search, word lists with frequencies and normalised data, n-grams, terminology extraction, diachronic analysis and trends, and parallel corpus (bilingual) facilities. Sketch Engine has been used by major British and other publishing houses for producing dictionaries such as well renown Macmillan English Dictionary, Dictionnaires Le Robert, Oxford University Press or Shogakukan and four of the UK's five biggest dictionary publishers use Sketch Engine[2].

[2] Wikipedia. URL: https://en.wikipedia.org/wiki/Sketch_Engine

3.2. Keywords and multi-word terms

In the context of corpus analysis *keywords* are single words which appear in the focus corpus more frequently than they would in general language represented by the reference corpus, in our case both corpora containing texts in Slovene language. Typically, the largest reference corpus containing most recent texts will be selected. On the contrary, *terms* are multi-word units (phrases) that fulfil two conditi-ons. Like keywords they appear in the focus corpus more frequently than they would in a reference corpus, and they have a structure allowed for terms in the language as set in the term grammar of the corpus, specific for the language concerned. The term grammar typically focusses on identifying noun phrases[3]. In both cases the keyness score of a word or multi-word term is calculated according to the formula taking into consideration the normalized frequency (per million) of the word/term in the focus corpus, the normalized frequency (per million) of the word/term in the reference corpus and the simple maths smoothing parameter. **Simple maths** is a method for identifying keywords of one corpus compared with another one. It includes a variable which allows the user to turn the focus either on higher or lower frequency words/terms[4]. The top keywords/terms identify and reflect the domain of the focus corpus very well and can be used to explore the vocabulary of the subject field. Key terms are multi-word noun phrases typical of a corpus, the first thousand ex- tracted key terms have been studied. The current research has been limited to multi-word terms exclusively.

3.3. Term extraction

Term extraction, often referred to as terminology extraction, is an automatic method of analysing text in order to identify phrases which fulfil the criteria for terms. Terminology extraction has its use in translation and terminology management but also in text analytics where it is used for topic modelling, data mining and information re-trieval from unstructured text. In our case it has primarily been used for terminology work and to augment the existing dictionary of libra-ry terminology.

Good term extraction provides a clean list of terms that requires very little manual cleaning. Many traditional methods relying mainly on the frequency in the focus text can only extract term candidates which the user has to go through and clean "manually". The manual cleaning can largely be

[3] Sketch Engine. URL: https://www.sketchengine.eu/user-guide/glossary/

[4] Sketch Engine. URL: https://www.sketchengine.eu/documentation/simple-maths/

avoided by making use of linguistic criteria in combination with statistics. If a phrase is to qualify as a term, it should fulfil two crucial criteria: it matches the structure that a term in the language can have, and it appears more frequently in the focus text than in general text[5].

In the context of text corpora, an n-gram, also called multi-word unit, will typically refer to a sequences of words. This study did not deal with unigrams, only polygrams (i. e. bigrams, trigrams and tetra- grams) have been studied. In some cases the items inside an n-gram may not have any relation between them apart from the fact that they appear next to each other, so it is not considered a term (frequent collocation only). Beside its importance for machine translation and language learning the study of n-grams is important and useful for terminology and lexicography.

4. Extracted terms

4.1. Preprocessing

To provide the highest level of trustworthiness, correct results and efficient analyses the corpus had been prepared respectively. For lemmatization of the tokens and part-of-speech tagging two existing tools, already successfully tested and implemented in some huge reference corpora of Slovene texts, performed well.

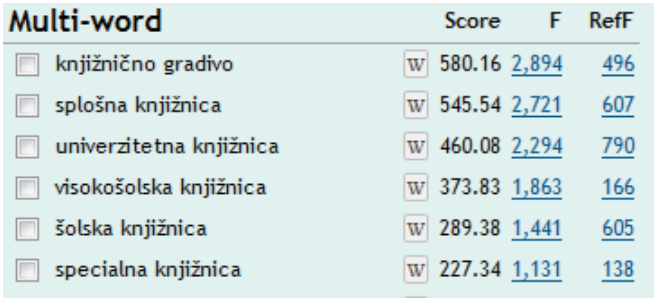
A language specific tokenizer provided by Sketch Engine proved adequate enough; only whitespace characters are recognized as token boundaries which is sufficient for the task. Regarding the complexity of the Slovene morphology and the rich inflections found in tokens the parts-of-speech tagging and term extractor proved an excellent combi-nation as the given result contains hardly any noise and does not require much manual cleaning. POS tagging provided tags with information about the part of speech and morphological and gramma-tical information regarding number, gender, case, tense, lower/upper case and the like. In addition lempos suffixes provide information whether a lemma represents a noun, verb, adjective, adverb, pronoun, adposition, conjunction, particle, interjection, numeral, abbreviation or some other possible residual. For this language specific processing specialized *MULTEXT-East Morphosyntactic Slovenian Specification version 4*[6] dataset was applied, available for Slovene corpora in Sketch Engine.

[5] URL: https://www.sketchengine.eu/user-guide/user-manual/term-extraction/

[6] URL: https://www.sketchengine.eu/slovene-tagset-multext-east-v4/

4.2. List of extracted terms

From the perspective of a terminolographer and lexicographer the term extractor is the most fascinating tool, a very useful basis and starting point for the recognition and stocktaking of technical termino-logy in recent texts where fresh, not yet established terms appear. Complying with supplied and incorporated specific linguistic instru-ctions and rules the module is fully competent to extract extensive lists of n-grams, qualified as technical terms, from a text or preferably a corpus. The proposed terms are an estimated proposal only of what interesting terms could be, a thorough human evaluation and conside-ration is still needed to correct eventual errors and exclude infor-mation noise. A list of (multi-)word terms is proposed (Fig. 1), it is up to the user to define the size of the list. We have chosen the first most relevant thousand terms. Each term is accompanied by helpful addi-tional information. (F) shows the frequency of occurence in the focus corpus, however, high frequency is not always a proof of a term as the items (words) inside an n-gram may not have any relation between them apart from the fact that they appear next to each other (frequent collocation). There have only been found 21 such cases in our study. (RefF) denotes the frequency of the proposed term in the reference corpus, most common a general one. (W) is a welcome helpful hint to the Wikipedia reference regarding the term or a related subject. The terms are sorted by score, which depends on the keyness score calcu-lated by the simple maths method (Score).



Multi-word		Score	F	RefF
knjižnično gradivo	W	580.16	2,894	496
splošna knjižnica	W	545.54	2,721	607
univerzitetna knjižnica	W	460.08	2,294	790
visokošolska knjižnica	W	373.83	1,863	166
šolska knjižnica	W	289.38	1,441	605
specialna knjižnica	W	227.34	1,131	138

Fig. 1. **List of extracted polygrams.**

4.3. Quantitative assessment

In general, the syntax of the Slovene language is rather loose concerning the word order, but in term construction two basic rules apply strictly: the noun is preceded by the adjective (e.g. *knjižnični katalog, virtualna knjižnica*)

and the verb is followed by the noun representing its object (e.g. *katalogizirati knjigo, indeksirati članek*). As a rule, the most frequent and important constituent element in n-grams, expected to form multi-word terms, would be nouns, adjectives and verbs, other parts of speech exceptionally only (e.g. adverb — *strokovno obdelati*, numeral — *format* 4^0). A brief statistical overview of the corpus shows the following potential elements for the formation of multi-word terms. Automatic part of speech tagging has identified lemmas as follows: 12.374 nouns, 4.999 adjectives, 2.204 verbs, 1.065 adverbs, and altogether a few hundreds of prepositions, numerals, con- junctions, abbreviations, particles and pronouns. It has to be stressed, however, that there might be minor differences in the count since the automatic part of speech tagging still cannot discern parti-cular forms of homographs resulting from inflections without human intervention (e.g. *dela* may be a form of the verb *delati*, a noun *delo* or *del*; *uporabnikov* may be a noun or an adjective, etc.). Nonetheless, elabo- rate grammatical rules governing the part of speech tagger help in resolving syntactical ambiguities successfully.

As stated, only polygrams have been studied. In the process of automatic terminology extraction the first thousand terms were represented as follows: 862 bigrams (e.g. *analiza citiranja, digitalna vsebina, elektronska knjiga, kataložni listek, odpis gradiva*), 109 trigrams (e.g. *abecedni imenski katalog, dostopnost knjižničnega gradiva, knjižnično informacijsko znanje*) and 29 tetragrams (e.g. *bibliotekarska in informacijska znanost, trajno ohranjanje digitalnih virov, uporabnik s posebnimi potrebami*). There have been no penta- grams extracted.

In this selection of 1000 most frequent n-grams their frequency ranges between 2.894 occurencies (*knjižnično gradivo*) and 47 oc- curencies (e.g. *specialni knjižničar, družba znanja*, etc.)

4.4. Quality assessment

The reliability of automatic extraction and quality of extracted terms had been set as primary goals. There have been some experiences in this field in the past but often failed in providing high quality results. Many polygrams extracted had no characteristics of terms and syntactical and/or morphographic features of constituent words were not correct. Often, the results were of very poor quality and did not encourage further processing.

The enhanced terminology extractor and elaborated grammatical rules have highly influenced the quality of extraction. The results were evaluated manually. Out of 1.000 extracted polygrams only 21 showed grammatical erors in their grammatical structure. They were rejected. 462 terms proved

syntactically and morphologically perfec but their semantical meaning did not fit into the subject field (library and information science). But the majority, 517 terms, were judged as correct and relevant terms in the subject field.

These chosen 517 terms were collated with the existing dictio-nary database showing that 312 of them (60,3 %) already figured in the dictionary as accepted terms, so the remaining 205 will be discussed as potential candidates to enter the dictionary. The quality and preci-sion of extraction and the relevance of results has surpassed the expectations by far.

5. Conclusion

Modern and up-to-date dictionaries, in particular dictionaries of specific subject fields, are bound to explore their potentional vocabu-lary by text mining. Recent developments in automatic terminology extraction have equipped terminographers and lexicographers with exceptionally efficient tools providing high quality results. A syn-chronous specialized text corpus, representing the technical language in the specific field of library and information science in Slovene language, has been constructed to support the lexicographers in following the dynamic changes in the language of technical and scinntific publications. The size of the corpus has exceeded 4 million words represented in 725 Slovene technical and scientific texts thus representing a rich and lexically diverse database for text mining. The aim of the research was to test the efficiency of selection and rele-vance of extracted terms through the process of automatic extraction. The results proved high reliability of extraction and excellent quality of terms extracted, evaluated afterwards manually and by comparison with the existing dictionary. Consequently, the method will become a part of the dictionary construction process.

Even though the corpus already covers a wide range of different types of documents ranging from doctoral dissertations and scientific articles to conference papers and has reached a rather wealthy selection of words used by numerous Slovene authors, a dynamic growth of the corpus by inclusion of recently published texts within the wider field of library and information science remains a further goal. Thus the corpus and automated term extraction will gain the representative role in the inventory and study of the Slovene library terminology and further lexicographic work.

References

1. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* (2004), The Sketch Engine. Berlin: ResearchGate. (accessed Februar 24, 2019) URL:https://www.researchgate.net/profile/Adam_Kilgarriff/publication/260387608_ITRI-04-08_the_sketch_engine/links/54e0d1210cf24d184b0de48f.pdf
2. *Simon N. I., Kešelj V.* (2018), Automatic Term Extraction in Technical Domain using Part-of-Speech and Common-Word Features. Proceedings DocEng'18. Halifax. (accessed Februar 24, 2019) URL: https://dl.acm.org/citation.cfm?id=3229100

Kanič Ivan
University of Ljubljana, School of Economics and Business (Slovenia)
E-mail: ivan.kanic@gmail.com

N. B. Krizhanovskaya, A. A. Krizhanovsky

SEMI-AUTOMATIC METHODS FOR ADDING WORDS TO THE DICTIONARY OF VEPKAR CORPUS BASED ON INFLECTIONAL RULES EXTRACTED FROM WIKTIONARY[1]

Abstract. The article describes a technique for using English Wiktionary inflection tables for generating word forms for Veps verbs and nominals in the Open Corpus of Veps and Karelian languages (http://dictorpus.krc.karelia.ru/). The information concerning Karelian and Veps Wiktionary entries with inflection tables is given. The operating principle of the Wiktionary static and dynamic templates is explained with the use of the *jogi* (river) dictionary entry as an example. The method of constructing the inflection table in the dictionary of the VepKar corpus according to the data of the dynamic template of the English Wiktionary is presented.
Keywords. Veps language, Karelian language, corpus, dictionary, Wiktionary.

1. Introduction

The morphological tagging process is one the most laborious works in the corpus linguistics. Large dictionaries with lemmas and word forms are used to perform the morphological tagging.

In the Leipzig Corpora Collection, texts from Internet were crawled and parsed in order to create 400 monolingual dictionaries [Goldhahn et al., 2012]. This is not our case since Veps and Karelian texts are almost absent in the Internet. However the Crúbadán project (Corpus Building for Minority Languages) shows a positive example of automatic search of texts in Internet for under-resourced languages [Scannell, 2007].

The following resources were at our disposal: Wiktionary and traditional dictionaries, native linguists who speak Veps and Karelian, and the programmer who developed the computer system VepKar. The abbreviation VepKar denotes the Open Corpus of Veps and Karelian languages.[2] Researchers at the Karelian Research Centre of RAS have been developing VepKar since 2009 [Zaitseva et al., 2017]. The separate paper [Krizhanovsky et al., 2019] was prepared about Karelian dialects and VepKar corpus in this proceeding.

The article describes a technique for using English Wiktionary inflection tables for generating word forms for Veps verbs and nominals. These generated word forms were added to the VepKar dictionary. The experience of extraction of dynamic templates data from the English Wiktionary is presented in the next section.

[1] The study was supported by the Russian Foundation for Basic Research, grant 18-012-00117.

[2] See http://dictorpus.krc.karelia.ru

2. Dynamic templates for Veps words and static templates for Karelian in English Wiktionary

The paper [Metheniti & Neumann, 2018] leads to the idea to extract word forms from English Wiktionary in order to add new lemmas and word forms to the VepKar dictionary. That paper presents the extraction of information from English Wiktionary in 150 languages, but only Veps and Karelian words are of interest for us. Wiktionary entries can contain inflection tables for nouns and verbs generated by using static or dynamic templates.

There is a big difference in an extraction of information from static templates (case of Metheniti & Neumann) and from dynamic templates (our case). The static template (Table 1) is coded in HTML code. It is need to process the HTML code to remove HTML formatting tags and to get word forms from an inflection table.

A completely different thing is a dynamic template, which includes a script (computer program) with inflectional rules. For each language (Veps language in our case) it was a simple matter to write a language-specific parser for dynamic template, but creating a language-independent parser, or filters for 150+ languages will be a major obstacle. This explains why [Metheniti & Neumann, 2018] extracted word forms only from static templates of Wiktionary.

Below we will compare static and dynamic Wiktionary templates and we will describe our approach, where the English Wiktionary served as a donor for the VepKar dictionary expanding.

2.1. Karelian language and the static template

There are about 650 Karelian lemmas in the English Wiktionary, but only 30 entries contain inflected forms, therefore no new lemmas with word forms were added to the Karelian dictionary of the corpus VepKar. There is only one template {{krl-decl}} for Karelian words in the English Wiktionary. Note that templates in wiki markup are indicated by double curly braces {{…}}. This static pattern {{krl-decl}} used to show declension tables for 30 Karelian nominals[3].

2.2. Veps language and dynamic templates

The situation with Veps words is different in English Wiktionary. There are exactly two dynamic templates: the template {{vep-decl-stems}} for nominals and the template {{vep-conj-stems}} for verbs.

[3] See pages that link to "Template:krl-decl" at w.wiki/3pb

There are 2000 Veps lemmas with about 1000 usages of the inflection-table template {{vep-decl-stems}} in the English Wiktionary. This is a dynamic template that calls a module with Lua programming code. It is worth trying to understand the program in the Lua programming language in order to write the same morphological rules in PHP programming language in VepKar and to get word forms from these 1000 inflection-tables for Veps nouns and adjectives.

2.3. Comparison of static and dynamic Wiktionary templates

The *jogi* [4] (*river*) entry in the English Wiktionary will show the difference between dynamic and static templates. This fact, that the word *jogi* exists in both the Karelian and Veps and that there are two types of templates in the Wiktionary entry *jogi*, makes it suitable as an example.

There are different sets of grammatical cases in Veps and Karelian languages (Fig. 1). The inflection tables generated via these templates (Fig. 1) do not show the difference between the static and dynamic templates… And this is correct, since the difference lies not in the tables themselves, but in the ways they are generated. The inflection table construction methods are presented in the Table 1.

Table 1. **Static and dynamic templates used in the Wiktionary entry *jogi* (*river*)**

N	Karelian language (static template)	Veps language (dynamic template)
1	Wiki page source code (wiki markup)	
2	====Declension==== {{krl-decl|title=jogi|jogi|joven|jogie|joven|joves|-jovespäi|jogih|jovel|jovelpäi|jovele|jovennu|jovekse|–|jovettah|jovenke|joveči|jovet|jogiloin|jogiloi|jovet|…}}	====Inflection==== {{vep-decl-stems|-jog|i|en|ed|id}}
3	Explanation of wiki markup	
4	The source code contains the *static template* {{krl-decl}} with an explicit listing of all 31 word forms, the template generates an inflection table of Karelian word forms.	The source code contains the *dynamic template* {{vep-decl-stems}} and only 5 template arguments (*jog, i, en, ed, id*). This template calls the Lua script for generating a table with 42 Veps word forms.
5	Generated inflection tables (see *Fig. 1a* and *Fig. 1b*)	

[4] See https://en.wiktionary.org/wiki/jogi

Declension of jogi		
	singular	plural
nominative	jogi	jovet
genitive	joven	jogiloin
partitive	jogie	jogiloi
accusative	joven	jovet
inessive	joves	jogilois
elative	jovespäi	jogiloispäi
illative	jogih	jogiloih
adessive	jovel	jogiloil
ablative	jovelpäi	jogiloilpäi
allative	jovele	jogiloile
essive	jovennu	jogiloinnu
translative	jovekse	jogiloikse
instructive	–	jogiloin
abessive	jovettah	jogiloittah
comitative	jovenke	jogiloinke
prolative	joveči	jogiloiči

Fig. 1a. **The inflection table of the Karelian noun *jogi* (*river*) generated by the static template {{krl-decl}}**

Inflection of *jogi*		
	singular	plural
nominative	jogi	joged
accusative	jogen	joged
genitive	jogen	jogiden
partitive	joged	jogid
essive-instructive	jogen	jogin
translative	jogeks	jogikš
inessive	joges	jogiš
elative	jogespäi	jogišpäi
illative	?	jogihe
adessive	jogel	jogil
ablative	jogelpäi	jogilpäi
allative	jogele	jogile
abessive	jogeta	jogita
comitative	jogenke	jogidenke
prolative	jogedme	jogidme
approximative I	jogenno	jogidenno
approximative II	jogennoks	jogidennoks
egressive	jogennopäi	jogidennopäi
terminative I	?	jogihesai
terminative II	jogelesai	jogilesai
terminative III	jogessai	—
additive I	?	jogihepäi
additive II	jogelepäi	jogilepäi

Fig. 1b. **The inflection table of the Veps noun *jogi* (*river*) generated by the dynamic template {{vep-decl-stems}}**

By Wiktionary convention[5], tables that show the forms of nouns are placed in a "Declension" section, while tables that show the forms of verbs are placed in a "Conjugation" section. Fig. 1a corresponds to the convention, but fig. 1b violates it. Why? "This happens because dictionaries are typically the product of several lexicographers' efforts and is constructed, revised, and updated over many years, inconsistencies… necessarily evolve" [Ide & Véronis, 1994]. English Wiktionary has 1600 active editors now[6].

[5] See Inflection-table templates conventions at w.wiki/3o4

[6] See English Wiktionary statistics at w.wiki/44E

2.4. An example of using the data of a dynamic template to add word forms to the Veps dictionary of the VepKar corpus

The template {{vep-decl-stems}}[7] calls the Wiktionary module "vep-nominals"[8]. The inflectional rules for the word form generation were extracted from this code in the Lua programming language and were coded in the PHP language in the VepKar system.

The editor manually copies a text with template name and parameters from the source code of Wiktionary entry (see Table 1, line 2). Then the editor inserts this text into the VepKar corpus to generate word forms (Fig. 2).

The string, which calls the dynamic template {{vep-decl-stems}} with 5 parameters (the stem *jog* and the four endings *i, en, ed, id*), is taken from the English Wiktionary in edit mode (dot-and-dash frame in the bottom

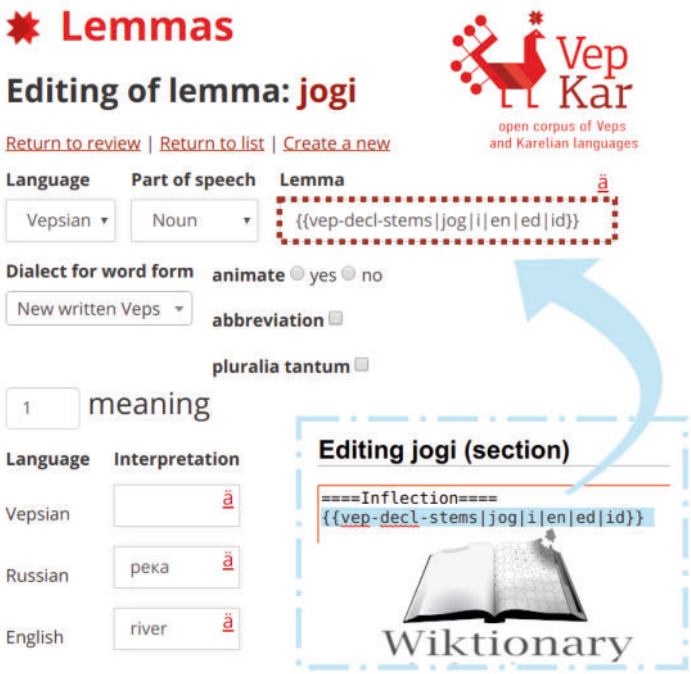


Fig. 2. **The method of constructing the inflection table in the dictionary of the VepKar corpus according to the dynamic template {{vep-decl-stems}} data of the English Wiktionary using the Veps noun *jogi* as an example**

[7] See https://en.wiktionary.org/wiki/Template:vep-decl-stems

[8] See https://en.wiktionary.org/wiki/Module:vep-nominals

of Fig. 2). This string is copied to the *Lemma* field in the VepKar website (dotted frame at the top of Fig. 2) to generate the inflection table of the Veps noun *jogi* (*river*). The generated inflection table in the VepKar corpus for the Veps noun *jogi* is available at: http://dictorpus.krc.karelia.ru/en/dict/lemma/858.

The rules for the word form generation in VepKar made it possible to add 42 word forms to the nominals (based on rules in the module "vep-nominals" [7] in Wiktionary), and 46 verb word forms at once (the template "vep-conj-stems"[9] and the module "vep-verbs"[10]). Thanks to this improvement, technical workers who do not speak Veps have significantly expanded Veps dictionary.

A text with template name and parameters was copied into the VepKar corpus manually (Fig. 2), in order to provide an additional control and verification during the creation of word forms. Since some lemmas in the system already existed, some lemmas had word forms without grammatical information. Thus, it was necessary to check and remove duplicates.

3. Discussion and conclusion

Wiktionary inflection tables were used to expand the Veps dictionary of the VepKar corpus. There are a number of reasons for choosing this approach.

- At the first stage, this approach does not require the participation of linguists.
- Rules for constructing inflection tables have already been developed by Wiktionary editors in Lua programming language. It was necessary to adapt these rules to our VepKar corpus system.

After programming inflectional rules in VepKar, the following procedure was applied.

1) Arguments of Wiktionary templates were manually copied into the VepKar dictionary editor, then the VepKar system generated about 40 word forms for each lemma.
2) Then, these rules, initially encoded in a Wiktionary dynamic template, were presented in natural language in the form of a table for discussion with linguists. These rules have been improved and corrected by our linguists.

[9] See https://en.wiktionary.org/wiki/Template:vep-conj-stems

[10] See https://en.wiktionary.org/wiki/Module:vep-verbs

This table with the rules was a push (and an example) for speeding up work on other Karelian dialects, since it is difficult for linguists to produce a formalized morphological model on their own. A computer program that generates word forms according to the rules is convenient for linguists, since the linguist sees the result and can correct the rule or create a new rule.

Generating word form rules for those grammatical categories that are absent in Wiktionary are going to be refined by Veps linguists in collaboration with programmers in the future. This applies, for example, the illative case for the nominals and all Veps verbs analytical forms.

References

1. *Goldhahn D., Eckart T., Quasthoff U.* (2012), Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages // LREC, Istanbul, Turkey, Vol. 29, pp. 759–765.
2. Ide N., Véronis, J. (1994). Machine Readable Dictionaries: What have we learned, where do we go // Future of Lexical Research, Beijing, China, pp. 137–146.
3. *Krizhanovsky A. A., Krizhanovskaya N. B., Novak I. P.* (2019), Predstavleniye dialektov v Otkrytom korpuse vepsskogo i karel'skogo yazykov (VepKar) [Presentation of dialects in the Open corpus of Veps and Karelian languages (VepKar)] // International scientific conference "Corpus linguistics". Saint Petersburg, 2019.
4. *Metheniti E., Neumann G.* (2018), Wikinflection: massive semi-supervised generation of Multilingual inflectional corpus from Wiktionary // Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway, (155), pp. 147–161.
5. *Scannell K. P.* (2007), The Crúbadán Project: Corpus building for under-resourced languages // Building and Exploring Web Corpora: Web as Corpus Workshop, Vol. 4, pp. 5–15.
6. *Zaitseva N. G., Krizhanovsky A. A., Krizhanovskaya N. B., Pellinen N. A., Rodionova A. P.* (2017), Otkrytyy korpus vepsskogo i karel'skogo yazykov (VepKar): predvaritel'nyy otbor materialov i slovarnaya chast' sistemy [Open corpus of Veps and Karelian languages (VepKar): preliminary selection of materials and dictionary of the system] // International scientific conference "Corpus linguistics". Saint Petersburg, pp. 172–177.

Krizhanovskaya Natalia
IInstitute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences (Russia)
E-mail: nataly.krizhanovsky@gmail.com

Krizhanovsky Andrew
IInstitute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences (Russia)
E-mail: andrew.krizhanovsky@gmail.com

М. Н Михайлов, Ю. В. Соума
M. Mikhailov, J. Souma

СОЗДАНИЕ ПАРАЛЛЕЛЬНОГО КОРПУСА МЕЖГОСУДАРСТВЕННЫХ ДОГОВОРОВ: ТЕХНИЧЕСКИЕ И МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ

COMPILING A PARALLEL CORPUS OF STATE TREATIES: TECHNICAL AND METHODOLOGICAL ISSUES

Аннотация. В статье рассматриваются особенности составления параллельного корпуса специальных текстов на примере корпуса «Parallel Electronic corpus of State Treaties» (PEST). Демонстрируются основные проблемы составления корпусов подобного типа и способы их решения. На основе предварительного анализа русско-финского материала авторы выделяют несколько возможных направлений дальнейшего исследования языковых и жанровых особенностей международных договоров.
Ключевые слова. Параллельный корпус, язык для специальных целей, язык международных договоров, русский язык, финский язык.

Abstract. The article presents the challenges of compiling a parallel corpus of specialist texts, the Parallel Electronic Corpus of State Treaties (PEST) as an example. The solutions of the typical problems of compiling text corpora of this kind are shown. The authors perform a pilot study of Russian-Finnish data and define possible directions for research of language and style of international treaties.
Keywords. Parallel corpora, language for special purposes, language of state treaties, Russian language, Finnish language.

1. Введение

Во многих видах практической и исследовательской деятельности требуются параллельные тексты, то есть оригинальные тексты и их переводы на другие языки (например, художественные произведения, газетные статьи, судебные материалы) или версии одних и тех же документов на разных языках (международные конвенции, технические инструкции). Параллельные тексты нужны для создания и тестирования систем автоматизированного перевода, в качестве «памяти переводов», как источник данных для составления словарей. Несмотря на быстрое развитие одноязычных языковых ресурсов, по-прежнему наблюдается нехватка параллельных текстов.

Двусторонние договоры между государствами — один из текстовых жанров, которые по определению являются дву- или многоязычными. Языковые версии документов должны быть максимально близки друг к другу, а их язык и стиль — достаточно хорошими. Меж-

государственные договоры являются публичными, открытыми для всех документами. Они могут представлять интерес при составлении параллельных корпусов текстов языка для специальных целей. Нам известно о корпусах текстов многоязычных документов международных и европейских организаций: напр., директивы Еврокомиссии (JRC-Acquis) и декларации ООН (MultiUN). Недостатком этих проектов является бедность метаданных и ошибки выравнивания (которые, впрочем, неизбежны при работе с таким большим количеством документов). Двусторонние договоры являются более трудным объектом для составления корпуса. Они не всегда хранятся в виде двуязычных документов, а старые договоры могут не существовать в электронной форме. В настоящей статье будет описан процесс работы над параллельным корпусом двусторонних межгосударственных договоров.

Прежде чем переходить к описанию собственно корпуса, необходимо перечислить некоторые особенности жанра договора, которые необходимо принимать во внимание при выполнении исследовательской работы и при использовании текстов для практических целей.

Согласно Венской конвенции о праве международных договоров (1969), в заключительных положениях договоров указывается, какие языковые версии имеют юридическую силу и к какой из них следует обращаться в случае возникновения разногласий. Эти языковые версии становятся «аутентичными текстами», то есть имеют статус «оригинала», самостоятельного текста. В то же время, все эти тексты в некоторой степени являются переводами, поскольку в процессе подготовки документа, как правило, происходит многократный перевод проекта договора с редактированием, доработками и обратным переводом. Таким образом, тексты двусторонних межгосударственных договоров являются одновременно и оригинальными текстами, и переводами. Эти тексты — плод коллективной работы больших групп специалистов: дипломатов, юристов, экспертов, переводчиков, редакторов. Такие тексты являются и результатом всевозможных компромиссов. Этим и интересен жанр международного договора как юридическое, культурное и языковое явление.

2. Parallel Electronic corpus of State Treaties (PEST) — параллельный корпус межгосударственных договоров

В Университете Тампере с 2015 г. ведется работа над параллельным корпусом международных договоров Parallel Electronic corpus of State

Treaties (PEST). Корпус составляется из полных текстов договоров за исключением приложений. На данном этапе завершен подкорпус договоров между Россией и Финляндией, заключенных с момента получения Финляндией независимости в 1917 г. по сегодняшний день. В дальнейшем планируется собрать договоры между Финляндией и Швецией, а также между Россией и Швецией. На следующем этапе в корпус будут добавлены многосторонние международные договоры (конвенции) с участием России, Финляндии и Швеции. Планируется также включить собранные русско-финские договоры в Корпус параллельных текстов Национального корпуса русского языка (НКРЯ).

На данный момент русско-финский подкорпус содержит 228 пар документов, это все соглашения, заключенные между двумя государствами, как действующие, так и утратившие силу, ведь юридическая сила документа не имеет значения для анализа языка. Объем русской части подкорпуса — 300 000 словоупотреблений, финской части — 250 000. Тексты выровнены на уровне предложений, лемматизированы и содержат морфологическую и синтаксическую разметку. По каждому тексту сохраняются метаданные (название документа, год подписания, тема и т. д.). Для работы с корпусом используется разработанный М. Н. Михайловым и Ю. Хярме пакет онлайн программ TextHammer, предназначенный для работы с параллельными корпусами текстов. Программы поддерживают работу с выровненными аннотированными текстами и позволяют получать конкордансы, частотные списки, списки коллокаций и т. п.

Подкорпус можно исследовать как единое целое или сравнивать разные его части. Группирование текстов подкорпуса может происходить по разным основаниям. Например, русско-финский подкорпус делится на три раздела по хронологическому принципу:

A (1917–1944) — от получения независимости от России до окончания «Войны-продолжения» и заключения Соглашения о перемирии между союзниками и Финляндией в 1944 г.;

B (1945–1991) — от окончания Второй мировой войны до распада Советского Союза;

C (1992–наши дни) — постсоветский период.

Статистика по разделам приводится в табл. 1. Разделы A, B и C имеют разные объемы как по количеству текстов, так и по количеству словоупотреблений, то есть нарушен классический принцип пропорциональности и сбалансированности. Раздел B существенно больше по объему, чем два других, поскольку второй период охватывает про-

Таблица 1. **Состав русско-финского подкорпуса PEST**

Раздел	Период	Количество пар текстов	Кол-во с/у русский язык	Кол-во с/у финский язык
A	1918–1944	46	81 246	67 511
B	1945–1991	128	141 751	115 190
C	1992–2016	54	81 586	906
Всего		228	304 583	247 607

межуток времени почти в 50 лет, и в течение этого периода отношения между странами были заметно более оживленными, чем в период A. Последствия каждого из важнейших политических событий немедленно отражались в заключаемых между странами договорах: увеличивалось или, наоборот, уменьшалось, их количество, которые к тому же они могли значительно отличаться от договоров предыдущих периодов. Тематика договоров сильно связана с требованиями текущего момента, например, мирные договоры есть только в разделе A, а договоры о торговле можно найти во всех разделах.

Параллельный корпус является принципиально другим типом данных, нежели одноязычный корпус. Параллельный корпус всегда более ограничен в объеме, поэтому в случае с корпусом PEST решить проблему асимметрии невозможно: чтобы добиться сбалансированности русско-финского подкорпуса, пришлось бы значительно сократить раздел B. Это привело бы к пропускам в важных тематических разделах и затруднило бы сопоставление данных. Более того, некоторые темы представлены не во всех разделах. Русско-финские договоры составляют только одну из четырех частей корпуса PEST, и попытки сбалансировать этот подкорпус привели бы к другим проблемам на уровне всего корпуса. Получение полностью сбалансированного и пропорционального корпуса текстов в данном случае крайне затруднительно, и даже в случае успеха чревато сильным обеднением материала и потерей ценных данных, которые могут быть полезными для части исследователей.

3. Использование корпуса

При наличии параллельного корпуса текстов, состоящего из большого количества выровненных документов, появляется возможность

исследовать квантитативными методами степень и характер участия сторон в составлении международных договоров, выявлять языковые особенности, не заметные «невооруженным глазом». Исследование международных договоров на материале параллельного корпуса может дать ответ на вопрос, насколько стандартным является язык международных договоров, и можно ли за казенными фразами разглядеть что-то еще? Является ли жанр международных договоров единообразным, и если нет, то что влияет на него больше: время заключения договора, тематика, страна, с которой заключается договор или политическая ситуация в мире и отношения между двумя странами?

Корпус дает возможность анализировать тексты на одном из языков или сравнивать языковые версии договоров между собой как в синхроническом, так и в диахроническом аспектах. При этом добавление каждой новой пары языков повышает объективность и универсальность полученных данных.

Одно из направлений исследования языка договоров — анализ частотных списков, которые позволяют узнать, какие слова чаще всего повторяются в текстах, и таким образом судить о тематике, стилистических пристрастиях автора, степени экспрессивности и т. п. [см., напр., Probirskaja 2009; Mikhailov, Santalahti 2017].

Параллельный корпус дает возможность исследовать, насколько совпадают или расходятся разные языковые версии договоров. Так, предварительное исследование выражения категории отрицания в финско-русских международных договорах [Souma et al 2017] показало, что нередко имеет место описание ситуаций с противоположных точек зрения, например, «запрещено X» ↔ «разрешено не X».

Другое направление исследования — интерференция языков и иные особенности, связанные с переводческой деятельностью, которые можно выявить также путем анализа различных морфологических и синтаксических особенностей текстов.

Анализ языка международных договоров может многое рассказать об отношениях между сторонами, о том, как шла работа над документами, об особенностях жанра международного договора как такового. Корпус текстов может быть интересен не только для лингвистов и переводчиков, но и для историков и специалистов в области международных отношений.

4. Расширение корпуса

У корпуса есть существенная проблема — маленький объем. Договоров между государствами не так много, например, между Россией и Финляндией было заключено немногим больше двухсот соглашений, а ведь это соседние государства. В результате получаем массивы данных довольно скромных размеров, что затрудняет выполнение исследований, требующих большого количества данных. Расширение корпуса путем включения новых пар государств позволяет смягчить проблему лишь отчасти, например, добавив договоры между Россией и Германией и между Финляндией и Германией, мы увеличим количество текстов на русском и финском языках, но при этом количество параллельных русско-финских текстов не изменится и останется незначительным.

Чтобы получить большее количество текстов для каждой языковой пары, можно собирать договоры более низкого уровня: между административными образованиями, городами, общественными организациями, крупными государственными компаниями и т. п. Такие документы не только позволят увеличить объем данных, но и будут давать и более разноплановую картину жанра договора. Работа по расширению корпуса уже начата и финансируется консорциумом FinCLARIN.

Источники

Исследовательский материал

PEST, параллельный корпус межгосударственных договоров. Университет Тампере, 2019. URL: puolukka.uta.fi/texthammer (для получения доступа обращаться к М. Н. Михайлову, mikhail.mikhailov@tuni.fi).

References

1. Vienna Convention on the Law of Treaties (1969), URL: http://legal.un.org/ilc/texts/instruments/english/conventions/1_1_1969.pdf
2. *Souma Ju. V., Kudashev I. S., Mikhailov M. N.* (*2017*), Negation in Russian and Finnish versions of the state treaties between Russia and Finland: a corpus-based research. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". Moscow: RGGU, pp. 393–402.
3. *Probirskaja, Svetlana* (2009), Rajankäyntiä: Suomen ja Venäjän kahdenväliset valtiosopimukset käännöstieteellisen avainsana-analyysin valossa. Väitöskirja. Tampere University Press.
4. *Mikhailov Mikhail, Santalahti Miia, Suoma Julia* (2019), PEST: A parallel electronic corpus of state treaties. В сб: Irene Doval, María Teresa Sánchez Nieto (ред.) Paral-

lel Corpora for Contrastive and Translation Studies. New resources and applications: John Benjamins, 183–195. (Studies in Corpus Linguistics 90).
5. *Mikhailov Mikhail, Santalahti Miia* (2017), Mistä kertoo valtiosopimusten kieli? Tapaustutkimus interferenssistä Suomen ja Venäjän välisissä valtiosopimuksissa. *MikaEL* 10, 73–87.

Михайлов Михаил Николаевич
Университет Тампере (Финляндия)
Mikhailov Mikhail
Tampere University (Finland)
E-mail: mikhail.mikhailov@tuni.fi

Соума Юлия Владимировна
Университет Тампере (Финляндия)
Souma Julia
Tampere University (Finland)
E-mail: julia.souma@tuni.fi

А. А. Новикова
A. Novikova

ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТА SKETCH ENGINE ДЛЯ ИЗВЛЕЧЕНИЯ ТЕРМИНОЛОГИИ

THE USE OF SKETCH ENGINE TOOL FOR TERM EXTRACTION

Аннотация. Извлечение терминологии — одна из актуальных задач корпусной лингвистики. Описание терминосистемы предметной области «Водоснабжение и водоотведение» представляется важным для решения задач научно-технического перевода. В статье описан опыт применения инструмента Sketch Engine для создания тестового корпуса текстов и извлечения терминологии для предметной области «Водоснабжение и водоотведение» на материале английского, немецкого и русского языков.
Ключевые слова. Корпусная лингвистика, корпус текстов, терминология, терминосистема, термины, извлечение терминологии.

Abstract. One of the main problems of corpus linguistics is term extraction. It is important to describe the term system of subject area "Water supply and water sewage" for solving scientific and technical translation problems. The paper describes the testing of the Sketch Engine tool for creation comparable text corpus and term extraction for the subject area "Water supply and water sewage" for English, German and Russian.
Keywords. Corpus linguistics, text corpora, terminology, term system, terms, term extraction.

1. Введение

Извлечение терминологии является одной из актуальных задач корпусной лингвистики. В условиях постоянного научно-технического прогресса растёт роль межкультурной коммуникации и на первый план выходит решение задач точного перевода научно-технических терминов и терминологии. К сожалению, успехи нейронного перевода не решают проблему точного и правильного перевода терминологии. Корпус текстов является проверенным средством, применяемым в терминологических исследованиях, управлении операциями с двуязычной и многоязычной терминологией и составлении на их основе терминологических словарей разного типа [Khurshid, Rogers 1992]. Наибольшее распространение для решения терминологических задач получили параллельные и сопоставимые корпусы текстов — массивы текстов одной предметной области на разных языках. Для параллельных корпусов характерно наличие текстов, являющихся переводами друг друга. В случае создания/использования сопоставимого корпуса, тексты необязательно должны быть переводами друг друга, достаточ-

но их принадлежности к одной предметной области. Данные, полученные с помощью корпусов разного типа, могут быть также использованы для описания терминосистемы определенной предметной области.

Исследование терминологии предметной области «Водоснабжение и водоотведение» и описание её терминосистемы для нескольких языков (английский, немецкий, русский) обусловлено экстралингвистическими и лингвистическими причинами. На данный момент отсутствуют комплексные описания терминосистем данной предметной области, что актуально для решения задач научно-технического перевода и составления многоязычных словарей, особенно в условиях постоянного взаимодействия с зарубежными инвесторами и обмена опытом. Представляется актуальным создание сопоставимого корпуса специальных текстов предметной области «Водоснабжение и водоотведение» для дальнейшего описания и сравнения терминосистем данной предметной области на материале английского, немецкого и русского языков.

2. Инструментарий и эксперимент

Для формирования тестового сопоставимого корпуса специальных текстов предметной области «Водоснабжение и водоотведение» используется инструмент Sketch Engine. Минимальный объём такого корпуса для работы терминолога по рекомендации ряда разработчиков системы Sketch Engine — 100 000 словоупотреблений.

С помощью функции Term Extractor получены списки ключевых слов и терминов-кандидатов предметной области «Водоснабжение и водоотведение» для трёх языков (английский, немецкий, русский). Для проверки полученных терминов-кандидатов выполнена формальная оценка по «эталонному списку» [Браславский, Соколов 2008]. В качестве «эталонного списка» для русского языка использовались ГОСТ по водоснабжению [ГОСТ 25151-82 1982] и Словарь-справочник терминов нормативно-технической документации [Academic.ru 2015]. Для английского языка — Англо-русский словарь по гидротехнике [Владимиров и др. 1983], для немецкого — Немецко-русский словарь по водному хозяйству [Krohmer et al. 2010]. Использование указанных словарей и ГОСТа в качестве «эталонного списка» для формальной оценки представляется целесообразным, потому что в них содержатся слова, относящиеся к профессиональной лексике. Отметим также, что профессиональный язык постоянно обогащается новыми терми-

нами и за последние несколько лет терминосистема предметной области «Водоснабжение и водоотведение» изменилась, добавились новые термины, не указанные в словарях и стандартах, поэтому представляется интересным корпусное исследование терминологии указанной предметной области. Эталонный список необходим для формальной оценки результатов, получаемых с помощью автоматизированных процедур.

Созданный с помощью Sketch Engine корпус содержит тексты из сети Интернет (корпус построен с помощью инструмента WebBootCat, который находит подходящие тексты в сети по ключевым словам, используя поисковую систему Bing). Объём полученного корпуса — 132 513 словоупотреблений. В состав корпуса входят три подкорпуса: для английского языка — 70 382 словоупотребления, для немецкого языка — 26 567, для русского — 35 564. Выборка терминов-кандидатов из корпуса составляет около 1000 лексических единиц для каждого языка соответственно; для тестового исследования решено ограничиться лексическими единицами с наибольшей частотой. Пример выборки терминов-кандидатов для английского, русского и немецкого языков с использованием инструмента Sketch Engine представлен в Таблице 1.

Анализ полученных результатов для английского языка показал совпадение терминов-кандидатов с «эталонным списком» на 65 %, при этом из 35 % не совпавших со списком терминов-кандидатов 10 % полученных терминов-кандидатов относятся к другим предметным областям (напр., *менеджмент, продажи, пожарная безопасность*). Что касается немецкого языка, то совпадение с «эталонным списком» составило только 16 %. Анализ оставшихся 84 % терминов-кандидатов показал, что они, как правило, не относятся к какой-либо предметной области. Эти лексические единицы являются названиями компаний и организаций (напр., *Ingenieur AG*), также в немецком корпусе присутствуют словосочетания, относящиеся к историческим текстам (напр., *römische Zeit, griechische Zeit* и *др.*). Совсем другая ситуация с русским корпусом. С «эталонным списком» совпали почти все термины-кандидаты (совпадение 77 %). Такие термины-кандидаты как *горячий водопровод, наружный водопровод в* списке отсутствуют, однако так или иначе относятся к термину *водоснабжение*.

Также с помощью инструмента Sketch Engine можно выделить лексические единицы, семантически близкие к тому или иному термину [Ковязина 2016]. Для этого в системе Sketch Engine используется

Таблица 1. **Пример выборки терминов-кандидатов для трёх языков с использованием инструмента Sketch Engine**

Английский				Немецкий				Русский			
Term	Score	Freq	R fr	Term	Score	Freq	R fr	Term	Score	Freq	R fr
water service	868.980	84	36	römische Zeit	489.490	16	0	внутренний водопровод	2206.850	106	153
fire flow	667.130	56	0	öffentlicher Brunnen	397.900	13	0	расход воды	1977.940	95	2723
water service provider	486.750	41	1	Ingenieur Ag	367.370	12	0	система водоснабжения	1291.210	62	3531
water supply	439.720	149	721	Müller Ingenieur	367.370	12	0	подача воды	916.640	44	2488
planning study	422.460	36	4	öffentliche Anlage	336.840	11	0	горячая вода	895.830	43	18007
fire service	413.710	42	50	Ersatz Wasserleitung	245.240	8	0	горячее водоснабжение	750.160	36	4503
recycled water	356.140	33	25	griechische Zeit	245.240	8	0	внутренняя канализация	729.350	35	199
demand management	344.760	34	42	Dielsdorf Müller	214.710	7	0	водопроводная сеть	708.540	34	578
backflow prevention	339.520	30	13	Von 200mm	184.180	6	0	пожарный кран	646.110	31	196
planning report	331.410	28	2	Anlage der Wasserversorgung	184.180	6	0	противопожарный водопровод	604.490	29	221

функция построения тезауруса. В рамках тестового исследования эта функция была применена только к термину *водопровод*. Результаты представлены в Таблице 2, где указана частота встречаемости слова (Freq) и статистическая мера (Score), показывающая семантическую близость полученных слов к ключевому слову. Полученный тезаурус в целом является точным, что видно из результатов, представленных в Таблице 2. Может вызывать сомнение наличие некоторых слов в автоматически построенном тезаурусе, поэтому, обычно необходима оценка экспертов. В тезаурус для ключевого слова *водопровод*, которое относится к сфере водоснабжения, попали слова *канализация, водоотведение*, которые относятся к сфере водоотведения. Это разные понятия, однако, все они относятся к одной предметной области «Водоснабжение и водоотведение», и, к тому же, взаимосвязаны: без систем водоснабжения невозможно функционирование систем водоотведения.

Затем, с помощью инструмента кластеризации мы получили список лексико-семантических групп для ключевого слова *водопровод*,

Таблица 2. **Результат построения тезауруса для ключевого слова «водопровод»**

Lemma	Score	Freq	Lemma	Score	Freq	Lemma	Score	Freq
канализация	0.442	297	дом	0.108	104	колодец	0.089	110
водоснабжение	0.306	316	использование	0.107	40	стояк	0.088	127
сеть	0.302	328	диаметр	0.105	111	назначение	0.087	26
система	0.279	419	здание	0.105	396	участок	0.083	95
трубопровод	0.200	191	кран	0.104	91	этаж	0.082	78
вода	0.146	680	установка	0.103	139	состав	0.080	25
труба	0.121	372	качество	0.100	72	точка	0.078	20
ввод	0.117	76	подача	0.098	62	место	0.078	55
водоотведения	0.109	48	кровля	0.094	34	расчет	0.078	61
дом	0.108	104	устройство	0.093	102	водоотведение	0.077	58

представленный на Рис. 1. При автоматической кластеризации были выделены следующие группы слов: *1*) *канализация, водоснабжение; 2*) *сеть, система; 3*) *трубопровод, труба, стояк; 4*) *водоотведение, отопление; 4*) *дом, здание; 5*) *использование, устройство, расчет, прокладка; 6*) *установка, бак; 7*) *подача, очистка; 8*) *кровля, этаж, высота, пример; 9*) *назначение, стена, расположение, конструкция; 10*) *состав, число, условие; 11*) *точка, место; 12*) *поверхность, водосток; 13*) *вариант, период.*

Всего в результате кластеризации получено 60 лексем, из которых 58 % относятся к предметной области, а 42 % нет.

Парадигматические связи внутри кластеров имеют разный характер: например, лексемы *дом, здание* — являются синонимами, *кровля, этаж* — согипонимами; лексемы *установка, бак* демонстрируют объектные отношения, а *качество, уровень* — синтагматическую связь. Далее требуется оценка эксперта, однако уже можно отметить, что такие лексемы как *канализация и водоснабжение, водоотведение и отопление* относятся к разным сферам деятельности: *канализация* относится к сфере водоотведения, а *отопление* — к сфере водоснабжения.

Как показывает анализ полученных данных, возникает очень важный вопрос о достаточном объеме и качестве корпуса. Следующая задача — увеличение объема корпуса до 100 000 словоупотреблений для всех трёх языков. Несмотря на неплохое качество полученного «миниполя» для *водопровода,* в нём не хватает специальной лексики. В любом случае, для комплексного описания терминосистемы той или иной предметной области недостаточно использовать только Web-корпусы, поэтому требуется пополнение корпуса текстами из разных источни-

водопровод *(noun)*
BCBO freq = 397 (8,261.53 per million)

Lemma	Score	Freq	Cluster
канализация	0.442	297	водоснабжение [0.306, 316]
сеть	0.302	328	система [0.279, 419]
трубопровод	0.200	191	труба [0.121, 372] стояк [0.088, 127]
вода	0.146	680	
ввод	0.117	76	
водоотведения	0.109	48	водоотведение [0.077, 58] отопление [0.056, 18]
дом	0.108	104	здание [0.105, 396]
использование	0.107	40	устройство [0.093, 102] расчет [0.078, 61] прокладка [0.058, 38]
диаметр	0.105	111	
кран	0.104	91	
установка	0.103	139	бак [0.058, 57]
качество	0.100	72	уровень [0.074, 38]
подача	0.098	62	очистка [0.066, 54]
кровля	0.094	34	этаж [0.082, 78] высота [0.073, 50] пример [0.069, 24]
колодец	0.089	110	
назначение	0.087	26	стена [0.075, 23] расположение [0.069, 20] конструкция [0.062, 30]
участок	0.083	95	
состав	0.080	25	число [0.074, 72] условие [0.063, 27]
точка	0.078	20	место [0.078, 55]
прибор	0.076	86	
подвал	0.074	18	
поверхность	0.074	20	водосток [0.063, 25]
вариант	0.070	14	период [0.068, 23]
источник	0.070	28	

Рис. 1. **Гнездо тезауруса с выделенными кластерами для ключевого слова «водопровод»**

ков (не только из сети Интернет). Однако использование таких корпусов всё-таки даёт представление о терминологии предметной области «Водоснабжение и водоотведение» и позволяет отработать методику автоматизированного выявления тематической лексики.

3. Заключение

С помощью инструмента Sketch Engine можно быстро создавать специальные корпусы текстов (на основе текстов из сети Интернет), получать данные о частотности лексических единиц, формировать списки ключевых слов и терминов-кандидатов, создавать глоссарии той или иной предметной области. Использование Sketch Engine сокращает работу терминолога, а сам инструмент хорошо зарекомендовал себя в работе по извлечению терминологии. Sketch Engine располагает мощными инструментами для дистрибутивно-статистического

анализа (тезаурус, кластеризация и др.) и выявления парадигматических связей между терминами. Однако, как отмечают исследователи, для совершенствования системы Sketch Engine необходимо совершенствование грамматики лексико-синтаксических шаблонов [Захаров 2015], на которой базируются инструменты выявления устойчивых синтаксических сочетаний (word sketches) и построения дистрибутивного тезауруса. С помощью этого инструмента можно получить предварительные данные, поэтому представляется целесообразным ручная проверка терминов-кандидатов на терминологичность, а также привлечение экспертов для установления парадигматических (онтологических) отношений между терминами.

Полученные результаты дают представление о терминологии, описывающей предметную область «Водоснабжение и водоотведение». Отметим, что корпус является тестовым, более того, для формирования корпуса использовались тексты, полученные только с помощью инструментов Web-поиска, встроенных в систему Sketch Engine, поэтому требуется продолжение исследования. В результате проведенного эксперимента разработана методология выявления терминологии предметной области на основе корпусов.

В дальнейшем для получения более точных и полных данных в рамках исследования требуется пополнение корпусов для обеспечения репрезентативности, а также экспертная оценка полученных результатов. Также планируется дальнейшая работа с немецким корпусом с целью повышения терминологичности автоматически получаемых списков терминов-кандидатов. Отдельная задача — сопоставление терминологии предметной области «Водоснабжение и водоотведение» на трёх языках.

Литература

1. *Браславский П. И., Соколов Е. А.* (2008). Сравнение пяти методов извлечения терминов произвольной длины. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008». Москва.
2. *Владимиров В. и др.* (1983). Англо-Русский Словарь по Гидротехнике (ок. 18 тыс. терминов). Русский язык: Москва.
3. *ГОСТ 25151-82.* (1982). Водоснабжение. Термины и определения. Государственный комитет СССР по стандартам. Москва.
4. *Захаров В. П.* (2015). Корпусно-ориентированный подход к построению тезаурусов и онтологий. Структурная и прикладная лингвистика. Вып. 11. СПб.: Изд-во С.-Петерб. ун-та. С. 123–141.

5. *Ковязина М. А.* (2016). Извлечение ключевых терминов на базе корпуса текстов о разработке нефтяных и газовых месторождений. Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. Том 2. Вып. 3. С. 61–69.
6. *Academic.ru*: словари и энциклопедии на «Академике». (2015). Словарь-справочник терминов нормативно-технической документации. URL: https://normative_reference_dictionary.academic.ru/ (дата обращения: 10.03.2019).
7. *Khurshid A., Rogers M.* (1992). Terminology management: a corpus-based approach. URL: http://www.mt-archive.info/90/Aslib-1992-Ahmad.pdf (дата обращения: 10.03.2019).
8. *Kilgarriff A.* (2014). The Sketch Engine: Ten Years On. Lexicography ASIALEX. Vol. 1, pp. 7–36. URL: http://link.springer.com/article/10.1007/s40607-014-0009-9 (10.03.2019).
9. *Krohmer R., Rumjanzev I. S., Nestmann F.* (2010). Russisch-Deutsches Wörterbuch für Wasserwirtschaft. Karlsruhe: KIT Scientific Publishing.

References

1. *Braslavski P. I., Sokolov E. A.* (2008). Sravnenie pyati metodov izvlechenija terminov proizvol'noj dliny. [Comparison of five methods for variable length term extraction]. Komp'juternaja lingvistika i intellectual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2008». [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2008»]. Moscow, 2008.
2. *Vladimirov V. i dr.* (1983). Anglo-Russkiy slovar' po gidrotehnike (ok. 18 tys. terminov). [English-Russian Dictionary on Hydraulic Engineering (about 18 000 entries)]. Russkiy yazyk: Moskva.
3. *GOST 25151-82.* (1982). Vodosnabjenie. Terminy i opredelenija. [Water Supply. Terms and Definitions]. Gosudarstvenniy komitet SSSR po standartam. Moskva.
4. *Zakharov V. P.* (2015). Korpusno-orientirovannji podhod k postrojeniju tezaurusov i ontologij [Corpus-based approach to thesaurus and ontology construction]. Strukturnaja i prikladnaja lingvistika. [Structural and applied linguistics]. Vip. 11. SPb. 2015. P. 123–141.
5. *Kovjazina M. A.* (2016). Izvlechenije kluchevih terminov na baze korpusa tekstov o razrabotke neftyanih i gazovih mestorozdenij. [Key term extraction based on a corpus of oil and gas field development discourse]. Vestnik Tyumenskogo gosudarstvennogo universiteta. Gumanitarnije issledovanija. Humanitates. [Tyumen State University Herald]. Tom 2. Vip. 3. 2016. P. 61–69.
6. *Academic.ru*: slovari i enciklopedii na «Akademike». (2015). Slovar'-spravochnik terminov normativno-tehnicheskoy dokumentacii. URL: https://normative_reference_dictionary.academic.ru/ (10.03.2019).
7. *Khurshid A., Rogers M.* (1992). Terminology management: a corpus-based approach. URL: http://www.mt-archive.info/90/Aslib-1992-Ahmad.pdf (10.03.2019).
8. *Kilgarriff A.* (2014). The Sketch Engine: Ten Years On. Lexicography ASIALEX. Vol. 1, pp. 7–36. URL: http://link.springer.com/article/10.1007/s40607-014-0009-9 (10.03.2019).

9. *Krohmer R., Rumjanzev I. S., Nestmann F.* (2010). Russisch-Deutsches Wörterbuch für Wasserwirtschaft. Karlsruhe: KIT Scientific Publishing.

Новикова Александра Алексеевна
Санкт-Петербургский государственный университет (Россия)
Novikova Alexandra
Saint-Petersburg State University (Russia)
E-mail: alexa.novikova707@gmail.com

С. Ю. Семенова, А. С. Панина
S. Yu. Semenova, A. S. Panina

ОПЫТ ИСПОЛЬЗОВАНИЯ ДАННЫХ НКРЯ ПРИ ОПИСАНИИ ПОЛИСЕМИИ В ПРИКЛАДНОМ СЕМАНТИЧЕСКОМ СЛОВАРЕ[1]

ON USING RNC DATA FOR RESTRICTED REPRESENTATION OF POLYSEMY IN AN NLP-ORIENTED SEMANTIC DICTIONARY

Аннотация. В АОТ-ориентированном семантическом словаре РУСЛАН значения полисемичных слов отражаются в отдельных статьях. Первые версии словаря получены в 2000-х гг. под руководством Н. Н. Леонтьевой; сейчас группой лексикографов с участием авторов ведутся работы по его расширению. В первых и в нынешних версиях выдерживается стратегия экономного представления полисемии (с эмпирическим ограничением не более 5 статей для слова). В первых версиях значения выделялись практически без корпусных технологий, на основе толковых словарей, с опорой на интуицию лексикографа. Модернизация в существенной мере опирается на корпусные данные; одним из критериев отбора «пятерки» представляемых лексем является встречаемость и частотность в первой сотне вхождений слова в основной корпус НКРЯ.
Ключевые слова: прикладной семантический словарь, полисемия, корпусные данные в компьютерной лексикографии.

Abstract. RUSLAN, a formal dictionary of Russian semantics for automated text processing originally created in mid-2000-ies by N. Leontyeva's group, is at present undergoing a major revision including its representation of lexical polysemy. In this work we rely on the Russian National corpus to find gaps and inconsistencies and to add lexical senses. Lexical ambiguity (polysemy and homonymy) in RUSLAN is uniformly represented by separate entries, with the technical limit of no more than 5 senses per word. In the previous version of the dictionary these were selected mainly by the lexicographer's introspection and subjective decision. The revised criterion for the primary 5 senses is that they need to occur within the first hundred search results in NRC.
Keywords: NLP-oriented semantic dictionary, polysemy, lexical ambiguity, corpora data in computational lexicography.

1. Общие сведения о словаре РУСЛАН

Словарь РУСЛАН, о котором идет речь в данной работе — формализованный семантический словарь, созданный на рубеже 1990-х — 2000-х гг. под руководством Н. Н. Леонтьевой [Леонтьева 2006; Леонтьева, Семенова 2002]. В настоящее время группой лексикографов с участием авторов проводятся работы по его обновлению и расши-

[1] Исследование выполняется при поддержке РФФИ: Проект №17-04-00594-ОГН «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика».

рению. Наряду с авторами в этих работах участвуют М. В. Ермаков, С. А. Крылов и Е. Г. Соколова.

При описании полисемии в нем, как в первых версиях, так и сейчас, была принята стратегия экономного представления полисемии, понимаемая как ограничение количества лексико-семантических вариантов: не более пяти значений для одного слова. Это эмпирическое ограничение обусловлено потенциальными сложностями распознавания значений при машинной обработке текста. На момент начала модернизации РУСЛАНа девять десятых представленных в нем слов были не более чем двузначны, и даже четырехзначных единиц набиралось меньше сотни на весь словник, насчитывавший ок. 12 тыс. единиц.

РУСЛАН создавался в 1990-е гг. как ресурс системы информационного анализа официальных документов РФ в Институте США и Канады РАН (тогда словарь имел сокращенное название РОСС — Русский Общий Семантический Словарь, а система анализа имела название ПОЛИТекст [Леонтьева 2006]). Первый словник формировался на основе текстов указов Президента и постановлений Правительства страны. Ориентиром для определения состава словника (и для построения синтаксического парсера) служило, в частности, официальное периодическое издание «Собрание законодательства Российской Федерации», с его лексикой и стилистикой. На основе ряда официальных документов РФ Ж. Г. Аношкиной был построен конкорданс, который определенное время служил для лексикографов справочным ресурсом. В конце 1990-х гг. словарь был перебазирован в НИВЦ МГУ (где получил свое нынешнее название РУСЛАН) и стал развиваться как общелексический. Наследием РОССа стали значительные пласты общественно-политической лексики при лакунах в предметных словах. Установление баланса между отраслевой и обиходной лексикой и стало одной из причин, побудившей нас обратиться к Национальному корпусу русского языка (НКРЯ).

2. Учет корпусных данных при словарном представлении полисемии

При пополнении словника РУСЛАНа новыми единицами используется не сам НКРЯ, а составленный на его основе Частотный словарь современного русского языка [Ляшевская, Шаров 2009]: добавляются слова, частотные характеристики которых в этом словаре больше единицы.

Что же касается добавления новых значений для слов, статьи которых уже имеются в РУСЛАНе, мы обращаемся непосредственно к НКРЯ и ориентируемся на встречаемость и частотность соответствующих лексем в первой сотне вхождений. Первым этапом описания неоднозначности и для новых, и для уже представленных в словаре единиц, впрочем, остается «предкорпусное» рассмотрение с особым вниманием к тезаурусным связям. Затем это умозрительное представление уточняется по толковым словарям; если словари выделяют у слова более пяти значений, их частотность приблизительно оценивается и ранжируется путем просмотра первой сотни вхождений в НКРЯ. Конечно, реальная глубина просмотра корпуса для семантически интересных слов бывает большей, но первая сотня вхождений рассматривается нами в качестве опорной выборки, позволяющей приблизительно оценить актуальность той или иной общелексической единицы.

Одновременно проверяется и другая словарная информация, в частности, валентностная структура слова. В словаре РУСЛАН предусмотрено два типа информации: классифицирующие и контекстные признаки [Леонтьева, Семенова 2002]. К контекстным признакам относятся поля словарной статьи, отражающие синтаксические свойства и лексическую сочетаемость: грамматические характеристики актантов, словосочетания с заглавным словом, контекстные лексические функции. Для снятия неоднозначности при обработке текста важны в первую очередь контекстные признаки, поэтому они должны особенно тщательно отражаться в словаре.

В большинстве случаев корпусные данные подтверждают ту картину полисемии, которая была реализована в словаре. Так, *кружок* в РУСЛАНе был описан и как графический объект, и как совокупность людей, причем второе значение с двумя вариантами реализации валентности на деятельность — корпус иллюстрирует обе этих реализации одновременно в примере, идущем одним из первых: *театральные кружки или кружки игры на гитаре*. Для графического объекта пример *ставится значок* (*например, кружок*) тоже находится на первой странице выдачи. Тут следует особо оговориться, что мы опираемся на сам факт присутствия некоторого значения в первой сотне примеров, хотя бы в виде единственного вхождения, и не требуем определенного количества или процента от этой сотни; модернизация РУСЛАНа не ставит цели полностью превратить его в частотный словарь.

Среди значений, добавляемых на основании данных НКРЯ, велика доля сравнительно новых словоупотреблений, которых не было в словарях, послуживших в свое время источником для РУСЛАНа. В качестве примера можно привести слово *тур — в* нашем словаре оно присутствовало только в значении этапа состязаний или выборов. Данные НКРЯ не просто иллюстрируют значение «поездка», но и показывают, что оно также не едино — *тур* как концертная поездка отличается от развлекательного путешествия по синтаксическим свойствам: у этого значения чаще реализуется валентность субъекта (*обычный тур главной рок-группы мира*), но нет валентности исходной точки (ср. *туроператоров, организующих туры из Европы в Крым*).

Отдельным вопросом при работе с НКРЯ стал выбор подкорпуса. Из общих соображений газетный подкорпус представлялся наиболее близким к исходным целям словаря РУСЛАН, а именно к ориентации на общественно-политические тексты. Фактически же мы убедились, что на содержание словарных статей сильнее повлияли не ограниченные конкретной отраслью словари, которыми пользовались составители. Результаты в газетном подкорпусе не оказались заметно ближе к лексике РУСЛАНа по сравнению с результатами общего поиска. Так, для слова *аппарат* в РУСЛАНе было представлено только административное значение; оно есть в обоих вариантах поиска в НКРЯ, как и значение «прибор, устройство» (в РУСЛАНе оно отсутствовало, прилагательное *аппаратный* и другие родственные единицы были описаны с отсылками на слово *аппаратура*). Общий поиск в дополнение к этим двум значениям дает также значение «совокупность средств», представленное в таких примерах, как *понимание корректного понятийного аппарата* или *использовать свой голосовой аппарат для производства речевых звуков*.

Разумеется, далеко не всегда результаты поиска в газетном подкорпусе отличаются от общего. Так, для слова *либеральный* терминологическое значение в газетном подкорпусе ожидаемо более частотно, чем «нестрогий» (хотя в его окружении попадается нетерминологическая лексика — например, *либеральная тусовка*), но и последнее тоже представлено. Слово *реставрация* имелось в РУСЛАНе только в физическом значении; в НКРЯ и в общем поиске, и в газетном подкорпусе фигурирует также значение политическое, как в примере *реставрация социализма*. Разделять их необходимо, поскольку у них разные синтаксические свойства: при физическом значении возможен агенс в творительном падеже (*после реставрации подлинного экспоната со-*

трудниками Государственного Исторического музея); при политическом объект способен реализоваться определением (с *монархической реставрацией*); правда, в обоих случаях для иллюстрации этих особенностей потребовался глубокий просмотр результатов поиска.

Тем не менее иногда именно газетный подкорпус помогает охватить всю гамму значений. Так, для слова *свидетель* в РУСЛАНе не было отражено значение «участник судебного процесса», только значение «очевидец». В газетном подкорпусе последнее тоже составляет большинство вхождений, зато для юридического термина близко к началу выдачи встречается его специфическая валентность: *свидетель защиты*, которую можно не вспомнить, если описывать это слово методом интроспекции. Один забавный пример был связан с омонимией: именно в газетном подкорпусе при проверке упомянутого выше слова *кружок* в выдачу попал его частичный омоним, *кружка*, благодаря чему мы заметили, что *кружка* в словаре отсутствовала.

3. Корпусные данные как источник для пополнения фразеологических полей словаря

Материал НКРЯ широко привлекается и при работе над другими типами словарных данных. С помощью корпуса выявляются фразеологические единицы, в том числе характерные именно для современного узуса и не указанные в классических словарях. Корпусные примеры служат основным источником текстовых иллюстраций. Эта деятельность подробно освещена в [Семенова 2017]; зона иллюстраций структурирована, и структура направляет отбор корпусных предложений — подбираются предложения, наиболее прототипично иллюстрирующие каждое из описываемых свойств данной единицы.

Корпус помогает нам и в определении устойчивых сочетаний с заглавным словом. Как правило, та же первая сотня вхождений показывает, какие сочетания полезно отразить в словарной статье (в полях ТЕРМ /термины/ и СЛСЧ /словосочетания/). Иногда обнаруживаются неожиданные сочетания, обладающие значительной частотностью. Например, в первой сотне вхождений слова *совещание* в основной корпус несколько раз встретилось сочетание *особое совещание*, обозначающее историческую реалию. Устойчивость его подтверждается и поиском по биграммам. Видимо, относительно высокая частотность обусловлена тем, что в корпусе в значительной мере (а возможно и гипертрофированно) представлена историческая проза и публицистика

определенной тематики. Еще один пример неожиданных сочетаний в первой сотне вхождений — *очистка от сезонности*. Такое сочетание мы не рассматриваем как фразеологизм в силу его принадлежности к профессиональному жаргону экономистов, но оно дает повод ввести абстрактную лексему *очистка2* наряду с обозначением физического действия *очистка1*.

Наконец, модернизация РУСЛАНа связана с корпусной идеологией еще в одном аспекте: сам словарь может в некотором смысле рассматриваться в качестве метатекстового корпуса. Его состав и сочетаемость единиц в нем может служить объектом анализа как лингвистическая модель.

Таким образом, обращение к корпусу является неотъемлемой частью деятельности в области прикладной семантической лексикографии на современном этапе. При всей важности таких предкорпусных этапов, как интроспекция и привлечение материала толковых словарей, корпус выступает как многофункциональный инструмент, определяющий и частности, и окончательный облик словарного описания.

Литература

1. *Леонтьева Н. Н.* (2006), Автоматическое понимание текстов: системы, модели, ресурсы: Учебное пособие. М.: Академия.
2. *Леонтьева Н. Н., Семенова С. Ю.* (2002), Об отражении полисемии в прикладном семантическом словаре. Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара ДИАЛОГ-2002. Т. 2., с. 489–496.
3. *Семенова С. Ю.* (2017) Об использовании данных Национального корпуса русского языка для иллюстрирования статей компьютерного семантического словаря. Труды международной конференции «Корпусная лингвистика — 2017». СПб., с. 321–324.

References

1. *Leont'yeva N. N.* (2006), Avtomaticheskoye ponimanie tekstov: sistemy, modeli, resursy [Automatic text understanding: systems, mogels, resources]. Moscow.
2. *Leont'yeva N. N, Semenova S. Yu.* (2002), Ob otrazhenii polisemii v prikladnom semanticheskom slovare [On ambiguity representation in an applied semantic dictionary]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2002» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2002»]. Moscow, Vol. 2, pp. 489–496.
3. *Semenova S. Yu.* (2017), Ob ispol'zovanii dannyh Natsional'nogo Korpusa Russkogo Jazyka dlja illjustrirovanija statej komp'juternogo semanticheskogo slovarja [On

choice and of RNC text examples in NLP-aimed semantic dictionary]. Trudy mezhdunarodnoj konferentsii «Korpusnaja lingvistika — 2017» [Proceedings of International Conference «Corpora Linguistics — 2017»], pp. 321–324.

Семенова Софья Юльевна
ИНИОН РАН, РГГУ (Россия)
Semenova Sophia Yu.
INION RAS, RSUH (Russia)
E-mail: sonya_sem@mail.ru

Панина Анна Сергеевна
ИВ РАН (Россия)
Panina Anna S.
Institute of oriental studies RAS (Russia)
E-mail: panina-anna@yandex.ru

R. R. Rebechi

'GOD', 'NATION' AND 'FAMILY' IN THE IMPEACHMENT OF A BRAZIL'S PRESIDENT: A CORPUS-BASED APPROACH TO DISCOURSE

Abstract. Brazil's Lower House of Congress voted the impeachment of President Dilma Rousseff on 17 June 2016. Internet users and online newspapers immediately commented on the vocabulary that most legislators used to justify their votes, in most cases associating words related to God, family and nation to the pro-impeachment speeches. By combining corpus linguistics with discourse studies, I investigated the transcripts of the speeches to confirm or not the general public's and the press's impressions. The recurring keywords showed that the lexical choice of pro- and counter-impeachment voters statistically coincides. However, their context of use sometimes differs.
Keywords. corpus-assisted discourse studies, political discourse, media discourse, impeachment, Brazil.

1. Introduction

On 17 April 2016, 511 Brazilian Deputies voted the impeachment of left-wing President Dilma Rousseff, who was accused of violating fiscal laws by using funds from state banks to cover budget shortfalls. Besides voting for, against, or simply abstaining from sending the case to the Senate, Brazil's Lower House of Congress's representatives were allowed to use the microphone for up to ten seconds to justify their votes. And these justifications reverberated in the press and social media in the form of articles and memes, in general with criticisms towards the lexical choice of the voters favorable to the process, with special emphasis to the so-called triad 'God', 'family', and 'nation'.

In light of such attacking tone towards the Deputies' lexical choices during the voting session, our interest arose in carrying out an analysis not supported solely by the frequency of words, with the purpose of confirming — or not — the representation of the Deputies' speeches by social and mass media. Employing a combination of quantitative and qualitative approaches based on computer-assisted discourse studies (CADS), I aim to semi-automatically examine the transcripts of the votes which resulted in Ms. Rousseff's suspension and subsequent impeachment.

2. A corpus-assisted discourse analysis

Even though discourse studies have long made use of corpus linguistics, this association has not resulted in a large volume of research, if compared to the contributions of CL in lexicography, translation, language teaching, and others. Nevertheless, this area can benefit from the methodology underlying

corpus linguistics, as criticisms to discourse analysis include its supposedly weak method of investigating only fragments of texts based on an analyst's preconceived ideas, which is not academically relevant [cf. Cheng 2003]. Combined with the pragmatism that a corpus-based methodology enables, discourse studies can have its inevitable degree of subjectivity lowered [Partington 2003].

2.1. Media discourse

Fairclough [1989] claims that power in the media is constructed from systematizations, that is, from repetition of information in media activities. As the fourth State [cf. Partington 2003), the media may influence important decisions and change a country's fate for the good and for the bad. Fairclough [1995] defends the concept of discourse representation, since, in general, in the publications there is no transparent report of what was said or written: what is observed is a decision making based on an interpretation and subsequent representation of the information that one wishes to transmit. Therefore, he distinguishes the primary discourse, which is the narrative proper, and the secondary discourse or the representation of the speech, which is permeated by interpretation. In spite of not being passive, the audience can be influenced by journalism, which produces new discourses and reformulates existing ones [Baker 2006].

2.2. Political discourse

Language and politics are strongly intertwined. «Language is necessary to any form of social activity, but politics is arguably the one that relies on language more than most to accomplish its goals» [Romagnuolo 2009: 1]. Several linguistic strategies are used by politicians in order to reach out the interlocutor and convince them of the veracity of their statements. Appealing to patriotism, to the cause of the disadvantaged and to the union, besides legitimizing the self's discourse, whereas delegitimizing the other's, are some of them.

3. Methodology

The research data in this paper is drawn from the transcripts provided by the Chamber as Google Spreadsheet. Each Deputy's talk was saved in TXT format in order to be processed by Wordsmith tools [Scott 2018] and the texts were subdivided into three subcorpora, according to the modes of vote — 'yes', 'no' and 'abstain' (Table 1):

Table 1. **The corpus**

	N. of speeches	N. of words (tokens)
Yes	367	19 249
No	137	7 836
Abstain	7	299
Total	511	27 384

Due to the very limited size of the 'abstain' subcorpus, the focus of the research lies on the 'yes' and 'no' subcorpora. I first applied the keyword technique to identify the most salient words in each subcorpus. This was done by comparing them with a 76 million word-corpus of other Brazilian Lower House sessions. This reference corpus is part of the *Corpus Brasileiro* (CB) [Berber Sardinha et al. 2010], a general language reference corpus of Brazilian Portuguese.

Keywords were calculated using the combination of a statistical test of significance (log-likelihood) with an effect-size measure (Log Ratio) [Hardie, *forthcoming*]. Log-likelihood was applied to identify words that were significantly more frequent in our study corpus in relation to the reference corpus, and Log Ratio was used to determine the difference between the frequencies of a given word in the two corpora. The minimum critical value of 6.63 ($p < 0.01$) was applied as a cut-off point for the log likelihood, and a minimum score of 2.0 for the Log Ratio calculation. In order to avoid selecting words restricted to a handful of examples, the analysis focused on words occurring at least five times in the 'no' subcorpus and twelve times in the 'yes' subcorpus, since their different sizes should be taken into account.

According to the criteria established, the quantitative analyses resulted, respectively, in 65 and 101 keywords. Keywords which recur in both lists, as well as those which are restricted to one of them, were manually investigated.

4. Discussion

4.1. YES or NO

Keywords retrieved from subcorpora 'yes' and 'no' corroborated the clear political polarization of the representatives' opinion. Following Chilton's [2004] dichotomy of strategies related to 'legitimization' and 'delegitimization' in political discourse, different forms of positive self-representation

and negative representation of the opponent were observed in both subcorpora. Pro-impeachment Deputies associated their votes for the continuation of the process with positive words, usually indicating celebration and expectations, such as *viva* [hurray], *esperança* [hope], *amor* [love], *futuro* [future], *favor* [favor], *mudança* [change], and *melhor* [better]. The ones contrary to the process, on the other hand, opted for a vocabulary which reveals defense to the then President and the legality of her government — *honrada* [honorable], *honesta* [honest], *legitimidade* [legitimacy], *urnas* [ballot boxes], *Constituição* [Constitution], etc.

Unsurprisingly, keywords in the 'no' subcorpus also demonstrate their more socialist view by praising people's accomplishments and defending assistance programs and minorities: *democracia* [democracy], *liberdade* [liberty], *soberania* (popular) [(popular) sovereignty], (*estado*) *democrático* (*de direito*) [democratic (rights)], (*reforma*) *agrária* [land (reform)], *trabalhadores* [workers], *pobres* [poor people], *companheiros* [companions], *luta* [fight], *juventude* [youth], *mulher* [woman], *classe* (*trabalhadora*) [(working) class]. *Golpistas* [coupists], *covardes* [cowards], *golpe* [coup], *farsa* [farce], *hipocrisia* [hypocrisy], *corruptos* [corrupts], and *ditadura* [dictatorship] are examples of accusations from the counter-impeachment Deputies against their opponents.

4.2. YES and NO

Among the keywords that recur in the 'yes' and 'no' keyword lists there are functional words, such as prepositions, pronouns and adverbs. These words are usually disregarded, whereas lexical, or content words, are privileged. However, in our analysis three categories deserve attention. They are (i) the first-person plural pronoun *nós* [we], (ii) the possessive adjectives *meu* and *minha* [my], and (iii) the combination of the preposition *per* and the definite article plural *os*, resulting in *pelos* [for]. Chilton [2004: 56] observes that, in political discourse "[…] the first person plural (we, us, our) can be used to induce interpreters to conceptualize group identity, coalitions, parties, and the like". Along with the first person singular *eu* [I], and possessive adjectives *meu*/*minha* [my], *nós* identifies the Self, the speaker, who is *here*, close to the interlocutor.

Along with phrases with *nome* [name] and *respeito* [respect], forming *em nome de* [in the name of] and *em respeito a* [in respect to], *pelos* [for] is used to legitimize the vote by associating it with a group, a place or a renowned character, characteristics which were observed in the two modes of votes. *Cometeu* [committed], *crime* (*de responsabilidade*) [breach of fiscal law] and

corrupção [corruption] are controversial words recurrently used by both sides of voters and which demanded an analysis that goes beyond keyness. The investigation of concordance lines of *cometeu* showed that in the 13 occurrences in the 'yes' subcorpus, the Deputies claim that the then President definitely committed the crime of breach of fiscal law. The only two times *não* [no/not] appears in the surroundings of the keyword, it is used with *só* [only] to include other accusations.

The analysis of few surrounding words of the node results even more unfruitful for *corrupção* [corruption]. *Combate* [fight (noun)], *combater* [fight (verb)], *contra* [against] and *fim* [end] collocate with the keyword in both subcorpora. Only a closer look at the co-contexts can distinguish between reciprocal accusations.

4.3. The triad

By comparing keywords of subcorpora 'yes' and 'no', I observed that *Deus* [God] and some words related to family members — *filho*(*s*) [sons; offspring][1] and *família* [*family*] — statistically recur in both subcorpora. *Nação* [nation], on the other hand, is a keyword only in the 'yes' subcorpus.

The language of political discourse is commonly intertwined with religious beliefs [cf. Chilton 2004]. Contrary to what mass media and Internet users published after impeachment voting, *Deus* [God] appears as a keyword in both subcorpora, even though the analysis of concordance lines showed that the contexts of use differ. In the 'yes' subcorpus, *Deus*, with 49 occurrences, has in its surroundings words associated with religious rituals, such as *abençoe* [bless], *Senhor* [Lord], *agradecer* [thank] and *pedir* [ask], resulting in appeals. As for the 'no' subcorpus, of the seven occurrences of the keyword, *Deus* is used (i) to criticize the speeches of those who supported impeachment (4 times), (ii) as an interjection (1 time) and (iii) to invoke divine help (two times). Therefore, I conclude that, in spite of being mentioned with statistically relevant frequency in both subcorpora, *Deus* is not used with the same intention.

By analyzing the political use of language, Chilton [2004: 117] concluded that, together along with fear, anger, a sense of security and loyalty, protectiveness towards the family is the kind of emotion that is stimulated, since it represents the center of social entities, in contrast with the «insiders» and

[1] In Portuguese, when masculine and feminine are together, the plural is usually formed in the masculine. So, for example, *filhos* can refer to daughter(s) and son(s) or to more than one son.

«outsiders» [Chilton 2004: 52]. Justifying the vote on behalf of family members was also identified as a recurrent strategy in the study corpus, being the keywords *filhos* [sons; offspring] and *família* [family] recurrent in both 'yes' and 'no' modalities of vote. A closer look at the co-contexts indicated a recurring tendency of justifying the vote on behalf of the offspring, as *pelos* [for], followed by the possessive adjectives *meus* [my], *seus* [your] and *nossos* [our], collocates with the search word *filhos* in both subcorpora.

The investigation of the keyword *família* demonstrated that, apart from its canonical meaning — a social group consisting of parent(s) and child(ren) –, used in all occurrences of the word in the 'yes' subcorpus, in three out of the eight occurrences of the word in the 'no' subcorpus, *família* is pronounced three times as part of the proper name *Bolsa Família*, a social program created by left-wing President Luiz Inácio Lula da Silva to provide financial aid to poor families in Brazil. A centerpiece of his administration, the program certainly played a center role in the election of Lula's successor, Dilma Rousseff. Therefore, it comes as no surprise that some Deputies opposed to the impeachment appealed to the program to justify their votes.

According to Chilton [2004: 204], appealing to patriotism is common in political discourse. Deputies pertaining to both modalities of vote declared respect to the states they represent. Nevertheless, reference to the country as a whole is frequent only in the 'yes' subcorpus, in which *Brasil* and *nação* [nation] are keywords.

5. Concluding remarks

By combining CL and discourse studies, in this paper I argue that, unlike what the mass media published, and social network users endorsed, the so-called triad *Deus*, *nação* and *família* does not always occur with statistically more significant frequency in pro-impeachment discourses, but in a larger number in these speeches, since the votes favorable to the process were 2.68 times bigger than those opposed to it. In addition, the crude counting of word occurrences is not enough to reach generalizations. The investigation of the surroundings of the keywords played a vital role in showing the real differences and similarities between the speeches.

References

1. *Baker P.* (2006), Using corpora in discourse analysis. London/New York.
2. *Berber Sardinha T., Moreira Filho J. L., Alambert E.* (2010), Corpus Brasileiro [Brazilian Corpus]. Sao Paulo.

3. *Cheng W.* (2013), Corpus-based linguistic approaches to critical discourse analysis. In *Chapelle C.* (*ed.*), The Encyclopedia of Applied Linguistics. London, pp. 1-8.
4. *Chilton P.* (2004), Analysing political discourse: Theory and practice. London and New York. [Abobe E-Reader Format]. Retrieved from http://voidnetwork.gr/wp-content/uploads/2016/10/Analysing-political-discourse-Theory-and-Practice-by-Paul-Chilton.pdf
5. *Fairclough N.* (1989), Language and Power. London.
6. *Fairclough N.* (1995), Critical discourse analysis: The critical study of language. London/New York.
7. *Hardie A.* (forthcoming), A dual sort-and-filter strategy for statistical analysis of collocation, keywords, and lockwords. [An informal introduction can be found at http://cass.lancs.ac.uk/?p=1133. Accessed 13/02/19].
8. *Partington A. S.* (2003), The linguistics of political argument: The spin-doctor and the wolf-pack at the White House. London.
9. *Romagnuolo A.* (2009), Political discourse in translation: a corpus-based perspective on presidential inaugurals. Translation and Interpreting Studies, 4(1), pp. 1–30.
10. *Scott M.* (2018), Wordsmith Tools, version 7. Oxford.

Rozane Rebechi
Universidade Federal do Rio Grande do Sul (Brazil)
E-mail: rozanereb@gmail.com

Линь Цзиньфэн, Д. М. Семёнова, С. Л. Пущин,
Т. Г. Петров, М. Н. Бабарико, С. В. Чебанов

Lin Jinfeng, D. M. Semyonova, S. L. Pushchin,
T. G. Petrov, M. N. Babariko, S. V. Chebanov

РУЧНАЯ РАЗМЕТКА КОРПУСА ДЛЯ ИЗУЧЕНИЯ СТАТИСТИКИ КОНЦЕПТОВ

MANUAL TAGGING OF THE CORPUS FOR STUDYING OF CONCEPT STATISTICS

Аннотация. Изучение статистики концептов предполагает работу с размеченными корпусами. В принципе, такая разметка может быть только ручной разметкой на основе экспертных оценок с привлечением нескольких экспертов. Однако, в ряде случаем такая возможность исключена и разметка делается автором исследования. Экспликация принципов разметки и воспроизводимые количественные закономерности (покрытие 80 % использования концептов 7±2 из них) дают основание считать такую разметку удовлетворительной.
Ключевые слова. Концепт, распределение концептов, квантитативная концептология, ручная разметка текста, соотношение Парето, магическое число Миллера.

Abstract. The study of concept statistics involves working with tagging corpuses. In principle, such a tagging can only be manual tagging based on expert assessments involving several experts. However, in some cases this possibility is excluded and the tagging is made by the author of the study. The explication of the principles of tagging and reproducible quantitative patterns (covering 80 % of the use of concepts 7 ± 2 of them) suggest that such tagging is satisfactory.
Keywords. Concept, concept distribution, quantitative conceptology, be manual text tagging, Pareto relation, Miller magic number.

Очевидным путём исследования концептов является разметка корпуса несколькими экспертами, но иногда привлечение нескольких экспертов невозможно. Далее в качестве экспертов выступали авторы ниже перечисленных исследований в сотрудничестве с С. В. Чебановым как руководителем.

Первый пример такой работы — исследование социальных институтов как типичных концептов в текстах, являющихся эталонными представлениями социальной реальности. Частоты упоминания институтов собраны по «Народным русским сказкам» А. Н. Афанасьева А. П. Чернышовой, кодексу Наполеона Ю. И. Ляпуновой, «Соборянам» Н. С. Лескова М. В. Кирилловой, «Истории одного города» и «Сказкам» М. Е. Салтыкова-Щедрина, «Жизни и необычайным приключениям солдата Ивана Чонкина» и «Москве 2042» В. Н. Войновича М. А. Смирновой, письменным работам школьников А. С. Курочкиной, что было

обобщено в статье С. В. Чебанова [2012]. Во всех случаях работа начиналась с выделения в тексте обозначений социальных институтов, сложность чего в их несформированности в русской культуре. Для их распознания требуется специальное обучение, что исключает использование нескольких экспертов. В результате обнаружена резкая неравночисленность упоминания разных концептов, но вид их распределения не изучался.

Такие распределения получены для концептов — описаний жестов в романе Л. Н. Толстого «Война и мир» [Семёнова, Чебанов 2012]. В ранговой форме распределение концептов разных компонентов жестов резко убывающее, а распределение жестов по героям — обычное ципфоподобное *H*-распределение (как и лексем в русских пословицах — рис. 1).

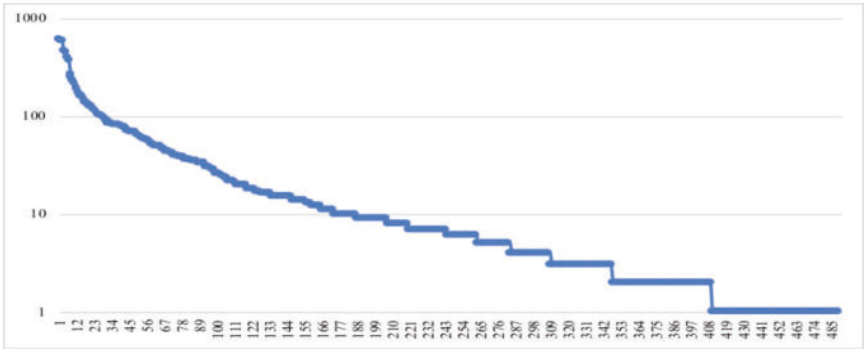


Рис. 1. **Частотно-алфавитное распределение лексем описания составности человека в СД. По оси абсцисс — ранги лексем, по оси ординат — логарифм частот**

Распределение концептов было основным объектом исследования Линь Цзиньфэн [2018] по русской (по собранию В. И. Даля ([Даль, 1862] — далее СД) и китайской (по Собранию китайских пословиц, 1961 [Собрание…, 1961] — далее СКП) пословичным картинам мира.

В этом случае разметка зависит от склонности эксперта к дихотомии, признающей в составе человека тело и дух, или к трихотомии, различающей тело, душу и дух, которую невозможно учесть при отборе экспертов. Поэтому разметка производилась автором, на протяжении двух с половиной лет занимавшимся этой проблемой, после чего он выбрал трихотомию как основу ручной разметки [Линь Цзиньфэн 2018], по данным которой была собрана статистика.

На основе анализа семантики токенов (лексем, слово-сочетаний, описательных конструкций, косвенных наименований и т. д.) и содержащих их пословиц были выделены концепты 1-го уровня, передающие составность человека и вычислены их частоты как суммы частот передающих их токенов. На основании отношения «часть-целое» выделены концепты 2-5-го уровней (для СД и СКП). Их частоты вычислялись как суммы частот, входящих в них концептов более низкого уровня (Табл. 1).

Таблица 1. **Число концептов разных уровней (N) и их число, приходящееся на 80 % покрытия (m)**

Уровень концептов	СД		СКП	
	N	m	N	m
1	49	8	37	12
2	36	8	27	9
3	24	6	21	7
4	18	6	15	5
5	5	2	3	1

Распределения этих частот резко неравночисленные и в ранговой форме резко убывающие (Рис. 2). Характер этих распределений оказывается однотипным для распределений частот концептов 1÷5 уровней русских и китайских пословиц. При этом, чем выше уровень концептов, тем их меньше и тем круче падение их частот. Динамика накопления частот концептов для всех распределений совершенно однотипна (Рис. 3).

При этом 80 % (соотношение Парето) покрытия частот употребления концептов достигается для концептов 1÷4 уровней русских и китайских пословиц за счёт 5–9 концептов (кроме 12) и слабо зависит от числа концептов (15÷49; Табл. 1), что соответствует магическому числу Миллера 7±2 компонентов, оптимальному для оперативной памяти [Миллер 2010].

Такая же картина получена [Бабарико, Чебанов 2015] для концептов чисел в пословицах СД, собрания В. М. Мокиенко с со-авторами ([Мокиенко и др., 2010] — далее СМ) и в СКП [Babariko, Jinfeng, Chebanov 2016]. 80 % покрытие числовых концептов достигается для 7 чисел всех трёх массивов — СД, СМ и СКП.

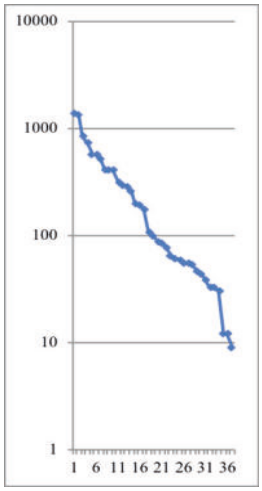


Рис. 2. **Частоты концептов, передающих составность человека, 1-го уровня в китайских пословицах. По оси абсцисс — ранг, по оси ординат — логарифм частот**

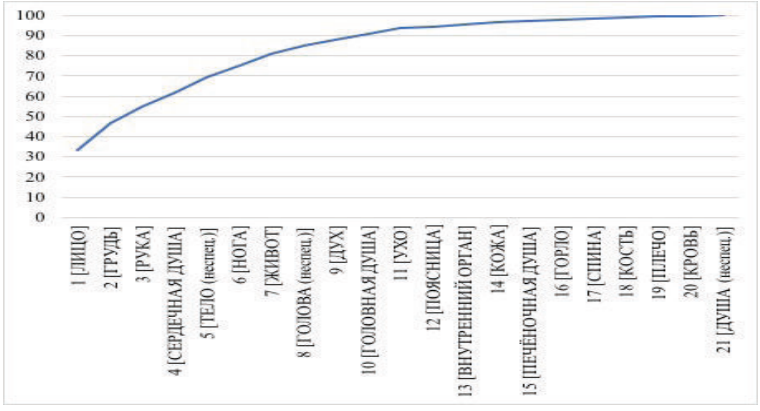


Рис. 3. **Накопленные частоты концептов 3-го уровня (китайских). По оси абсцисс ранги и соответствующие им лексемы, по оси ординат — накопленные частоты (в процентах)**

Другое направление квантитативной концептологии –изучение частот токенов, выражающих концепты. Вариант их изучения наме-

чен в связи с понятием пойнтер-точки [Кудрин 2007: 25–33]. Видовым (спектровым) представлением *H*-распределения является функция $\Omega(x) = W_0/x^{1+\alpha}$, где Ω(x) — коли-чество знакотипов с одинаковым количеством знакоупотребле-ний, x — количество знакоупотреблений каждого знакотипа, α — характеристический показатель распределения, а за W_0 принима-ется численность самого частого знакотипа. Пойнтер-точка $\mathcal{R}$ — точка перегиба, такая что, «Гипербола делится точкой $\mathcal{R}$ на две ветви: слева i = 1, 2, …, $\mathcal{R}$ — неоднородные касты <классы знако-типов, в данном случае лексем, с одинаковой частотой>, где каж-дая каста представлена множеством видов <лексем>; справа i = $\mathcal{R}$ + 1, $\mathcal{R}$ + 2, …, K — однородные <содержание только одну лексе-му> касты … (i соответствует числу особей этого вида)» [Кудрин 2007: 29]. Е. Б. Кудрина показала, что по частоте упоминаний ге-роев в «Мастере и Маргарите» М. А. Булгакова к пойнтер-точке $\boldsymbol{R}$ = 34 примыкают Левий Матвей, Гелла, Н. И. Босой, Варенуха, Римский, Стёпа Лиходеев и Га-Ноцри, которые составляют бул-гаковскую специфику повествования и его отличие от «Фауста» Гёте [Кудрин 2007: 31]. Поэтому по методике С. Л. Пущина [2014: 21–28] была найдена пойнтер-точка распределения токенов, описывающих человека в СД и СКП.

Для СД пойнтер-точка соответствует частоте 186 (ближай-шие лексемы Воля — 190 и Сердце — 175), вблизи пойнтер-точки оказываются *добро, видеться, добрый, воля, сердце, чёрт, грех*, характеризующие [СЕРДЕЧНУЮ ДУШУ], [ГОЛОВНУЮ ДУ-ШУ], [ДУХ], а через *видеть* и [ТЕЛО], что соответствует образу русского человека как живущего душой и духом. Для СКП пойн-тер-точки равна 92 (ближайшие лексемы 听Слышать — 96, 腿Нога от пояса до стопы — 87), вблизи которой 饱 *сыт,* 死 *уме-реть,* 饥 *голод,* 讲 (*книжн.*) *говорить,* 食*еда* (*книжн.*), 腰 *поясни-ца*, передающие потребность в еде как важнейшую потребность [ТЕЛА], отмечая важную для китайской культуры часть [ТЕЛА] [ПОЯСНИЦУ] и включая только две лексемы, обозначающие действия [ГОЛОВНОЙ ДУШИ] — *слышать* и *говорить*. Такая контрастная картина соответствует клише образов представителей двух народов.

Для концепта [ВЗГЛЯД] по материалам Д. М. Семёновой удалось рассмотреть распределение концептов его вариантов (средств передачи концепта [ВЗГЛЯД]) и рассчитать для него пойнтер-точку, которой соответствует концепт [СМОТРЕТЬ (ПУСТО)] с частотой 9 в окружении [ОПУСТИТЬ ГЛАЗА (ПУС-ТО)] и [ВЗГЛЯД (ПУСТО)], что отражает не только значение концепта [ВЗГЛЯД], но и всё содержание текста, выражающего взгляд Л. Н. Толстого на светское общество.

Другой способ количественного анализа выражающих кон-цепт токенов — метод ***RHA*** Т. Г. Петрова. Он заключается в том, что после получения ранговой формулы ***R*** (частотный словарь выражающих концепты токенов), вычисляется информационная энтропия Шеннона (***H*** = $-\Sigma p_i * \ln p_i$ где p_i — нормированная к 1 частота i-ой лексемы — [Петров, Фарафонова 2005: 48]), характе-ризующая равномерность распределения токенов, и анэнтропии (***A*** = $-[(\Sigma \ln p_i)/n] - \ln(n)$, где n — число токенов, представляющих концепт — [там же, с. 61]), введённая для характеристики нерав-номерности доли компонентов (токенов) в распределение. При этом ***H и A*** рассчитываются для полных (что отражает индивиду-альность) или усечённых составов (для их сопоставления).

Были вычислены ***H и A*** для токенов, представляющих кон-цепты [ТЕЛО], [ДУША], [ДУХ] и их совокупности в СД и СКП. Расчёты произведены по полным и усечённым (по 20 самым частым токенам, что равно минимальному числу токенов, выражающих концепт — [ДУХ] *в* СКП) словарям. В итоге получены ожидаемые результаты, вызывающие частные вопросы.

Изменение ***H*** однотипно для полных и усечённых словарей СД и СКП. Для каждого из четырёх сопоставимых комплектов данных ***H*** является минимальной для токенов концепта [ДУХ], следующими являются ***H*** токенов концепта [ДУША], а ещё больше для концепта [ТЕЛО]. ***H*** для токенов, передающих совокупность концептов [ТЕЛО]+[ДУША]+[ДУХ], является максимальным. Для полных и усечённых составов все величины ***H*** для СД сдвинуты в сторону больших величин, по сравнению с СКП, свидетельствуя о большей дифференциации представления о человеке в русских пословицах. При этом основной вклад в ***H*** совокупности концептов [ТЕЛО]+[ДУША]+[ДУХ] вносят токе-ны, передающие концепт [ТЕЛО], причём этот вклад заметно больше для китайских пословиц. Для усечённых словарей ***A*** уменьшается с увеличением словаря, т. е. с ростом *H*, что ожидае-мо. Неожиданно то, что все точки, (исключение — [ДУХ] русских пословиц), лежат на одной прямой. Для полных словарей ***A*** в несколько раз больше, чем для усечённых, и есть тенденция роста ***A*** с увеличением объёма словаря и ростом *H*.

В итоге обнаруживается новый статистический объекта со следующими характеристиками.

— Имеется некоторый текстовой массив, в котором выделен набор однотипных концептов. Тогда

— распределение частот этих концептов в ранговой форме — это резко убывающее распределение (показано для 12 массивов) такое, что:
— Миллеровское число 7±2 концептов покрывает 80 % употреблений концептов из этого набора (выполняется закон Парето 80:20; показано для 5 массивов).
— Частота самых редких концептов измеряется несколькими единицами — первыми десятками (если концепты связаны с тематикой текста) или единицами (если тематика текста иная чем семантика концепта; напр., числа в пословицах), но они не образуют длинного хвоста *hapax legomena* (как в *H*-распределениях токенов), что позволяет квалифицировать их как распределения с толстыми хвостами [Фуфаев 1996; Anderson 2006], которые обрезаны.
— Распределение токенов, выражающих этот набор концептов, аппроксимируется *H*-распределением, вблизи пойнтер-точки ***R*** которого концентрируются токены, наиболее полно представляющие данный набор концептов (показано для 4 массивов, для 3 — впервые).
— Для усечённых словарей токенов существует обратное соотношение ***H*** и ***A*** (показано для 2 массивов).
— Для полных словарей токенов наблюдается прямое отношение ***H*** и ***A***, причём ***A*** полнее отражает особенности текста (показано для 2 массивов).

Итак, на изученных с разной детальностью 12 массивах, рассматриваемых как размеченные корпуса, выявлены однотипные закономерности распределения концептов. Это позволяет утверждать, что ручная семантическая разметка корпуса одним экспертом приемлема для изучения концептов.

Литература

1. *Бабарико М. Н., Чебанов С. В.* (2015), Русская паремиологическая арифмология XIX-XXI веков // Структурная и прикладная лингвистика. Вып. 11. СПб.: СПбГУ, с. 186–219.
2. *Даль В. И.* (1862), Пословицы русского народа. М.: В Университетской типографии, 883 с.
3. *Линь Цзиньфэн.* (2018), Концепты [ТЕЛО], [ДУША], [ДУХ] в русской и китайской языковых картинах мира (антропологическая трихотомия в пословичной картине мира). Дисс. … к. филол. н. Спец. 10.02.21. Т. 1-2. СПб.: СПбГУ.

4. *Миллер Дж. А.* (2010), Магическое число семь плюс минус два. О некоторых пределах нашей способности перераба-тывать информации. URL: http://www.ebbinghaus.ru/wpcontent/uploads/2010/02/Miller_564-580.pdf.
5. *Мокиенко В. М., Никитина Т. Г., Николаева Е. К.* (2010), Большой словарь русских пословиц. М.: ОЛМА Медиа Групп, 1026 с.
6. *Петров Т. Г., Фарафонова О. И.* (2005), Информацион-но-компонентный анализ. Метод RHA: СПб.: СПбГУ, 168 с.
7. *Пущин С. Л.* (2014), О трёх теоремах Б. И. Кудрина // Ценологические исследования. М.: Техника. Вып. 53, с. 11–28.
8. *Семёнова Д. М., Чебанов С. В.* (2012), Ценоз описаний кинесики романа Л. Н. Толстого «Война и мир». Ценологические исследования. М.: Техника. Вып. 46: Специфика ценологических представлений разных школ, с. 181–203.
9. *Фуфаев В. В.* (1996), Основы теории динамики струк-туры техноценозов // Ценологические исследования. Абакан: Центр системных исследований. Вып. 1: Математическое описание ценозов и закономерности технетики, с. 156–193.
10. *Чебанов С. В.* (2012), Полнотекстовые базы данных как инструмент понимания (на материале русской лингвосоциологии) // Понимание и рефлексия в коммуникации, культуре и образовании: материалы Международной научно-практической Интернет-конференции, посвященной 70-летию Факультета иностранных языков и международной коммуникации Тверского государственного университета. 1 октября — 15 де-кабря 2011 г. Тверь: Тверской государственный университет, с. 185–197.
11. *Anderson Ch.* (2006), The Long Tail: Why the Future of Business Is Selling Less of More. N. Y.: Hyperion, 238 pp.
12. *Babariko M., Jinfeng L., Chebanov S.* (2016), Idealized Cognitive Model (ICM) of Numbers in the Chinese (C) and Russian (R) Linguistic World Picture (LWP) as a Basis of Conceptual Map-ping. 3rd International Congress of Numanities (ICoN 2016). Creati-vity, Diversity, Development. Program and abstracts. Kaunas, Interna-tional Semiotics Institute, Kaunas University of Technology. pp. 43–45.
13. 中国谚语资料, 中国文艺研究会资料室主编, 兰州艺术学院文学系55级民间文学小组, 上中下三册,上海文艺出版社, 1961年, 1111页 (Собрание китайских пословиц (1961) / Китай-ская научная библиотека искусств, фольклорная группа факуль-тета литературы Ланьчжоуского института искусств. Шанхай: издательство Шанхайской литературы и искусств. Т. 1–2. 1111 с)

References

1. *Babariko M. N., Chebanov S. V.* (2015), Russian paremiolo-gical arithmetic of XIX-XXI centuries.\ // Structural and applied linguistics. Issue. 11. SPb.: SPbSU, pp. 186–219.
2. *Dal V. I.* (1862), Proverbs of the Russian people. M.: In the University printing house, 883 pp.
3. *Lin Jinfeng.* (2018), Concepts [BODY], [SOUL], [SPIRIT] in Russian and Chinese language worldview (anthropological trichoto-my in the proverbial worldview). Diss. … candidate of philological sciences. Spec. 10.02.21. Vol. 1–2. SPb.: SPbSU.

4. *Miller Dzh. A.* (2010), Magic number seven plus or minus two. Some limits on our ability to process information, URL: http://www.ebbinghaus.ru/wpcontent/uploads/2010/02/Miller_564-580.pdf.
5. *Mokienko V. M., Nikitina T. G., Nikolaeva E. K.* (2010), Large dictionary of Russian Proverbs. M. : OLMA Media Group, 1026 p.
6. *Petrov T. G., Farafonova O. I.* (2005), Information-component analysis. RHA-method: SPb.: SPbSU, 168 p.
7. *Pushchin S. L.* (2014), On three theorems of B. I. Kudrin // Cenological research. M. : Technique. Issue. 53, pp. 11–28.
8. Semenova *D. M., Chebanov S. V.* (2012), Coenosis of descriptions of kinesics of the novel L. N. Tolstoy's «War and peace». Cenological research. M.: Technique. Issue. 46: Specific of cenological views of different schools, pp. 181–203.
9. *Fufayev V. V.* (1996), Fundamentals of the theory of the dynamics of the structure technocenosis // Cenological research. Abakan: Center for system research. Issue 1: Mathematical description of coenoses and laws of technetics, pp. 156–193.
10. *Chebanov S. V.* (2012), Full-text databases as a tool of un-derstanding (by the material of Russian linguosociology) // Understan-ding and reflection in communication, culture and education: materi-als of the International scientific and practical Internet-conference dedicated to the 70th anniversary of the Faculty of foreign languages and international communication of Tver state University. October 1 — December 15 2011. Tver: Tver state University, pp. 185–197.
11. *Anderson Ch.* (2006), The Long Tail: Why the Future of Business Is Selling Less of More. N. Y.: Hyperion, 238 p.
12. *Babariko M., Jinfeng L., Chebanov S.* (2016), Idealized Cognitive Model (ICM) of Numbers in the Chinese (C) and Russian (R) Linguistic World Picture (LWP) as a Basis of Conceptual Map-ping. 3rd International Congress of Numanities (ICoN 2016). Creati-vity, Diversity, Development. Program and abstracts. Kaunas, Interna-tional Semiotics Institute, Kaunas University of Technology, pp. 43–45.
13. Collection of Chinese Proverbs (1961) / Chinese scientific library of arts, folk literature group of Lanzhou Institute of arts. Shanghai: publishing house of Shanghai literature and arts. Vol. 1–2. 1111 p.

Линь Цзиньфэн
Ланчжоуский политехнический университет (Китай)
Lin Jinfeng
Lanzhou University of Technology (China)
***E-mail:* linjinfeng1990@163.com**

Семёнова Дарья Михайловна
ООО «Интеллиджер» (Россия)
Semenova Daria Mikhailovna
LLC "Intelliger" (Russia)
***E-mail:* dasha.glc@gmail.com**

Пущин Сергей Львович
ООО «ТИПОГРАФИЯ КСИ-ПРИНТ» (Россия)
Pushchin Sergey Lvovich
LLC "TIPOGRAFIYA KSI-PRINT" (Russia)
E-mail: **z1q813@mail.ru**

Петров Томас Георгиевич
ООО «Соколов» (Россия)
Petrov Thomas Georgievich
LLC "Sokolov" (Russia)
E-mail: **tomas_petrov@rambler.ru**

Бабарико Максим Николаевич
Министерство экономического развития (Россия)
Babariko Maxim Nikolaevich
Economic development Ministry (Russia)
E-mail: legal.insolvency@gmail.com

Чебанов Сергей Викторович
Кафедра математической лингвистики филологического факультета СПбГУ (Россия)
Chebanov Sergey Viktorovich
Department of mathematical linguistics, faculty of Philology, Saint-Petersburg State University (Russia)
E-mail: **s.chebanov@gmail.com**

В. А. Шульгинов, В. А. Шульгинов

V. A. Shulginov, V. A. Shulginov

КОРПУСНОЕ ИССЛЕДОВАНИЕ АВТОРСКОЙ РЕЦЕПЦИИ В СТРУКТУРЕ ЭЛЕКТРОННОГО ГИПЕРТЕКСТА[1]

CORPUS DRIVEN RESEARCH OF AUTHOR RECEPTION IN STRUCTURE OF HYPERTEXT

Аннотация. В данной статье электронный гипертекст рассматривается как нелинейная коммуникативно-познавательная единица, отвечающая всем критериям текстуальности. Нами разрабатывается корпусный подход к количественному анализу семантики гипертекстового перехода в парах «ссылка/целевой текст», «предложение/целевой текст». В статье описывается принцип сбора и организации база данных гипертекстовых структур, а также алгоритм разметки по морфологическому признаку и степени семантической близости. В ходе исследования полученных данных выявляются механизмы авторской рецепции при организации электронного гипертекста.
Ключевые слова. Электронный гипертекст, пресуппозиция, дистрибутивная семантика, корпус, семантическая близость.

Abstract. The paper presents hypertext as a non-linear communicative-cognitive phenomenon, which has all the signs of textuality. We propose a corpus study of the semantics of hypertext transitions, based on the analysis of semantic proximity pairs link/target text, sentence/target text. The paper presents the principle of parsing and organizing a database of hypertext structures, as well as a markup algorithm for the POS-tagging and the semantic proximity. The article analyzes the features of the author's reception in the organization of hypertext.
Keywords. Hypertext, presupposition, distributive semantics, corpus, semantic proximity.

Введение

Понятие «электронный гипертекст» имеет множество интерпретаций как в гуманитарных, так и технических науках, что возводит его в ранг мифологемы современного научного знания. Мы считаем методологически верным акцентировать внимание на текстовой природе электронного гипертекста и, вслед за Р. К. Потаповой, рассматривать электронный гипертекст как особый тип текста. В нашем понимании электронный гипертекст представляет собой «коммуникативно-познавательную единицу нового типа, которая, с одной стороны, отвечает всем критериям текстуальности (целостность, связность, намеренность, приемлемость, информативность, ситуативность, ин-

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-312-00010

тертекстуальность), с другой — характеризуется сложной структурой и нелинейными связями между фрагментами» [Шульгинов 2016: 233].

При этом гипертекстуальность оказывается имманентным признаком интернет-среды, в которой функционирует электронный гипертекст. Погружая текст в электронную среду, автор эксплицирует заложенную в нем интертекстуальность с помощью гипертекстовых ссылок, которые обеспечивают возможность гипертекстового перехода между фрагментами. Ключевым критерием существования электронного гипертекста становится наличие семантики гипертекстового перехода, который выражается в семантическом взаимодействии источника ссылки и его объекта» [Дедова 2009: 196]. Семантика гипертекстового перехода отражается во взаимодействии номинации ссылки с целевым текстом, которая может быть рассмотрена в двух аспектах:

— от ссылки к целевому тексту: гипертекстовый переход становится способом прояснения референта высказывания, необходимого для адекватного восприятия текста читателем, что особенно актуально при толковании терминологической лексики в электронных энциклопедиях [Клочкова 2015: 53];
— от целевого текста к ссылке: целевой фрагмент гипертекста рассматривается в качестве текста-стимула, а номинация ссылки и контекст ее существования становятся носителями вербализованной реакции автора.

Таким образом, электронный гипертекст представляет собой совокупность трехчастных элементов, являющихся результатом рецепции автора, которые включают в себя исходный текст (текст-носитель ссылки), целевой текст (текст, к которому осуществляется переход при активации ссылки) и ссылку, обеспечивающую связность текстовых фрагментов. Для номинации данных единиц мы предлагаем использовать термин «гипертекстема». Мы используем корпусный подход[2] для исследования семантических связей в структуре гипертекстемы.

1. Принцип построения корпуса гипертекстем

Разработка корпуса гипертекстем включает в себя сбор гипертекстовых единиц, нормализацию текстов, аннотирование по лингвистическим и экстралингвистическим принципам.

[2] Доступ к тестовой версии корпуса можно получить по адресу: http://hypercorpus.ru/

Формирование базы данных происходит в автоматическом режиме с помощью парсера (разработан на языке программирования Python), алгоритм работы которого включает в себя два этапа. На первом этапе парсер обходит заданные электронные ресурсы для индексации всех ссылок на страницах, в результате чего формируется база данных, состоящая из следующих пар: исходная/целевая страница ссылки, а также адреса их доменов. На втором этапе парсер анализирует полученные пары страниц для того, чтобы обнаружить в них полные текстовые фрагменты. При обнаружении нужных тегов все элементы исходного и целевого текста загружаются в базу данных.

В состав основных источников вошли следующие электронные СМИ: «Коммерсант», «Известия», «РБК», «Новая газета», «ТАСС», «Дождь», «Новая газета», «Ведомости», «Интерфакс». В настоящий момент база данных включает в себя 31 тыс. уникальных текстов, входящих в состав 51 тыс. гипертекстем (12 млн. лексем).

На этапе инвентаризации мы выделяем следующие значимые элементы гипертекстемы: целевой текст; исходный текст, из которого извлекается предложение-носитель ссылки и номинация ссылки. Такое членение исходного текста позволяет проанализировать минимальный контекст существования ссылки. Затем предобработку проходит каждый текстовый фрагмент: удаляются стоп-слова, проводится токенизация и лемматизация значимых слов. Кроме того, в составе исходного и целевого текста выделяются ключевые слова, характеризующие их тематическую отнесенность. Мы пришли к выводу, что при работе со стилистически однородными текстами более точные данные дает метрика TF-IDF, которая считается по формуле

$$TFIDF_{x,y} = tf_{x,y} \times \log , \quad (1)$$

где *tf* — частота слова в данном документе, N — общее число документов в корпусе, df_x — количество документов, содержащих слово. Данная мера позволяет выявлять те лексемы, которые оказываются уникальны для конкретного текста относительно общего массива текстов.

Лингвистическая разметка включает полное морфологическое аннотирование (используется инструмент Pymorphy2), а также показатель семантической близости в парах ссылка/целевой текст, предложение/целевой текст, исходный текст/целевой текст. Семантическая разметка строится с помощью открытого фреймворка «WebVectors», который строит векторные модели дистрибутивной семантики слов. Дистрибутивный подход основан на вычислении степени семантиче-

ской близости между языковыми единицами с учётом их сочетаемости: чем чаще лексемы образуют одинаковые коллокации, тем ближе они друг к другу по значению [Kutuzov 2017: 155]. Таким образом, мы получаем количественные показатели семантической близости от 0 до 1, где 0 означает отсутствие семантических пересечений, а 1 — абсолютную синонимию .

2. Результаты исследования

Частеречный анализ показал, что большинство ссылок включает в свой состав глагольные (61 %), субстантивные (38 %) и адъективные (15 %) номинации. Ссылки, характеризующиеся максимальной степенью семантической близости с целевым текстом, указывают на конкретный референт сообщения. Несмотря на потенциальное тематическое разнообразие целевых текстов, набор номинаций оказывается ограничен спецификой публицистического дискурса. Мы выделяем следующие тематические группы (в скобках указана общее число вхождений и средняя семантическая близость в парах ссылка/текст): судебно-административная сфера (*судья* (1/0,54), *прокуратура* (6/0,31) *задержать* (297/0,26), *арестовать* (152/0,30)); происшествия (*выброс* (1/0,43), *пострадать* (93/0,18), *погибнуть* (52/0,26)); политика (*праймериз* (1/0,53), *импичмент* (1/0,53), *агитация* (1/0,49), *ратифицировать* (4/0,46)); финансово-экономическая сфера (*баррель* (4/0,61), *доходность* (1/0,62), *трейдер* (1/0,46), *пошлина* (10/0,43) *подешеветь* (6/0,37)); волеизъявление (*одобрить* (91/0,20), *поручить* (60/0,20), *требовать* (66/0,20)).

Семантика гипертекстового перехода оказывается детерминирована темой целевого текста и локализована в номинации ссылки. Таким образом, использование ссылок с сильной семантической связью, обусловлено установкой автора на актуализации референта в целевом тексте, что позволяет читателю получить знание о пресуппозиции высказывания автора. Кроме того, функции ссылок могут расширяться за счет выражения модальной оценки содержания целевого текста. Так, в предложении *Мегин Келли заявила, что считает Владимира Путина очень умным человеком, которого не получится «перемудрить»* в качестве источника ссылки использована лексема *перемудрить*, что выражает модальность стимулирует читателя к совершению гипертекстового перехода. Социальная направленность номинации ссылки выражается в лексемах с семантическим компонентом усиление

(*ужесточение*, *ужесточить*), репрессии (*рабство*, *линчевать, обезглавить*) или преодоления (*прорвались*).

Ссылки со слабой семантической связью выполняют эвиденциальную функцию: они маркируют гипертекстовый переход, который подтверждает достоверность информации в исходном тексте. Большинство таких ссылок являются глагольными и относится к лексико-семантической группе «сообщение»: *заявить* (1320/0,21), *заявлять* (202/0,20), *сообщить* (1020/0,20), *сообщать* (722/0,20), *писать* (450/0,15), *объявить* (366/0,19). Использование глагольных ссылок объясняется тем, что они представляют собой предикатный центр предложения, содержащий в себе в свёрнутом виде отдельный морфосинтаксический паттерн языка. Таким образом, при использование глагольных ссылок референциальные связи выражаются с помощью всей аргументно-предикативной конструкции. Например, в предложении *напомним, на Страстном бульваре в Москве у памятника писателю Александру Твардовскому 12 человек 10 сентября заявили о бессрочной акции протеста*, где ссылка имеет номинацию *заявили*, на семантику гипертекстового перехода указывает субъект действия (12 человек), объект (*акция протеста*), обстоятельства места (*на Страстном бульваре в Москве у памятника писателю Александру Твардовскому*) *и* времени (*10 сентября*). Обращает на себя внимание и тот факт, что целевой текст незаконно занимает позицию объекта по отношению к ссылке-предикату. Таким образом, валентность глагола реализуется и в линейном, и гипертекстовом пространстве.

Субстантивные ссылки эвиденциального типа представлены рядом номинаций, указывающих на формат целевого текста: *публикация* (188/0,22), *интервью* (113/0,20), *сайт* (58/0,17), *сообщение* (44/0,19), *информация* (33/0,20), *заявление* (27/0,22), *материал* (27/0,19). К данной группе примыкают ссылки-имена собственные, которые чаще всего указывают на название ресурса-первоисточника: *ТАСС*, *РБК*, *Интерфакс и др*. Например: *Соответствующие изменения вносятся в закон «О порядке выезда из Российской Федерации и въезда в Российскую Федерацию», пишет ТАСС.* Таким образом, субстантивная ссылка входит в состав субъектно-предикатной конструкции, аккумулируя в себе тем самым лексическое наполнение всего высказывания.

Заключение

Мы выявили две основные стратегии автора при создании электронного гипертекста. Во-первых, номинация ссылки может содержать реакцию на содержание целевого текста. В этом случае семантика гипертекстового перехода отражается в номинации ссылки, которая вступает в паратекстуальные отношения с целевым текстом. Во-вторых, может быть использована используется стратегия подтверждения эвиденциональности исходного текста, за счет отсылки к первоисточнику. В этом случае ссылка указывает на вид деятельности, формат или жанр целевого текста, а семантика гипертекстового перехода распределяется по всем компонентам аргументно-предикатной конструкции. Перспективой нашего исследования является расширение корпуса, за счет концептуально-устных жанров интернет-коммуникации. Это позволит провести сопоставительный анализ и выявить степень влияния стиля речи на проявление рецепции автора в выборе номинации ссылки.

Литература

1. *Дедова О. В.* (2008) Теория гипертекста и гипертекстовые практики в Рунете. М.
2. *Потапова Р. К.* (2002) Новые информационные технологии и лингвистика, М.: МГЛУ
3. *Клочкова Е. С.* (2015) Субстантивные гиперссылки как средство расширения тезауруса и создания ассоциативного поля. Материалы IV международной научно-практической конференции «Гипертекст как объект лингвистического исследования». С. 52–59 с.
4. *Шульгинов В. А.* (2016) Когнитивная модель электронного гипертекста. Вестник Кемеровского государственного университета. 2016. № 4 (68). С. 233–238.
5. *Kutuzov A., Kuzmenko E.* (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham.

References

1. *Dedova O. V.* (2008) Teoriia giperteksta i gipertekstovye praktiki v Runeteve [The theory of hypertext and hypertext practice in RuNet]. Moscow.
2. *Potapova R. K.* (2002) Novye informacionnye tekhnologii i lingvistika [New information technologies and linguistics]. Moscow.
3. *Klochkova E. S.* (2015) Substantivnye giperssylki kak sredstvo rasshireniya tezaurusa i sozdaniya associativnogo polya [Substantive hyperlinks as a means of expanding the thesaurus and creating an associative field]. Materialy IV mezhdunarodnoj nauchno-

prakticheskoj konferencii «Gipertekst kak objekt lingvisticheskogo issledovaniya» [Proceedings of the IV International Scientific Practical Conference "Hypertext as an object of linguistic research"], pp. 52–59.
4. *Shulginov V. A.* (2016) Kognitivnaia model' elektronnogo giperteksta [Cognitive model of hypertext]. Vestnik Kemerovskogo gosudarstvennogo universiteta [Bulletin of Kemerovo State University], (4), pp. 233–238.
5. *Kutuzov. A., Kuzmenko E.* (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham.

Шульгинов Валерий Александрович
Дальневосточный федеральный университет (Россия)
Shulginov Valery
Far Eastern Federal University
E-mail: prostovalera@yandex.ru

Шульгинов Вадим Александрович
Ростелеком (Россия)
Shulginov Vadim
Rostelecom (Russia)
E-mail: vadim.shulginov@yandex.ru

ДИАЛЕКТНЫЕ И ИСТОРИЧЕСКИЕ КОРПУСЫ

DIALECTAL AND HISTORICAL CORPORA

И. В. Азарова, Е. Л. Алексеева, А. М. Лаврентьев, Е. А. Рогозина, К. В. Сипунин

I. V. Azarova, E. L. Alexeeva, A. M. Lavrentiev, E. A. Rogozina, K. V. Sipunin

ПРЕДСТАВЛЕНИЕ И АНАЛИЗ БИБЛЕЙСКИХ, СВЯТООТЕЧЕСКИХ И ЛИТУРГИЧЕСКИХ ЦИТАТ В КОРПУСЕ СКАТ

REPRESENTATION AND ANALYSIS OF BIBLICAL, PATRISTIC AND LITURGICAL QUOTES IN *SCAT*

Аннотация. В докладе рассматривается форма представления цитат из Библии, сочинений Отцов Церкви и литургических текстов в составе житий Санкт-Петербургского корпуса агиографических текстов (СКАТ). Формат разметки цитат опирается на рекомендации консорциума TEI (Text Encoding Initiative). Перенос корпуса на платформу TXM позволяет применить к XML-представлению цитат ряд стастистических методов, такие как вычисление специфичности и факторный анализ соответствий.
Ключевые слова. Санкт-Петербургский корпус агиографических текстов СКАТ, платформа TXM, древнерусская агиография, библейские цитаты, святоотеческие цитаты, устойчивые сочетания лексем.

Abstract. The paper deals with the representation of quotes from the Bible, patristic and liturgical sources within hagiographic texts comprising the SCAT Corpus (St Petersburg Corpus of Hagiographic Texts). The format we use follows the TEI recommendations (Text Encoding Initiative). Since the corpus have been imported to the TXM platform, a number of statistical procedures can be applied to the XML tagging of quotes, including the specificity calculation and correspondence analysis.
Keywords. SCAT, St Petersburg Corpus of Hagiographic Texts, TXM platform, Old Russion hagiography, Bible quotes, patristic quotes, set phrases.

1. Представление цитат в корпусе СКАТ

Санкт-Петербургский корпус агиографических текстов (СКАТ, http://project.phil.spbu.ru/scat) содержит свыше 20 севернорусских

житийных текстов XV–XVII вв. В настоящее время в сотрудничестве с лабораторией IHRIM (Лион) осуществляется перенос корпуса на платформу TXM (http://textometrie.org), что делает, в частности, доступным комплекс текстометрических процедур, разработанных для этой платформы [Лаврентьев и др. 2018; Azarova et al. 2018].

Одним из интересных направлений корпусных исследований является выявление интертекстуальных отношений, в связи с чем была начата работа по обозначению библейских цитат и цитат из святоотеческих сочинений и литургических текстов в житиях корпуса.

При разметке цитат в формате XML мы следуем рекомендациям консорциума Text Encoding Initiative (TEI). Для обозначения границ цитат используется тег <q>, который не только позволяет отметить начало и конец цитаты, но и привести источник цитирования с помощью атрибута @source, ссылающегося на внешний файл с источниками цитат. Дополнительно мы используем тег <seg>, чтобы отделить слова, вводящие цитату, от собственно цитаты и атрибут @type, могущий принимать значения: author — авторская речь, speaker — цитируемый автор, modified — текст цитаты перестроен, allusion — аллюзия на соответствующий фрагмент Библии или другого источника. При работе с TXM XML-разметка цитат позволяет создавать подкорпусы и разбивки и применять такие статистические методы, как вычисление специфичности и факторный анализ соответствий (Correspondence analysis) [Guillot et al. 2013]. Формат разметки реализован на материале Жития Дионисия Глушицкого.

В агиографической литературе использование цитат выполняет очень важную функцию: цитата своим авторитетом подтверждает и подчеркивает значительность описываемого события. Исследователи житий давно обратили внимание на то, что построение текста следует определенному канону: автор находит ближайший прототип для святого, чью жизнь и подвиг он должен описать, и следует ему [Панченко 2003]. Неслучаен и выбор цитат. М. К. Кузьмина показала, что для каждого сюжетного фрагмента характерен свой набор цитат и что любая конкретная цитата употребляется только в определенных сюжетах [Кузьмина 2017].

2. Выявление особенностей цитатного материала

Выделение цитат в текстах корпуса позволяет поставить ряд исследовательских задач.

1. Обычное наблюдение показывает, что текстовые свойства цитат не вполне совпадают со свойствами авторского текста: в цитатах могут встречаться архаичные формы слов (например, Им. пад. ед. ч. *любы* в цитате из Евангелия от Иоанна, в то время как в других случаях используется только форма *любовь*), особая лексика (*при исходищих водных из* Псалтыри). В лаборатории IHRIM имеется опыт работы по исследованию особенностей прямой речи в старофранцузском корпусе [Guillot et al. 2015], мы продемонстрируем результаты анализа специфичности и факторного анализа соответствий на материале лексики и морфологических тегов цитат в контрасте с общим житийным корпусом.

2. Представляют несомненный интерес устойчивые сочетания лексем в библейских цитатах. Рассмотрим словосочетание *жизнь вечная*. Анализ употребления компонентов этого сочетания лексем в текстах Нового и Ветхого Заветов (далее — НЗ и ВЗ) показывает, что в НЗ это сочетание встречается 9 раз, но чаще (22) оно встречается в синонимичном перефразировании: *живот вечный*, поскольку *жизнь* в этом значении выглядит как инновация, хотя и употребляется в текстах Нового и Ветхого Заветов 95 раз в этом значении (против 124 употреблений *живот*). Более показательным является сочетаемость прилагательного *вечный* (всего 238 употреблений, из которых 71 в НЗ). В ветхозаветных текстах это прилагательное обладает довольно широкой сочетаемостью: *вечный закон/ завет/ бог/ огонь, вечное имя/ поношение/время; вечная память/ правда/ пустыня; вечные роды/ соли/ холмы/ горы* и т. п. В новозаветных текстах оно реже появляется, но при этом из 71 вхождения 31 — это сочетания со словами *жизнь* и *живот*.

Таким образом, необходимо исследовать такие устойчивые сочетания лексических единиц, которые будут представлены в житийных цитатах библейских текстов. Синтаксические сочетания лексем, представленных в цитатах, необходимо оценить при помощи стандартных индексов извлечения коллокаций, например MI-score [Азарова и др. 2005], однако важным является, что в условиях становления жанровых особенностей текстов мы видим кодификацию не столько сочетания лексем, сколько способов выражения смысла «ключевых клише». Основными параметрами исследования подобной «устойчивости» будут те, которые продемонстрированы в примере: — ограничение сочетаемости лексемы в новозаветных текстах; — появление особых сочетаний лексем, не встречающихся в ветхозаветных текстах или встречающиеся редко; — наличие синонимических вариантов для

одного из структурных элементов сочетания. Безусловно, такие специфические новозаветные сочетания могут получаться за счет появления лексем, которые вовсе не встречались в ветхозаветных текстах.

Литература

1. *Азарова И. В., Синопальникова А. А., Смрж П.* (2005), Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet. Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2005 (Звенигород, 1–7 июня 2005 г.), с. 11–17.
2. *Кузьмина М. К.* (2017), Канон преподобнического жития сквозь призму библейских цитат. М.
3. *Лаврентьев А. М, Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.* (2018), Новый комплекс инструментов автоматической обработки текста для платформы ТХМ и его апробация на корпусе для анализа экстремистских текстов. Вестник НГУ. Серия Лингвистика и межкультурная коммуникация, 16 (3), с. 9–31.
4. *Панченко О. В.* (2003), Поэтика уподоблений (к вопросу о «типологическом» методе в древнерусской агиографии, эпидейктике и гимнографии). ТОДРЛ, Т. 54, с. 491–534.
5. *Azarova I., Alexeeva E., Lavrentiev A, Sipunin K., Rogozina E.* (2018), Using TXM Platform to optimize textual information retrieval and representation in the St. Petersburg corpus of hagiographic texts (SCAT). El'Manuscript 2018. Abstracts, Participants, Programme, Vienna, Krems, p. 56.
6. *Guillot C., Lavrentiev A., Pincemin B., Heiden S.* (2013) Le discours direct au Moyen Âge: vers une définition et une méthodologie d'analyse. D. Lagorgette, P. Larrivée (ed.). Représentations du sens linguistique 5, Chambéry, Université de Savoie, p. 17–41.
7. *Guillot C., Heiden S., Lavrentiev A., Pincemin B.* (2015), L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique. // Kirsten Jeppesen Kragh; Jan Lindschouw (ed.) Les variations diasystémiques et leurs interdépendances dans les langues romanes. Actes du Colloque DIA II à Copenhague (19–21 nov. 2012), Éditions de linguistique et de philologie, p. 15–28.

References

1. *Azarova I. V., Sinopal'nikova A. A., Smrzh P.* (2005), Predstavlenie ustojchivyh leksicheskih sochetanij v komp'juternom tezauruse RussNet [Representation of multiword expressions in computer thesaurus RussNet]. Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnoj konferencii Dialog'2005 (Zvenigorod, 1–7 ijunja 2005 g.) [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog'2005 (Zvenigorod, 1–7 June 2005)], pp. 11–17.
2. *Kuz'mina M. K.* (2017), Kanon prepodobnicheskogo zhitija skvoz' prizmu biblejskih citat [Canon of Monastic Vitae through a Prism of Bible Quotes]. Moscow.

3. *Lavrent'ev A. M, Solov'ev F. N., Suvorova M. I., Fokina A. I., Chepovskij A. M.* (2018), Novyj kompleks instrumentov avtomaticheskoj obrabotki teksta dlja platformy TXM i ego aprobacija na korpuse dlja analiza jekstremistskih tekstov [A New Toolkit for Natural Text Processing with the TXM Platform and its Application to a Corpus for Analysis of Texts Propagating Extremist Views]. Vestnik NGU. Serija Lingvistika i mezhkul'turnaja kommunikacija [NSU Vestnik Journal, Series: Linguistics and Intercultural Communication.], 16 (3), pp. 9–31.
4. *Panchenko O. V.* (2003), Poetika upodoblenij (k voprosu o «tipologicheskom» metode v drevnerusskoj agiografii, jepidejktike i gimnografii) [Poetics of Assimilation (on the 'typological' method in the Old Russian Hagiography, epideictic oratory and hymnography)]. TODRL, Vol. 54, pp. 491–534.
5. *Azarova I., Alexeeva E., Lavrentiev A, Sipunin K., Rogozina E.* (2018), Using TXM Platform to optimize textual information retrieval and representation in the St. Petersburg corpus of hagiographic texts (SCAT). El'Manuscript 2018. Abstracts, Participants, Programme, Vienna, Krems, p. 56.
6. *Guillot C., Lavrentiev A., Pincemin B., Heiden S.* (2013) Le discours direct au Moyen Âge: vers une définition et une méthodologie d'analyse. D. Lagorgette, P. Larrivée (ed.). Représentations du sens linguistique 5, Chambéry, Université de Savoie, p. 17–41.
7. *Guillot C., Heiden S., Lavrentiev A., Pincemin B.* (2015), L'oral représenté dans un corpus de français médiéval (9e-15e): approche contrastive et outillée de la variation diasystémique. // Kirsten Jeppesen Kragh; Jan Lindschouw (ed.) Les variations diasystémiques et leurs interdépendances dans les langues romanes. Actes du Colloque DIA II à Copenhague (19–21 nov. 2012), Éditions de linguistique et de philologie, p. 15–28.

Азарова Ирина Владимировна
Санкт-Петербургский государственный университет (Россия)
Azarova Irina
Saint Petersburg State University (Russia)
E-mail: ivazarova @gmail.com

Алексеева Елена Леонидовна
Санкт-Петербургский государственный университет (Россия)
Alexeeva Elena
Saint Petersburg State University (Russia)
E-mail: el.alexeeva @gmail.com

Лаврентьев Алексей Михайлович
Национальный центр научных исследований, лаборатория IHRIM (Франция)
Lavrentiev Alexei
Centre national de la recherche scientifique, IHRIM
(Institut d'Histoire des Représentations et des Idées dans les Modernités) (France)
E-mail: alexei.lavrentev@ens-lyon.fr

Рогозина Елена Андреевна
Санкт-Петербургский государственный университет (Россия)
Rogozina Elena
Saint Petersburg State University (Russia)
E-mail: renehorn.r@gmail.com

Сипунин Константин Владимирович
Санкт-Петербургский государственный университет (Россия)
Sipunin Konstantin
Saint Petersburg State University (Russia)
E-mail: 79818514079@yandex.ru

В. А. Баранов
V. A. Baranov

ПОИСК И ДЕМОНСТРАЦИЯ ДАННЫХ В ИСТОРИЧЕСКОМ КОРПУСЕ «МАНУСКРИПТ»[1]

DATA RETRIEVAL AND DEMONSTRATION IN THE HISTORICAL CORPUS "MANUSCRIPT"

Аннотация. Представлены пользовательские модули исторического корпуса «Манускрипт» (manuscripts.ru), позволяющие сформировать запрос и визуализировать данные славянских средневековых кодексов X–XV веков. Рассказано о назначении модулей, описаны базовые параметры запросных форм. Основное внимание уделено возможностям модуля n-грамм и модуля статистики. Первый, давая возможность пользователю указать количество компонентов сочетания, расстояние между ними, порядок следования, статистическую меру и многие другие структурные и лингвистические параметры, позволяет получить сведения о статистических характеристиках n-грамм. Второй предоставляет возможность выявить распределение символов, словоформ, лемм в кодексе, сопоставить их количество в нескольких подкорпусах или получить статистические сведения о лингвистических единицах. Приведены примеры запросов и результатов их выполнения.
Ключевые слова. Корпусная лингвистика, корпусный менеджер, лингвистическая статистика, средневековые славянские рукописи.

Abstract. The article deals with the user modules of the historical corpus "Manuscript" (manuscripts.ru) providing means for query formation and visualization of data of the Slavonic medieval codices of the 10th – 15th centuries. The article details the purpose of the modules and describes the main parameters of the query forms. The main attention is given to the possibilities of the module of n-grams and the module of statistics. The first giving to the user the possibility of indication of the number of components in the combination, the distance between them, the sequence order, statistic measure and many other structural and linguistic parameters allows getting data on the statistic characteristics of the n-grams. The second gives the possibilities of revealing the distribution of symbols, word forms and lemmas in the codex and comparing their number in several sub-corpora or obtaining statistic data on the linguistic units. There are given examples of queries and results of their execution.
Keywords. Corpus linguistics, corpus manager, linguistic statistics, medieval Slavonic manuscripts.

1. Стандартные формы поиска и визуализации данных

Необходимыми составляющими любого корпуса являются размеченные тексты и процедуры обработки, поиска и демонстрации данных — корпусный менеджер.

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (РФФИ) в рамках проекта «Лингвостатистический анализ однокомпонентных и многокомпонентных лексических единиц исторического корпуса "Манускрипт"» (грант № 18-012-00463).

Исторический корпус «Манускрипт» (manuscripts.ru), содержащий полные транскрипции текстов средневековых славянских рукописей X–XV веков[2], предоставляет пользователю несколько способов доступа к данным — одношаговый и двушаговый интерфейсы, поиск материала в нескольких или только в одном документе, особую запросную форму для параллельных корпусов, а также специальные интерфейсы для извлечения n-грамм и для получения количественных и статистических сведений о единицах корпуса.

Созданные в разное время, предназначенные для решения различных задач, запросные формы имеют тем не менее несколько идентичных параметров: указание маски лингвистической единицы и ее грамматических значений, выбор алфавита визуализации и формы вывода — текста, перечней, конкордансов. Если текст и/или рукопись имеют аналитическую разметку, то пользователь может увидеть соответствующие фрагменты[3] или построить указатели их словоформ или лемм. Использование многотекстовой запросной формы позволяет получить сопоставительные перечни лингвистических единиц, которые дают возможность увидеть различия в лексическом составе кодексов.

Особое внимание уделено средствам нечеткого поиска, который необходим для нивелирования различий в написании одной и той же словоформы. Пользователь может 1) выбрать одну из предложенных степеней точности маски, что позволяет приравнивать или различать строчные, надстрочные и инициальные буквы, учитывать или не учитывать их вариантность, унифицировать или устранять диакритику и титла, раскрывать лигатуры и под.; 2) использовать при создании маски регулярные выражения; 3) создать маску с помощью современного кирилловского алфавита; 4) выбрать поиск на основе лемм.

Отличаются от современных корпусов и способы доступа к документам[4]: пользователь может выбрать для работы отдельную рукопись или уже сформированную коллекцию, которые доступны на отдельных страницах портала, создать подкорпус или сделать запрос по всему корпусу.

---

[2] О корпусе см., например, [Баранов 2015].

[3] Например, главы и стихи библейских книг, песнопения гимнографических текстов, погодные записи летописей и др.

[4] Основными единицами базы данных являются физический экземпляр рукописи или ее сохранившийся отрывок, текст (произведение) и символ. Иерархическая модель позволяет осуществлять структурную, аналитическую, морфологическую и др. разметки «промежуточных» единиц.

2. Специализированные формы демонстрации данных

Корпус имеет две специализированные возможности доступа к данным — модуль n-грамм и модуль статистики.

2.1. Модуль n-грамм

Модуль[5] предназначен для получения сведений о сочетаемости лингвистических единиц. Особенностью запросной формы является наличие параметров, обеспечивающих получение количественной и статистической информации о разнообразных многокомпонентных сочетаниях:

— базовые — количество компонентов и статистическая мера;
— структурные — расстояние между компонентами, их следование, точность совпадения маски и текстового прецедента и нек. др.;
— лингвистические — тип компонентов (словоформа / лемма), их грамматические значения, учет стоп-слов и границ конструкций и др.

В настоящее время пользователю доступны семь статистических мер — Mutual Information score (и вариант Pointwise Mutual Information score), T-score, Log-Likelihood score (и вариант $\text{Log-Likelihood}_{\text{Dunning}}$), Dice coefficient (и вариант logDice), Chi-squared test, C-value, Inside[6].

Одной из целей анализа n-грамм в корпусе, как известно, является поиск устойчивых сочетаний — коллокаций и коллигаций. Модуль позволяет проводить различные эксперименты, используя сочетания параметров и их значений в зависимости от решаемых задач. Так, выявление симметричных сочетаний возможно двумя способами — с помощью меры Dice и с помощью параметра «Симметрия», который так же, как статистическая мера, извлекает из подкорпуса n-граммы, компоненты которых встречаются только вместе.

Одновременно могут использоваться несколько параметров. Так, результатом запроса, приведенного на рисунке 1[7], стали биграммы *страхъ ѓ҃нь* (значение меры 2,82), (*отъ*) *лица страха* (1,99), *страсть*

---

[5] Адрес запросной формы: http://manuscripts.ru/mns/cred_ngr.stat .

[6] О статистических мерах см., напр., [Evert 2004; Mima et al. 1998].

[7] Подкорпус двух списков Паримейника (http://manuscripts.ru/mns/portal.main?p1=57), мера T-score, маска ^стра[хшс][^ст].*, значение *существительные*, свободный порядок следования, ислкючение стоп-слов.

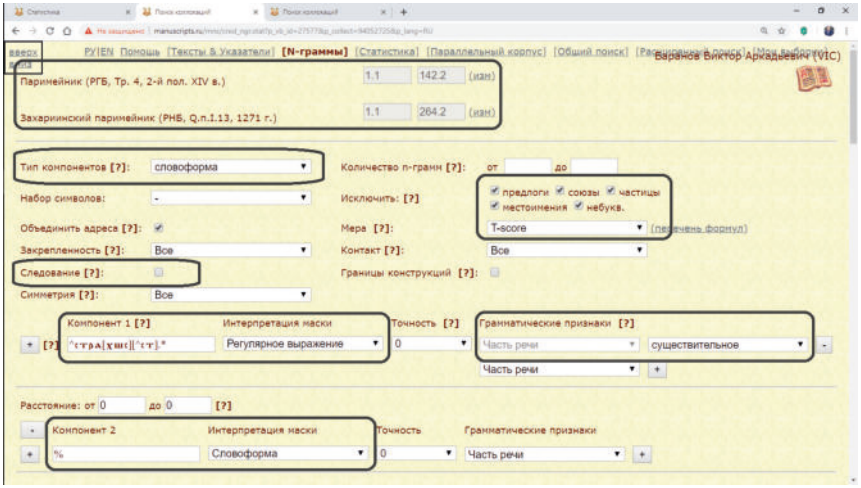


Рис. 1. **Извлечение биграмм с компонентом *страх*(*ш*) из подкорпуса Паримейников**

трепетъ (1,41), *прм̑дрости страхъ* (1,41), *w*^{m}*врьзаютъ страхы* (1,00) и нек. др.

2.2. Модуль статистики

Второй специализированный модуль[8] сегодня имеет несколько режимов работы: а) режим выявления распределения единиц (символов, словоформ, лемм) в рукописи(ях), б) режим количественной оценки единиц (символов, словоформ, лемм, фрагментов, текстов) в пределах подкорпуса(ов), в) режим статистической оценки лингвистических единиц подкорпуса в сопоставлении со средней их частотностью в контрастном подкорпусе.

На рисунке 2 показан результат запроса по поиску распределения словоформ с начальным *страх*(*ш*) в трех списках летописей (Лаврентьевском, Ипатьевском, Радзивилловском[9]). Графики, выровненные по погодным записям, демонстрируют бóльшую частотность этих слов в некоторых фрагментах второй половины списков, например в погодных записях с 1095 по 1098 гг.

Использование второго режима позволяет сравнить абсолютные и относительные частоты одной единицы в нескольких подкорпусах.

[8] Адрес запросной формы: http://manuscripts.ru/mns/!cred2.stat .

[9] Страница коллекции: http://manuscripts.ru/mns/portal.main?p1=23 .

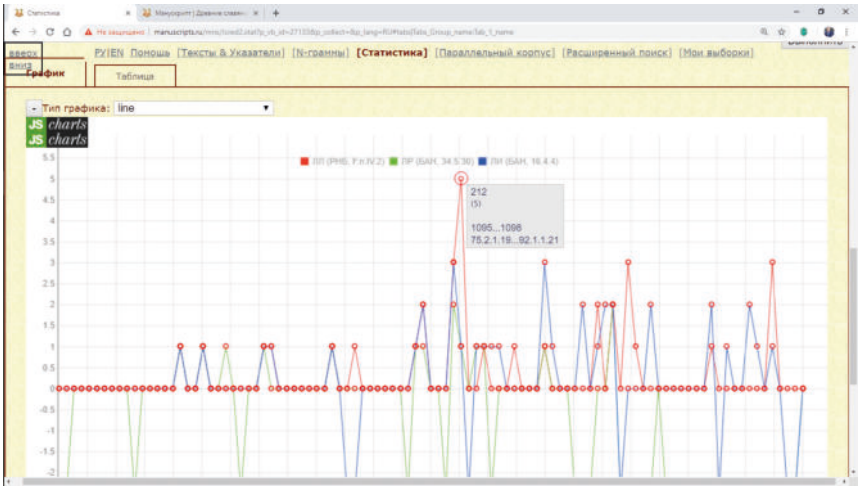


Рис. 2. **Распределение словоформ с начальным *страх*(*ш*) в трех списках летописей**

При использовании неточной маски словоформы могут быть показаны как списком, так и суммированно. В таблице 1 приведено начало списка слов с начальным *страх*(*ш*) и их абсолютное и относительное

Таблица 1. **Словоформы с начальным *страх*(*ш*) в летописях и Паримейниках (начало списка)**[10]

Единицы	Ранг	Всего		Подкорпус летописей		Подкорпус Паримейников	
		436 452		349 913		86 539	
		F	f	F	f	F	f
страхъ	1	35	0,00008	19	0,00005	16	0,00018
страха	2	28	0,00006	15	0,00004	13	0,00015
страшно	3	19	0,00004	17	0,00005	2	0,00002
страшенъ	4	8	0,00002	7	0,00002	1	0,00001
страхомъ	5	7	0,00002	6	0,00002	1	0,00001
страшна	6	5	0,00001	3	0,00001	2	0,00002

[10] В таблице: F — абсолютное, f — относительное количество.

количество в подкорпусах летописей и Паримейников. Режим суммирования позволяет получить относительное количество тех же слов, которое для этой выборки равно 0,00026 и 0,00060 соответственно. Оба представления показывают бóльшую частотность этих слов в текстах Паримейников.

Третий режим предназначен для статистического сопоставления частот лингвистических единиц с их средними частотами в контрастном подкорпусе. Пользователь имеет возможность использовать меры Log-Likelihood, T-score и Weirdness (и их варианты, обеспечивающие вычисления и при отсутствии формы в контрастном подкорпусе)[11]. Например, оценка слов с начальными *стра*х(*ш*) в подкорпусах летописей и Паримейников на фоне большей части текстов исторического корпуса показывает более чем двукратное превышение значения меры T-score в Паримейниках — 0,00185 vs 0,00081.

3. Подготовка подкорпуса и интеграция модулей

В связи с тем, что лингвистический анализ часто требует сопоставления материалов подкорпусов, обладающих различными характеристиками, важной составляющей корпусного менеджера является обеспечение формирования подкорпусов на основе метаразметки. Особенностью корпуса «Манускрипт» является возможность подготовки подкорпуса не только на основе характеристик текстов и рукописей, но и на основе разметки фрагментов, что позволяет исследовать, в частности, текстологически неоднородные по составу кодексы. Так, могут быть созданы подкорпуса стихир и канонов служебных миней, подкорпуса глав евангельских книг и т. п. В таблице 2 представлены первые 10 наиболее частотных слов и форм тропарей и стихир нескольких служебных миней XI–XIV вв.[12]

После разработки двух новых модулей — модуля статистики и модуля n-грамм исторически сложившееся разнообразие способов доступа потребовало их интеграции, которая бы устранила необходимость при переходе из одного модуля в другой вновь формировать подкорпус.

В настоящее время подготовленный подкорпус может быть сохранен и использован в различных модулях. При переключении между

[11] О статистических мерах см., например, [Ahmad, Gillam, Tostevin 1999; Rayson, Garside 2000; Roelleke 2013].

[12] Коллекция миней: http://manuscripts.ru/mns/portal.main?p1=10 .

запросными формами загруженные подкорпуса остаются доступными для выборки, а в модуле статистики часть из них может быть удалена или к ним могут быть подгружены другие.

Таблица 2. **Наиболее частотные словоформы в тропарях и стихирах славянских служебных миней XI–XIV вв.**[13]

Единицы	Ранг	Подкорпус миней. Всего		Стихиры миней		Тропари миней	
		15 160		8 843		6 317	
		F	f	F	f	F	f
и	1	603	0,03978	382	0,04320	221	0,03498
въ	2	191	0,01260	118	0,01334	73	0,01156
ти	3	128	0,00844	72	0,00814	56	0,00886
яко	4	126	0,00831	60	0,00679	66	0,01045
съ	5	104	0,00686	76	0,00859	28	0,00443
тѧ	6	94	0,00620	47	0,00531	47	0,00744
на	7	92	0,00607	50	0,00565	42	0,00665
сѧ	8	87	0,00574	48	0,00543	39	0,00617
ѥси	9	85	0,00561	39	0,00441	46	0,00728
же	10	84	0,00554	47	0,00531	37	0,00586

Заключение

При создании интерфейса доступа к корпусу должны быть решены многообразные технологические задачи, обеспечены функциональность и дружественность форм запросов и вывода. За более чем пятнадцатилетнюю работу над корпусом коллектив проекта создал несколько версий каждого из модулей, совершенствуя одни функции и отказываясь от других. Некоторые версии модулей так и остались экспериментальными. Существенных нововведений потребовали предоставление пользователю возможности работы с аналитической разметкой и создание статистических модулей.

[13] Веб-форма позволяет сортировать данные таблицы по значениям любого столбца.

Поиски наиболее оптимальных форм запросов и вывода данных исторического корпуса продолжаются.

Литература

1. *Баранов В. А.* (2015), Исторический корпус как цель и инструмент корпусной палеославистики. Scripta & e-Scripta, Vol. 14–15, с. 39–62.
2. *Ahmad K., Gillam L., Tostevin L.* (1999), University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER), Proc. 8th Text Retrieval Conference TREC, pp. 717–724.
3. *Evert S.* (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations, Dissertation, Stuttgart, available at: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf .
4. *Mima H., Frantzi K. T., Ananiadou S.* (1998), The C-value/Example-based Approach to the Automatic Recognition of Multi-Word Terms for Cross-Language Terminology, The Pacific Rim International Conferences on Artificial Intelligence (PRICAI'98), available at: https://www.researchgate.net/publication/304254311 .
5. *Rayson P., Garside R.* (2000), Comparing corpora using frequency profiling, Proceedings of the Comparing Corpora Workshop at ACL 2000, pp. 1–6, available at: http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf .
6. *Roelleke T.* (2013), Information Retrieval Models: Foundations and Relationships, Synthesis Lectures on Information Concepts, Retrieval, and Services, Vol. 5, no. 3, pp. 1–163, available at: https://www.morganclaypool.com/doi/abs/10.2200/S00494ED-1V01Y201304ICR027 .

References

1. *Baranov V. A.* (2015), Istoricheskij korpus kak cel' i instrument korpusnoj paleoslavistikiIstoricheskij korpus kak cel' i instrument korpusnoj paleoslavistiki. [Historical Corpus as the Goal and Tool for Corpus Paleoslavistics]. Scripta & e-Scripta, Vol. 14–15, pp. 39–62.
2. *Ahmad K., Gillam L., Tostevin L.* (1999), University of surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER), Proc. 8th Text Retrieval Conference TREC, pp. 717–724.
3. *Evert S.* (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations, Dissertation, Stuttgart, available at: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf .
4. *Mima H., Frantzi K. T., Ananiadou S.* (1998), The C-value/Example-based Approach to the Automatic Recognition of Multi-Word Terms for Cross-Language Terminology, The Pacific Rim International Conferences on Artificial Intelligence (PRICAI'98), available at: https://www.researchgate.net/publication/304254311 .
5. *Rayson P., Garside R.* (2000), Comparing corpora using frequency profiling, Proceedings of the Comparing Corpora Workshop at ACL 2000, pp. 1–6, available at: http://ucrel.lancs.ac.uk/people/paul/publications/rg_acl2000.pdf .

6. *Roelleke T.* (2013), Information Retrieval Models: Foundations and Relationships, Synthesis Lectures on Information Concepts, Retrieval, and Services, Vol. 5, no. 3, pp. 1–163, available at: https://www.morganclaypool.com/doi/abs/10.2200/S00494ED1V01Y201304ICR027 .

Баранов Виктор Аркадьевич
Ижевский государственный технический университет
имени М. Т. Калашникова (Россия)
Baranov Victor
Kalashnikov Technical State University (Russia)
E-mail: victor.a.baranov@gmail.com

С. С. Земичева
S. S. Zemicheva

НОВЫЕ ТЕМЫ ДИАЛЕКТНОГО ДИСКУРСА (НА МАТЕРИАЛЕ ТОМСКОГО ДИАЛЕКТНОГО КОРПУСА)[1]

NEW TOPICS OF THE DIALECT DISCOURSE (BASED ON THE TOMSK DIALECT CORPUS MATERIAL)

Аннотация. На материале Томского диалектного корпуса рассматривается возможности исследования тематической организации диалектного дискурса. Описываются фрагменты, относящиеся к темам «Техника» и «Экология». Анализируется специфика их описания в диалектном дискурсе с учетом темпорального и гендерного факторов, связи с традиционными установками сельского социума. Структурируется содержание данных тематических фрагментов. Анализируются причины их актуализации в дискурсе.
Ключевые слова. Томский диалектный корпус, русские говоры Сибири, тематическая разметка.

Abstract. The possibilities of studying the thematic organization of the dialect discourse are considered on the material of the Tomsk dialect corpus. Fragments of the topics «Technique» and «Ecology» are described. The specificity of their description in the dialect discourse is analyzed, taking into account temporal and gender factors, and the connection with the traditional attitudes of rural society. The content of these thematic fragments is structured. The reasons for their actualization in the discourse are analyzed.
Keywords. Tomsk dialect corpus, Siberian dialects of Russian, topic annotation.

Томский диалектный корпус создаётся с 2017 г. на материале диалектологических экспедиций в среднеобский регион (Томская, центральная часть Кемеровской области). Объём ресурса на сегодня — более миллиона словоупотреблений. Такие особенности корпуса, как большой временной охват (70 лет) и детальная тематическая разметка (73 темы), дают возможность постановки новых исследовательских задач с опорой на эти параметры [Корпус. Демо-версия].

Наряду с сохранением констант традиционной культуры, материалы корпуса отражают перемены, связанные с её трансформацией. Так, тема «ТЕХНИКА», хотя и не относится к числу частотных, однако обладает определённым потенциалом для исследования мировоззрения сельских жителей, в том числе в микродиахроническом ключе.

[1] Исследование выполнено при поддержке гранта РФФИ № 19-012-00320 «Томский диалектный корпус как новый ресурс для изучения народно-речевой культуры».

Таблица 1. **Распределение количества текстов, включающих тематический фрагмент «Техника», по временным периодам**

Период	50-е	60-е	70-е	80-е	90-е — 2010-е	Итого
Фрагмент «Техника»	3	23	49	57	33	165
Всего текстов	77	163	257	349	145	991
%	4	14	19	16	23	17

Количественные данные (табл. 1) позволяют выявить отчётливую тенденцию — возрастание числа фрагментов, посвящённых данной теме, как в абсолютных, так и в относительных цифрах. Особенно резкий скачок наблюдается на рубеже 50-х и 60-х гг., когда активно проводилась политика механизации села.

Закономерно, что относительное число текстов, где актуализируется данная тема, несколько выше в речи мужчин: у них она встречается в 20 % текстов (53 из 264), тогда как у женщин — в 15 % (112 из 727).

Анализ текстовых фрагментов позволяет выделить повторяющиеся мотивы, раскрывающие содержание данной темы. Наиболее типичны в диалектном дискурсе контексты с оппозицией «Раньше/Теперь», где сопоставляется жизнь до широкого использования сельскохозяйственной техники и после: *Но'нче всё машины работают, а тода' и жали, и молотили, и косили — всё руками. Ра'зе что коровёнка либо лошадёнка пособит чё и всё. Тра'хторы-то уж в колхозе появились при советской власти. С* ***ыми-то много легче работать, тута-ка даже и сравнить грешно.*** (Батурино Шегарский р-н, 1967). Таким образом, описываются процессы индустриализации нашей страны, свидетелями которых были сами говорящие: *Теперь что, имеем и плиту га'зову, и свет. Теперь, говоришь, если бы поднялись старики: дак это что! В двадцатых годах у нас, нет в тридцатых, у нас провели телефон, а чё раньше? деревня глухая* (Зырянское, 1978); *В Пико'вке у нас радио не было. Я пися'т пятом году уехала, оттуда, у нас провели радио.* (Колпашево, 2016). Как правило, появление сельскохозяйственных машин оценивается положительно, так как они облегчают труд крестьянина: *Тапе'рь у нас совхоз.* ***Посмотришь — жи'зней залюбуешься.*** *Машины кругом, тра'хторы, комбайны, а раньше этого ничего не было*. (Малое Бабарыкино, 1967). Однако иногда отмечаются негативные последствия внедрения техники, например, что люди стали более ленивыми: *А сейчас машинами ко'сют, сгребают, а работать никто не хочет*. (Кулаково, 1974).

Часто встречаются рассказы о появлении первых технических приспособлений: *Первый трактор приехал — народ аж убился — бежал смотреть. Из-за речки трактор приехал…. Самолёт прилетел, так тож ведь народ из Молчановой весь в эропо'рт.* (Молчаново, 1975).

Отмечаются и перемены в крестьянском быту, связанные с покупкой новой техники: *У нас утюгов и не видели почти что, а потом уже маленько утюг купили с трубой, с углями, углями грелся, ну, такой огромный, туды' угли, гладит. Вам уже не приходилась видеть такие утюги, потому что вы выросли в такой жизни, с пядеся'т девятого году уже электрические, у нас уже никого не было в деревни, уже машины стиральные стали, с пядеся'т, да пядеся'т шестого или пядеся'т седьмого появились машины стиральные в деревнях, а раньше не видели, ну вот корыта, даже ванны не видели, не было.* (Кривошеино, 1978).

Первыми реакциями на технические нововведения у носителей традиционной культуры стали, с одной стороны, любопытство, с другой — страх: *У нас одни жили, как он торговал, дак все, так у него была машина* [*сельскохозяйственная*]. *Мы даже не видели, кака' она, только слышали, что трещит. Всё просились у отца, посмотреть как она работает. Что за машина.* (Нарым, 1983); *Мы с ним работали, когда первые машины были. Как нам сказали «машины», так мы с* ***испугу*** *борона побросали.* (Зырянское, 1978).

Корпус фиксирует и суеверия, связанные с внедрением техники: *У нас потом купили машину зажи'тошны. Идём мы на покос: «Тя'тя, погляди, чёрт ездит»* (Луговская,1972); *Правильно у нас старики говорили, что чёрт в углу будет говорить — то про радио, хорошо живём.* (Старая Шегарка, 1982). Такие представления типичны для народной культуры (довольно подробный анализ представлен в работе [Никитина 1999]).

Попытки осмыслить стремительные техногенные трансформации в отдельных случаях приводят к актуализации эсхатологических представлений, при этом развитие научно-технического прогресса связывается с падением нравственности: *Это идёт все по Писанию. Раньше у нас тётка жила, была. Она Евангиле читала, вот я тода' ничё не понимала, теперь вспоминаю. Как они говорили? Во, ребятишки, погодите, белый свет тенётами опутается, стальны' будут птицы летать, стальны' эти будут кони землю пахать. Это оно так и есть. Весь белый свет этими тенётами перепутался, так щас провода-то кругом запутали тенётами. Будет народ бесстыжий, беззаконствой, как скотина…* (Вершинино, 1991).

Ещё одним свидетельством перемен в жизни сельского общества является появление в диалектном дискурсе темы «Экология». Первые тексты подобного рода зафиксированы в 70-х гг., более распространены они в записях, сделанных в 80-е и позже. Хотя эта тема встречается в ограниченном количестве текстов (4 % всех), она отражает важные, на наш взгляд, изменения. Тема «Экология» имеет отчётливо выраженную гендерную специфику и актуализируется в речи мужчин в 2 раза чаще, чем в речи женщин [Земичева 2018]. Вероятно, это связано с тем, что в диалектном социуме сохраняется традиционное разделение труда, и охота, рыбалка, кедровый промысел, как и осмысление связанных с ними проблем, более свойственны мужчинам.

Текстовые фрагменты с темами «Техника» и «Экология» нередко пересекаются, так как крестьяне указывают на взаимосвязь данных явлений. При этом если отдельные технические нововведения, как уже отмечалось, воспринимаются информантами положительно, то общее влияние технического прогресса на трансформацию природы нередко получает негативную оценку: *А сейчас этими машинами вы'ковырят её* [*землю*]*, тракторами-то. А раньше пахали плугами, на лошадях. В глубину скажем пятнадцать-двадцать санти'метров, значит, это точка. А он щас где-то возьмёт бе'лу глину, тронет. Трактор — его же не будешь убавлять ка'жну минуту, гидравлика-то работает. А где не совсем не хватит. А потом они осенью щас вот зябь эту вспашут, а весной то'ко они заборо'нют её. Ну земля не стала плодить ничего.* (Парабель, 1985). В некоторых случаях вмешательство в природные процессы со стороны государства (связанное, в частности, с добычей нефти и газа как актуальными для региона отраслями промышленности) вызывает у сельских жителей отторжение, вступая в противоречие с традиционным укладом жизни: ***Мать-сырой земле нет покою,*** *всё изрыли. Геологи. В болоте нефть <задушили> и жгли её. Хоть бы хлеб родил.* (Инкино, 1972).

В зону осмысляемых экологических событий попадают также природные изменения, вызванные техногенными факторами: *А сейчас откуда рыбы много, матушка моя? Кругом нефть да газ. В реке раньше, вот, например, стерлядь взять. В Кете' она никогда не была, потому что это боло'тна река, там и вода-то чёрная. А сейчас деваться некуда, она и туда лезет, потому что, видишь, какое движение стало? Раньше что, на угле ходили, на дровах. От угля-то никакого вреда-то нет. А теперь ходят на этом, как его называют, на керосине. Вот и нет рыбы-то.* (Колпашево, 1983). Отмечается и изменение климата:

Как вот раньше: как с юга пасмурно — на другой день погода испортится. А теперь не сходится. Месяц на рогу, да плашмя лежит — какой урожай уродится. А теперь не сходится. Климат переменился — говорят. Раньше были страшны морозы, до пятьдесят два градуса. А теперь слабые. Зима морозна, лето — тепло, раньше говорили. А теперь зима — тепло, лето — сухо. Как-то не сходится по предметам разным. Не'которы говорят, что спутники лета'т. Все говорят, на погоду де'йствут, и на всё. Кто как говорят. (Колпашево, 1983). Как можно видеть, и здесь актуальна оппозиция «Раньше/Теперь».

В рамках темы «Экология» упоминается также загрязнение воды, воздуха, уменьшение количества рыбы, птицы, грибов и ягод. При этом осознание экологических проблем может происходить естественно, благодаря жизненному опыту крестьян: [*Сейчас уже не ходите к речке?*] *Нет, ну тут как-то раза два приходилось, не было в водокачке воды, дак ходили. А теперь её всю замусорили, всё кидают. Я говорю, всю жизь этой речкой пользовались, а теперь чё попало, весь хлам, весь сор туда, какие-то отходы и всё к берегу.* (Батурино Томский р-н, 2008).

В других случаях интерес к данной проблематике возникает, по-видимому, под влиянием средств массовой информации: ***Счас учёные сказали,*** *что нельзя бить кедру' колотом. Кедра от этого начинает болеть. <…> Лесные жители, вот скажем, к которым ближе кедрач, следили за ним больше, чем лесники. Раньше за жителем закрепляли участок, он его берёг, не трогал. Счас совсем не так относятся. Срубают ни за чё деревья. Лишь бы загубить. Нет бережного, любовного отношения к природе. Бережливость счас почему-то не воспитана в человеке. Неблагородно к кедрачу относятся* (Парабель, 1985); *Вот у нас, видишь, кедро'вник-то весь поуничтожили, леспромхоз. Что-то такое* ***государственный министр лесной промышленности*** *как-то проглядел. Скомандовал, чтоб кедра'ч у речек ува'ливали. Вот щас нача'ли разводить его у нас же, лесхоз нача'л. Разводит. Сколько-то гектар они уже. Вот насажда'ют и ро'стют тут вот. Там участие принимают вот школа, ученики, вот в посадке-то, насажде'нии кедрача'.* (Парабель, 1985); *Сосновый лес был такой. Раньше кедр не трогали. В* ***газетах борьба идёт,*** *да всё ровно срублены, брошены. Лежит. Какая-то бесхозяйственность. Раньше лес не рубили. Нет, конечно, рубили. Зря лесину-то срубить — как-то не было. Раньше как-то к лесу бережно относились. А теперь с пелёнок просят ветку. Мать <щиплет> ему. Конечно, варвар вырастет. А так, дак все деревья общипаны. В Сочи*

ездила: сады, лес такой красивый, а у нас всё ломать, да бросать. (Колпашево, 1983). В таких фрагментах текста чувствуется влияние официального дискурса, о чём свидетельствуют вкрапления книжной лексики (*бережливость, неблагородно, насаждение, бесхозяйственность, варвар*), речевые штампы (*борьба идёт; участие принимают*). Показательны и ссылки на авторитеты (*учёные, государственный министр лесной промышленности*), указание источника информации (*прочитаешь,* в *газетах*). Подобные явления подробно проанализированы И. В. Тубаловой [Тубалова 2016].

Рассуждения об экологии появляются также в контексте деятельности государства — прежде всего, определённых запретов: *Щас ведь это запрещённо — бот* [*орудие кедрового промысла*]. *Потому что когда бьют* [*шишку*], *аж шкура слетает, оббивают. Значит, она иссыхает и погибает.* (Парабель, 1985); *На охоте, значит, всяких это тоже мелочёвку. Ну бывает и соболей, и барсуков. И слушай, ты тут моуо не пиши, а то чтоб меня не посадили ещё, а то меня посадят, это скажут,* ***браконьер какой-то, а? Ты чё вычеркни там чё-нибудь.*** *А то вишь вот барсуков, например, ловить нельзя, а я их ловил, а ты записала уже, да?* [*Смеётся*] (Парабель, 2012); *Самоло'вы — это тоже запрещённый вид лова. Категорически запрещён. Им только промышляют только на Оби. Да, на больших реках. Добывают как например, стерлядку, вот это на самоловы. Ну это очень редко у нас, кто занимается,* ***за это очень строго, очень строго. За одну рыбку штрафуют до семидесяти пяти, до ста рублей*** (Нарым, 1980). Лес, река кормят сельских жителей, поэтому запрет пользоваться дарами природы воспринимается неоднозначно, особенно в ситуации, когда вырубка леса и ловля рыбы осуществляются государством в несравненно бо'льших масштабах.

Гораздо реже упоминается о природоохранной деятельности: *Щас ещё рано грибы-ягоды собирать. В* ***зелёной зоне*** *собираем, там специально ни лес, ничё не вырубают* (Белый Яр, 1980).

Выводы: тексты Томского диалектного корпуса отражают колоссальные исторические изменения XX в., повлиявшие как на материальные аспекты жизни крестьян, так и на их мироощущение. Трансформации обнаруживаются, в частности, при анализе тем «Техника», «Экология», выделенных в корпусе. Тематическая разметка позволяет проводить анализ связных фрагментов диалектного дискурса, выявляя определённые мотивы (повторяющиеся смысловые фрагменты), свойственные описанию данной темы в диалектной коммуникации.

Для тем «Экология» и «Техника», в силу новизны соответствующей проблематики, важна миромоделирующая оппозиция «Раньше/Теперь». Тема «Техника» раскрывается также через мотивы её первого появления и связанных с этим эмоций и суеверий. Рассуждения на экологические темы связаны как с непосредственными наблюдениями, личным опытом сельских жителей, так и с деятельностью государства и СМИ. Обнаруживается взаимосвязь исследуемых тем. Нередко технический прогресс воспринимается в диалектных текстах как причина природных изменений.

Корпус также даёт возможность выявить взаимосвязь экстралингвистических факторов и количества тематических фрагментов. Так, при анализе темы «Техника» обнаруживается влияние темпорального фактора, темы «Экология» — гендерного.

Литература

1. *Земичева С. С.* (2018), Взаимосвязь тематики диалектного текста и пола говорящего (на материале Томского диалектного корпуса), Актуальные проблемы и перспективы русистики (Материалы по итогам Международной конференции русистов в Барселонском университете), Барселона, с. 491–500.
2. *Никитина С. Е.* (1999) Убери магнитофон!, Мир звучащий и молчащий: Семиотика звука и речи в традиционной культуре славян. М., с. 325–329.
3. Корпус. Демо-версия. URL http://losl.tsu.ru/corpus/demo (дата обращения 22.02.2019).
4. *Тубалова И. В.* (2016) Полифонический текст в устных личностно-ориентированных дискурсах. Томск.

References

1. Nikitina S. E. (1999) Uberi magnitofon! [Take away the tape recorder!], Mir zvuchashchij i molchashchij. Semiotika zvuka i rechi v tradicionnoj kulture slavyan [World Of Sounds and Silence. Semiotics of Sound and Speech in the Traditional Culture of the Slavs], Moscow, pp. 325–329.
2. Corpus demo-versiya. [Corpus. Demo version]. URL http://losl.tsu.ru/corpus/demo (request date 02/02/2019).
3. Tubalova I. V. (2016) Polifonicheskij tekst v ustnyh lichnostno-orientirovannyh diskursah [Polyphonic text in oral personality-oriented discourses]. Tomsk.
4. Zemicheva S. S. (2018) Vzaimosvyaz tematiki dialektnogo teksta i pola govoryashchego na materiale Tomskogo dialektnogo korpusa [The correlation between the topic of a dialect text and the speaker's gender (based on the materials of Tomsk dialect corpus)], Aktualnye problemy i perspektivy rusistiki. Materialy po itogam Mezhdunarodnoj konferencii rusistov v Barselonskom universitete [Current Trends and Future

Perspectives in Russian Studies. Proceedings of the International Conference on Russian Studies at the University of Barcelona], Barcelona, pp. 491–500.

Земичева Светлана Сергеевна
Томский государственный университет (Россия)
Zemicheva Svetlana
Tomsk State University (Russia)
E-mail: optysmith@gmail.com

А. А. Крижановский, Н. Б. Крижановская, И. П. Новак

A. A. Krizhanovsky, N. B. Krizhanovsky, I. P. Novak

ПРЕДСТАВЛЕНИЕ ДИАЛЕКТОВ В ОТКРЫТОМ КОРПУСЕ ВЕПССКОГО И КАРЕЛЬСКОГО ЯЗЫКОВ (ВЕПКАР)[1]

DIALECTS IN OPEN CORPUS OF VEPS AND KARELIAN LANGUAGES (VEPKAR)

Аннотация. В статье описывается организация работы с диалектами карельского языка в Открытом корпусе вепсского и карельского языков (http://dictorpus.krc.karelia.ru/). Обсуждаются вопросы, связанные с необходимостью обеспечения технической поддержки словарной части корпуса по заполнению и указанию связей между фонетическими вариантами лемм разных диалектов карельского языка. Показаны различные способы пополнения словаря словоформами.
Ключевые слова. карельский язык, корпус, словарь.

Abstract. The article describes the work with dialects of the Karelian language in the Open Corpus of Veps and Karelian languages (http://dictorpus.krc.karelia.ru/). The issues related to the variety of Karelian dialects are discussed. Several ways of the dictionary extending with word forms and the connection of phonetic variants of lemmas are shown.
Keywords. Karelian language, corpus, dictionary.

1. Введение

Открытый корпус вепсского и карельского языков (ВепКар)[2] содержит словари и тексты прибалтийско-финских языков народов Карелии.

Проект ВепКар является продолжением работ по Корпусу вепсского языка (http://vepsian.krc.karelia.ru/), который разрабатывался в 2009-2016 годы сотрудниками Карельского научного центра РАН. В 2016 году проект расширился тремя наречиями карельского языка и пополнился новым функционалом [Зайцева и др. 2017]. На 2019 год в корпусе более 21 тысячи словарных статей и почти 2 тысячи текстов на 27 диалектах вепсского и карельского языков. В проект включены еще 18 диалектов карельского языка, но они пока не отражены в корпусе.

Сотрудники Карельского научного центра РАН заполняют словарь и добавляют тексты в Корпус вепсского и карельского языков. Корпус карельского языка включает три подкорпуса в соответствии с делени-

[1] Работа выполнена при поддержке РФФИ (проект 18-012-00117).

[2] См. http://dictorpus.krc.karelia.ru/

ем языка на три наречия: собственно карельское и ливвиковское, обладающие собственными младописьменными формами, а также людиковское, над нормированным вариантом которого ведется работа в настоящее время (рис. 1).

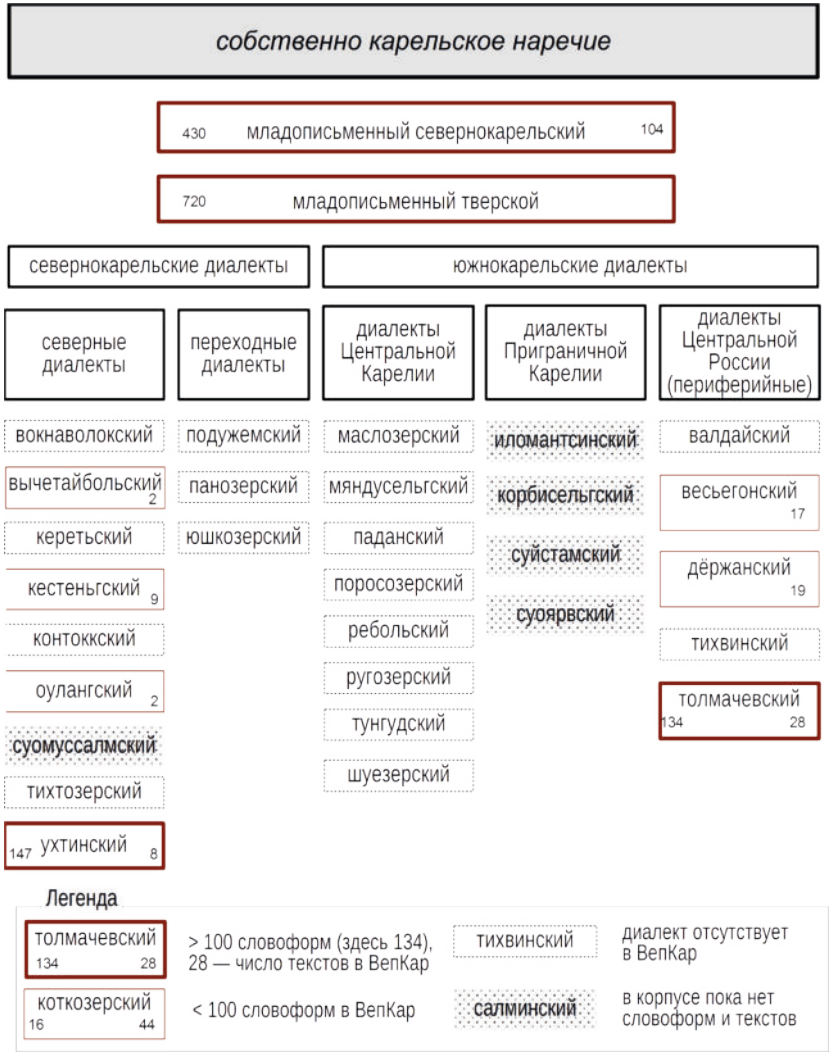


Рис. 1.а. **Диалекты собственно карельского наречия карельского языка, число словоформ (слева) и число текстов (справа) в корпусе ВепКар**

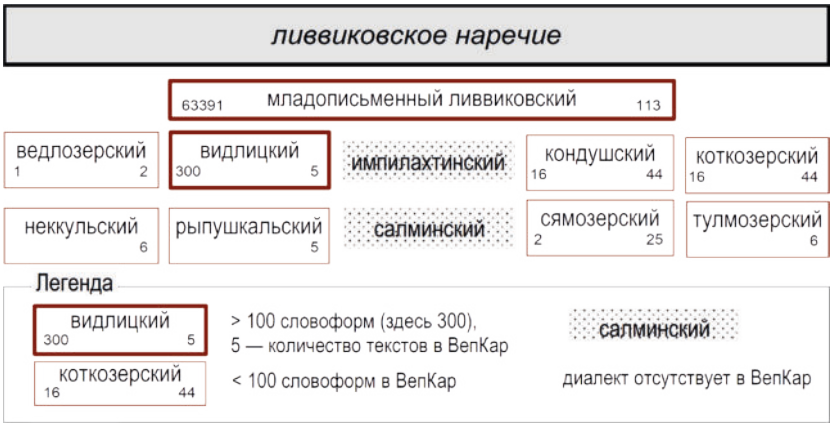


Рис. 1б. **Диалекты ливвиковского наречия карельского языка, число словоформ (слева) и число текстов (справа) в корпусе ВепКар**

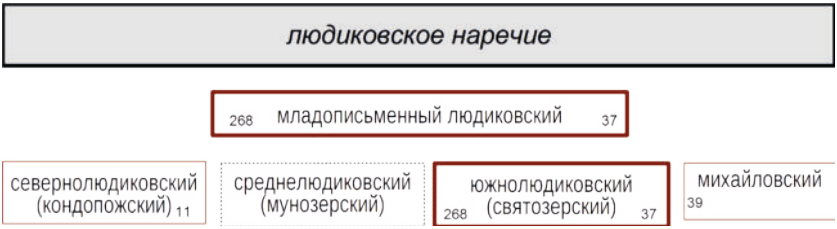


Рис. 1в. **Диалекты людиковского наречия карельского языка, число словоформ (слева) и число текстов (справа) в корпусе ВепКар**

Для отображения многообразия диалектов и говоров карельского языка появилась возможность при редактировании леммы указывать словоформы по отдельному диалекту. Для установления связей между одинаковыми лексемами в разных диалектах одного наречия был предложен новый вид отношений между леммами — фонетический вариант.

2. О диалектах карельского языка

В корпусе ВепКар три карельских наречия даны отдельно, а не представлены единым языком по ряду причин. Формирование этих наречий произошло в процессе разновременных волн переселений карелов, носителей древнекарельских диалектов, на новые обширные

территории и контактов с их автохтонным населением. Таким образом, разделение языка на наречия явилось результатом влияния со стороны как родственных вепсского и финского языков, так и иноязычной системы русского языка, которого невозможно было избежать в условиях многовекового экономического и культурного взаимодействия карельского населения с русским. Основному воздействию оказались подвержены лексическая и фонетическая системы карельских наречий. Развившиеся в них существенные различия диктуют необходимость использования в корпусе ВепКар трех разных словарей карельского языка (данные на февраль 2019 года):

- ливвиковское наречие (17 тысяч лемм),
- людиковское наречие (450 лемм),
- собственно карельское наречие (130 лемм).

Трудности разработки корпуса ВепКар вызваны необходимостью поддерживать многообразие диалектов и говоров карельского языка (рис. 1), сохранить в корпусе и представить читателю наличие взаимосвязей между словами разных диалектов. В частности, в корпусе ВепКар реализованы следующие возможности:

- поиск в словаре по языкам и по наречиям;
- добавление словоформы для отдельного диалекта в процессе редактирования леммы;
- использование «фонетического варианта» для связи одной и той же леммы, имеющей в диалектах разное написание.

Технически «фонетический вариант» реализован в корпусе следующим образом. Корпус ВепКар имеет тезаурус (http://dictorpus.krc.karelia.ru/ru/dict/lemma/relation). К восьми семантическим отношениям был добавлен «фонетический вариант» для связывания лексем, имеющих отличия в звучании в разных наречиях и диалектах карельского языка.

Примеры: kaži (севернокарельский, ливвиковский, людиковский) / kazi (с.к.) / kasi (с.к.) ‘кошка’ [NOM.SG], kodi (с.к., ливв., люд.) / koti (с.к.) ‘дом’ [NOM.SG], aiga (с.к.) / aika (с.к.) / aigu (ливв.) / aige (люд.) / aig (люд.) ‘время’ [NOM.SG], andua / andoa (с.к., ливв.) / antua (с.к.) / andada (люд.) / antta (люд.) ‘давать’ [INF], jogi / d’ogi (с.к., ливв.) / joki (с.к.) / d’og (люд.) ‘река’ [NOM.SG].

Схема наречий и диалектов карельского языка (рис. 1) построена по данным ВепКар, использована классификация от 2001 года, пред-

ложенная П. М. Зайковым [Зайков, 2001, с. 27], и подробная таблица 2019 года [Новак и др., 2019].

В российском прибалтийско-финском языкознании в карельском языке принято выделять три наречия, каждое из которых распадается на диалекты. Традиционная классификация диалектов карельского языка, основанная на административном принципе (волостное деление территории Карелии начала XX в.), вызывает целый ряд вопросов. В имеющихся вариантах классификации количество диалектов варьируется от 35 до 45. При этом между выделенными диалектами далеко не всегда удается обнаружить наличие каких-либо существенных отличий, а изоглоссы диалектных явлений могут не совпадать с волостными границами. В традиционной классификации, таким образом, речь идет скорее о группах говоров, а проблема диалектного членения оставлена нерешенной. Диалектные материалы корпуса ВепКар в совокупности с его техническими возможностями могут быть использованы в качестве базы для решения этой проблемы.

Корпус на настоящий момент представлен 29 собственно карельскими, 10 ливвиковскими и 4 людиковскими диалектами из традиционной классификации. Всего получается 43 диалекта карельского языка и четыре младописьменные формы (севернокарельский, тверской, ливвиковский и людиковский).

3. Запросы в поисковой системе корпуса ВепКар

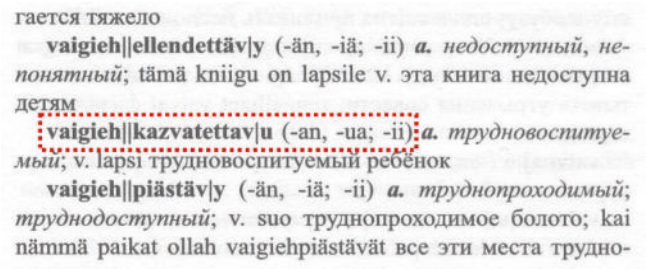
При формировании запроса можно уточнить язык, диалект (выбрать вариант из списка), либо искать по всему словарю. При поиске по лемме, по словоформе, толкованию можно указать шаблон. Например, *ka%i*.

В диалектных текстах часто присутствует разнообразное смягчение речи с помощью символа ', поэтому в системе помимо словарного написания словоформы, хранится запись для поиска без смягчений. Это не только расширяет поиск по словарю через поисковую форму, но и упрощает автоматическую разметку текстов.

4. Словарь корпуса ВепКар и генерация словоформ

По числу словоформ в словаре корпуса ВепКар лидируют младописьменный вепсский язык (28874 словоформы) и младописьменный ливвиковский (63391)[3].

[3] http://dictorpus.krc.karelia.ru/ru/dict/dialect?limit_num=45



гается тяжело

vaigieh||ellendettäv|y (-än, -iä; -ii) ***a.*** *недоступный, непонятный*; tämä kniigu on lapsile v. эта книга недоступна детям

vaigieh||kazvatettav|u (-an, -ua; -ii) ***a.*** *трудновоспитуемый*; v. lapsi трудновоспитуемый ребёнок

vaigieh||piästäv|y (-än, -iä; -ii) ***a.*** *труднопроходимый; труднодоступный*; v. suo труднопроходимое болото; kai nämmä paikat ollah vaigiehpiästävät все эти места трудно-

Рис. 2. **Фрагмент страницы книги «Большой карельско-русский словарь (ливвиковское наречие)», используемый для генерации словоформ в корпусе ВепКар, см. словарную статью для сложного прилагательного vaigiehkazvatettavu 'трудновоспитуемый', содержащую автоматически сгенерированные словоформы**

vaigiehkazvatettavu

язык: карельский: ливвиковское наречие

часть речи: прилагательное

1 значение

- **русский:** трудновоспитуемый

словоформы

No	грамматические признаки	Младописьменный ливвиковский
	Единственное число	
1.	номинатив	vaigiehkazvatettavu
2.	генитив	vaigiehkazvatettavan
3.	партитив	vaigiehkazvatettavua
	Множественное число	
4.	партитив	vaigiehkazvatettavii

Рис. 3. **Словарная статья для сложного прилагательного vaigiehkazvatettavu[4] в корпусе ВепКар содержит список словоформ (выделен пунктиром), сгенерированных по окончаниям -an, -ua, -ii, представленным в «Большом карельско-русском словаре (ливвиковское наречие)»**

[4] См. http://dictorpus.krc.karelia.ru/ru/dict/lemma/21943

Лучшая проработка вепсской части словаря и вепсских текстов объясняется тем, что корпус ВепКар, разрабатываемый с 2016 года, включил в себя и стал продолжением проекта "Корпус вепсского языка", начало которому было положено в 2009 году [Зайцева, 2017]. Младописьменный ливвиковский занимает сейчас лидирующее положение в корпусе по числу лемм и словоформ, поскольку в 2018 году в словарь корпуса почти в полном объёме был добавлен лексический материал издания Бойко Т. П. (Большой карельско-русский словарь (ливвиковское наречие). Петрозаводск, 2016), что составило около 17 тысяч лемм и 63 тысячи словоформ (рис. 2).

Для ускорения процесса ввода ливвиковских слов были добавлены новые модули автоматического разбора текста словарной статьи. Если раньше нужно было сначала заполнить отдельно поля формы создания леммы, сохранить, потом открыть форму редактирования словоформы, вручную внести четыре словоформы из словарной статьи, то теперь достаточно в поле леммы вставить фрагмент словарной статьи вместе с окончаниями (например, "**vaigieh||kazvatettav|u** (-an, -ua; -ii)") (рис. 2) и система автоматически разбирает текст, выделяет лемму, убирает разделители |, затем по основе леммы и окончаниям формируются словоформы (рис. 3) и записываются в базу данных корпуса.

Заключение

В статье описана организация работы с диалектами в корпусе ВепКар, в том числе указание связей между словами разных диалектов карельского языка. Показаны способы полуавтоматического пополнения словаря словоформами. Представлена схема наречий и диалектов карельского языка, построенная на основе данных корпуса ВепКар и ряда источников. Традиционно в языке выделяют три наречия, каждое из которых включает ряд диалектов: 29 собственно карельских, 10 ливвиковских и 4 людиковских. Таким образом, карельский язык представлен 43 диалектами, а также четырьмя младописьменными формами (севернокарельской, тверской, ливвиковской и людиковской).

Проблема диалектного членения карельского языка до настоящего момента не решена. Современная классификация карельских диалектов основана на административном принципе, а не на лингвисических критериях. И именно диалектные материалы корпуса и его поисковые возможности, значительно ускоряющие процесс работы с ними, обеспечат языковедов данными для численных и теоретических выводов о классификации диалектов карельского языка.

Литература

1. *Зайков П. М.* (2001), Глагол в карельском языке. Петрозаводск: ПетрГУ.
2. *Зайцева Н. Г., Крижановский А. А., Крижановская Н. Б., Пеллинен Н. А., Родионова А. П.* (2017), Открытый корпус вепсского и карельского языков (ВепКар): предварительный отбор материалов и словарная часть системы // Труды международной конференции «Корпусная лингвистика — 2017», СПб., с. 172–177.
3. *Новак И., Пенттонен М., Руусканен А., Сиилин Л.* (2019), Карельский язык в грамматиках. Сравнительное исследование фонетической и морфологической систем. Петрозаводск: КарНЦ РАН, с. 22.

References

1. *Zaikov P. M.* (2001), Glagol v karel'skom yazyke [Verb in Karelian language]. Petrozavodsk: PetrSU.
2. *Zaitseva N. G., Krizhanovsky A. A., Krizhanovskaya N. B., Pellinen N. A., Rodionova A. P.* (2017), Otkrytyy korpus vepsskogo i karel'skogo yazykov (VepKar): predvaritel'nyy otbor materialov i slovarnaya chast' sistemy [Open corpus of Veps and Karelian languages (VepKar): preliminary selection of materials and dictionary of the system] // International scientific conference "Corpus linguistics". Saint Petersburg, pp. 172-177.
3. *Novak I., Penttonen M., Ruuskanen A., Siilin L.* (2019), Karel'skiy yazyk v grammatikakh. Sravnitel'noe issledovanie foneticheskoy I morfologicheskoy sistem — Petrozavodsk: KarRC RAS, p. 22.

Крижановский Андрей Анатольевич
Институт прикладных математических исследований
Карельского научного центра РАН
Krizhanovsky Andrew
Institute of Applied Matthematical Research, Karelian Research Centre, RAS
E-mail: andrew.krizhanovsky@gmail.com

Крижановская Наталья Борисовна
Институт прикладных математических исследований
Карельского научного центра РАН
Krizhanovskaya Natalia
Institute of Applied Matthematical Research, Karelian Research Centre, RAS
E-mail: nataly.krizhanovsky@gmail.com

Новак Ирина Петровна
Институт языка, литературы и истории Карельского научного центра РАН
Novak Irina
Institute of Linguistics, Literature and History, Karelian Research Centre, RAS
E-mail: bel.irina@rambler.ru

А. А. Лебедев, А. А. Рогов, К. А. Кулаков, Н. Д. Москин
A. A. Lebedev, A. A. Rogov, K. A. Kulakov, N. D. Moskin

К ПРОБЛЕМЕ СОЗДАНИЯ РАЗМЕЧЕННЫХ КОРПУСОВ ТЕКСТОВ В ГРАФИКЕ XIX ВЕКА[1]

TO THE PROBLEM OF CREATING OF MARKED CORPUSES OF TEXTS IN THE GRAPHICS OF THE XIX CENTURY

Аннотация. В статье рассмотрены проблемы, которые возникают при разработке корпуса русских публицистических текстов второй половины XIX века СМАЛТ. Первая связана с большой орфографической вариативностью текстов (наблюдается различное написание одних и тех же слов). Вторая проблема связана с обработкой составных словоформ (они могут быть по-разному проанализированы с точки зрения их размещения в корпусе).
Ключевые слова. СМАЛТ, атрибуция текстов, разметка текстов, Ф. М. Достоевский.

Abstract. This article discusses the problems that arise during the development of the corpus of Russian publicistic texts of the second half of the XIX century SMALT. The first is associated with a large orthographic variation of texts (there is a different writing of the same words). The second problem is related to the processing of compound word forms (they can be analyzed differently from the point of view of their placement in the corpus).
Keywords. SMALT, text attribution, text markup, F. M. Dostoevsky.

Одна из отличительных особенностей корпуса СМАЛТ — это его нацеленность на оригинальные публицистические тексты, написанные во второй половине XIX века (шестидесятые-семидесятые годы), что противопоставляет данный корпус большинству других из числа современных русскоязычных лингвистических корпусов (тексты которых ориентированы на современную орфографию) [Котов 2014]. Подобного рода подход вызывает очевидные затруднения, связанные в первую очередь с проблемами автоматической обработки таких текстов.

Основу корпуса СМАЛТ составляют тексты статей журнала «Время» (1861–1863 гг.), представленные в дореформенной графике. При этом часть представленных в корпусе публицистических произведений имеет установленного автора (Ф. М. Достоевский, А. А. Григорьев, М. И. Владиславлев и ряд других), в то время как авторство других текстов однозначно не определено (такие тексты имеют пометку Dubia) [Захаров 2000; Котов 2012].

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-012-90026.

Тексты в корпусе СМАЛТ хранятся в дореформенной графике, но при этом пользователю предлагается и современное их написание. Тексты второй половины XIX века демонстрируют большую орфографическую вариативность, что затрудняет подготовку материала для подобного корпуса (в том числе и в автоматическом режиме). К примеру, в статье «Вопросъ объ университетах» мы можем встретить два разных написания слова «профессор» — через одну и через две «с»:

1) …*по поводу ея* профессоръ *Костомаровъ написалъ свою* статью…
   Корпорація ***профессоровъ*** *сохраняется во всей силѣ…*
2) …*явилось третье* мнѣніе г професора *Стасюлевича…*
   Что обязывало професора *имѣть въ виду* воспитательно-учебныя цѣли

Безусловно, учет подобных особенностей-разночтений полезен с точки зрения целого ряда аспектов: корпус СМАЛТ позволяет учесть изменение орфографических норм, а также проследить за общей логикой развития грамматической системы русского языка XIX века.

Для решения проблемы многозначного описания в программной реализации корпуса СМАЛТ выполнено разделение написания слов и их словоформ. Таким образом, разные написания слов могут иметь одну общую словоформу. Кроме этого, для каждого слова в словоформе приведено современное написание, что позволяет выполнять поисковые запросы без использования дореформенной графики.

Один из проблемных и сложных для решения вопросов, связанных с морфологической разметкой слов в лингвистическом корпусе — обработка составных словоформ. Рассмотрим несколько примеров, которые могут быть по-разному проанализированы с точки зрения их размещения и обработки в корпусе:

1) *Въ первомъ* нумерѣ газеты *День* **помѣщена была** статья *объ* университетахъ покойнаго *Хомякова…*
2) …умноженіе *у насъ въ* Россіи воспитательныхъ заведеній *съ болѣе широкимъ* преподаваніемъ *чѣмъ въ* гимназіи *принесетъ намъ* пользу…
3) … *на эту* тему *можно написать* ***тридцать пять*** печатныхъ листовъ…

В ранних разборах текстов Ф. М. Достоевского данная проблема для приведенных случаев решалась следующим образом: выделенные

компоненты разбирались как одно слово, которому приписывались все грамматические категории.

1) *Въ первомъ* нумерѣ газеты *День* **помѣщена была** статья *объ* университетахъ покойнаго *Хомякова*

   Слово: помѣщена была (id=55056)
   Начальная форма: ПОМѢЩЕННЫЙ
   Часть речи: Причастие

2) …умноженіе *у насъ въ* Россіи воспитательныхъ заведеній *съ болѣе широкимъ* преподаваніемъ *чѣмъ въ* гимназіи *принесетъ намъ* пользу…

   Слово: болѣе широкимъ (id=55101)
   Начальная форма: ШИРОКІЙ
   Часть речи: Прилагательное

3) *на эту* тему *можно написать* ***тридцать пять*** печатныхъ листовъ

   Слово: тридцать пять (id=55392)
   Начальная форма: ТРИДЦАТЬ ПЯТЬ
   Часть речи: Числительное

В поздних вариантах разбора, которые легли в основу корпуса СМАЛТ, такие же контексты разбирались пословно, и у пользователя корпуса есть возможность посмотреть грамматические категории каждого из слов. На текущий момент для некоторых текстов в корпусе СМАЛТ присутствует морфологическая разметка в двух вариантах (условно называемых «старый» и «новый»), несколько различающаяся по набору частей речи и по грамматическим категориям, однако вне зависимости от выбранного варианта элементы в данных контекстах разбираются как два отдельных слова.

Рассмотрим разборы со старым набором атрибутов:

1) *Въ первомъ* нумерѣ газеты *День* **помѣщена была** статья *объ* университетахъ покойнаго *Хомякова*
   Слово: помѣщена (id=12210)

   Начальная форма: помѣщенъ
   Часть речи: Причастие

Слово: была (id=369)
Начальная форма: быть
Часть речи: Глагол
Отвлеченный глагол-связка: Да

2) …умноженіе *у насъ въ* Россіи воспитательныхъ заведеній *съ болѣе широкимъ* преподаваніемъ *чѣмъ въ* гимназіи *принесетъ намъ* пользу…

Слово: болѣе (id=3763)
Начальная форма: болѣе
Часть речи: Наречие
Вспомогательная часть сложной степени сравнения: Да

Слово: широкимъ (id=12611)
Начальная форма: (более) широкій
Часть речи: Прилагательное

3) *на эту* тему *можно написать* ***тридцать пять*** печатныхъ листовъ

Слово: тридцать (id=14819)
Начальная форма: тридцать
Часть речи: Числительное
Разряды по значению: Количественное
По способу образования: Часть составного

Слово: пять (id=14820)
Начальная форма: пять
Часть речи: Числительное
Разряды по значению: Количественное
По способу образования: Часть составного

Рассмотрим разборы с новым набором атрибутов:

1) *Въ первомъ* нумерѣ газеты *День* **помѣщена была** статья *объ* университетахъ покойнаго *Хомякова*

Слово: помѣщена (id=10571)
Начальная форма: помѣщенный
Часть речи: Причастие

Слово: была (id=10572)
Начальная форма: быть
Часть речи: Глагол

2) …умноженіе *у насъ въ* Россіи воспитательныхъ заведеній *съ болѣе широкимъ* преподаваніемъ *чѣмъ въ* гимназіи *принесетъ намъ* пользу…

   Слово: болѣе (id=5341)
   Начальная форма: болѣе
   Часть речи: Наречие
   Степень сравнения: Компонент сравнительной аналитической степени сравнения

   Слово: широкимъ (id=10889)
   Начальная форма: широкій
   Часть речи: Прилагательное
   Форма: Полная
   Степень сравнения: Компонент сравнительной аналитической степени сравнения

3) *на эту* тему *можно написать* ***тридцать пять*** печатныхъ листовъ

   Слово: тридцать (id=12656)
   Начальная форма: тридцать
   Современное написание: тридцать
   Часть речи: Числительное

   Слово: пять (id=1116)
   Начальная форма: пять
   Современное написание: пять
   Часть речи: Числительное

Второй подход, при котором каждая подобная словоформа разбирается по отдельности, видится более удобным с точки зрения хранения и поиска данных. В то же время, разумным представляется указание в ходе разбора на то, что некоторое слово является, к примеру, частью составного союза:

мы и сами не предлагаемъ никакихъ реформъ ***да и*** *о предложенныхъ* реформахъ *не очень распространимся*

Слово: да (id=2179)
Начальная форма: да
Часть речи: Союз
По составу: Часть составного союза

Слово: и (id=2180)
Начальная форма: и
Часть речи: Союз
По составу: Часть составного союза

Каждый из представленных подходов имеет свои достоинства и недостатки и не может быть выбран в качестве основного. Таким образом, разумным решением является предоставление гибких инструментов анализа текстов. Программная реализация корпуса СМАЛТ позволяет выполнять морфологическую разметку текста в удобном для исследователя формате. Также планируется внедрить возможность использования авторской морфологической разметки и последующего сопоставления морфологических разборов.

Литература

1. *Захаров В. Н., Леонтьев А. А., Рогов А. А., Сидоров Ю. В.* (2000), Программная система поддержки атрибуции текстов статей Ф. М. Достоевского. Труды Петрозаводского государственного университета. Серия «Прикладная математика и информатика». Петрозаводск, № 9, с. 113–122.
2. *Котов А. А., Минеева З. И., Рогов А. А., Седов А. В.,* Сидоров Ю. В. (2014), Лингвистические корпусы. Петрозаводск.
3. *Котов А. А., Некрасов М. Ю., Седов А. В., Рогов А. А.* (2012), Информационная система для создания размеченных корпусов малой размерности // Ученые записки Петрозаводского государственного университета. Петрозаводск, № 8-1, с. 108–112.

References

1. *Zaharov V. N., Leontev A. A., Rogov A. A., Sidorov Y. V.* (2000), Programmnaya sistema podderzhki atribucii tekstov statej F. M. Dostoevskogo [Software support system for the attribution of texts of articles by F. M. Dostoevsky] // Trudy Petrozavodskogo gosudarstvennogo universiteta. Seriya «Prikladnaya Matematika i Informatika» [Proceedings of Petrozavodsk State University. Series «Applied Mathematics and Computer Science»]. Petrozavodsk, no. 9, pp. 113–122.
2. *Kotov A. A., Mineeva Z. I., Rogov A. A., Sedov A. V., Sidorov Y. V.* (2014), Lingvisticheskie korpusy [Linguistic Corpuses]. Petrozavodsk.
3. *Kotov A. A., Nekrasov M. Y., Sedov A. V., Rogov A. A.* (2012), Informacionnaya sistema dlya sozdaniya razmechennyh korpusov maloj razmernosti [Information system for creating small-sized marked corpuses] // Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta [Proceedings of Petrozavodsk State University]. Petrozavodsk, no. 8-1, pp. 108–112.

Лебедев Александр Александрович,
Петрозаводский государственный университет (Россия)
Lebedev Aleksandr Aleksandrovich,
Petrozavodsk State University (Russia)
E-mail: perevodchik88@yandex.ru

Рогов Александр Александрович,
Петрозаводский государственный университет (Россия)
Rogov Aleksandr Aleksandrovich,
Petrozavodsk State University (Russia)
E-mail: rogov@petrsu.ru

Кулаков Кирилл Александрович
Петрозаводский государственный университет (Россия)
Kulakov Kirill Aleksandrovich,
Petrozavodsk State University (Russia)
E-mail: kulakov@cs.petrsu.ru

Москин Николай Дмитриевич,
Петрозаводский государственный университет (Россия)
Moskin Nikolai Dmitrievich,
Petrozavodsk State University (Russia)
E-mail: moskin@petrsu.ru

И. С. Николаев
I. S. Nikolaev

КОРПУСНОЕ ИССЛЕДОВАНИЕ ТОПОНИМОВ В ИЖОРСКИХ НАРОДНЫХ ПЕСНЯХ

CORPUS STUDY OF TOPONYMS IN IZHORIAN FOLK SONGS

Аннотация. В статье обсуждается использование топонимов в ижорских эпических песнях по данным корпуса «Древние песни финского народа». На материале вариантов эпической песни на сюжет «Большой дуб» выявлены топонимы и соответствующие им географические объекты Ингерманландии и прилегающих территорий. Построена иерархия их значимости для исполнителей народных песен и носителей ижорского языка.
Ключевые слова. Топонимы, корпусное исследование, эпические песни, ижорские народные песни, Ингерманландия.

Abstract. Toponyms in the Izhorian epic songs are discussed in the article. The study is based on the corpus «Old songs of the Finnish People». Variants of the epic song «The Big Oak-tree» are investigated to list the toponyms and geographical objects of Ingermanland and adjacent territories. We built a hierarchy of toponyms' importance to the singers of the songs and to the speakers of the Izhorian language.
Keywords. Toponyms, corpus study, epic songs, Izhorian folk songs, Ingermanland.

Материалом нашего исследования являются ижорские народные песни, которые еще в XIX веке были широко распространены на территории исторической Ингерманландии на западе Ленинградской области в местах проживания ижоры — прибалтийско-финского народа — носителей ижорского языка.

Народные песни прибалтийско-финских народов России, Финляндии и Эстонии были записаны, собраны и опубликованы в течение XIX и XX веков фольклористами и лингвистами, сохранившими богатое устное наследие до того, как древние песенные традиции у некоторых из этих народов прервались в связи с сокращением носителей и исчезновением их языков.

В конце XX века на основе архивов в России, Финляндии и Эстонии были созданы цифровые коллекции, которые положили начало созданию корпусов текстов народных песен и дали возможность проводить разнообразные корпусные исследования.

В наших предыдущих докладах на конференции «Корпусная лингвистика» мы уже обсуждали детально эти архивы и корпусы, их воз-

можности и описывали проекты грамматической базы данных ижорского языка и экспертной системы по ижорскому языку (Николаев 2011, 2017).

В этом докладе мы хотели бы обратиться к исследованию топонимов, которые встречаются в текстах ижорских эпических песнях. Работа с корпусами народных песен и грамматической базой данных дает нам возможность выявить некоторое количество географических названий, которые отражают пространственные и географические представления ижорцев, которые песенная традиция связывает с определенными сюжетами. Обращение к эпическим песням связано с тем, что такого рода песни посвящены доисторическим мифологическим сюжетам, распространенным у всех прибалтийско-финских народов. Но у каждого народа оказывается свой набор топонимов, связанных с их территорией проживания и ближайшими ареалами, на которые распространяется их хозяйственно-экономическая деятельность.

Надо сказать, что общий набор топонимов, встречающихся в народных песнях совсем невелик. Намного больше число антропонимов. Однако даже у представителей одного народа, в данном случае — у ижорцев, можно наблюдать некоторую вариативность в использовании топонимов в разных вариантах одной и той же песни, записанных в разных районах Ингерманландии у носителей разных диалектов: нижнелужского, сойкинского и хэвасского.

В качестве примера, где встречается наибольшее разнообразие топонимов, мы возьмем сюжет «Большой дуб», который считается древнейшим космогоническим мифом у прибалтийско-финских народов. В варианте сюжета, распространенного в Ингерманландии (НПИ, 489), речь идет о большом дереве (чаще всего дубе), которое появляется благодаря хорошему удобрению земли (например, пивом) и вырастает выше неба, загораживая солнце. Поэтому его необходимо срочно срубить, и люди ищут дровосека, способного с этим справиться. Из дерева затем делают необходимые в хозяйстве постройки (например, баню) и разную утварь.

В корпусе «Древние песни финского народа» Общества финской литературы насчитывается более 300 записей песен на этот сюжет, из которых мы выбрали 50 текстов, записанных финским лингвистом В. Поркка в 1881-1883 годах (SKVR). Вариант этой эпической песни, записанный в 1968 году карельскими лингвистами Э. Киуру и Э. Кюльмясу, опубликован с русским переводом в сборнике «Народные песни Ингерманландии» (НПИ, 27; НПИ, 254).

Приведем в русском переводе короткий отрывок из этой эпической песни о поиске человека, который может срубить разросшийся большой дуб (в данном варианте сестра ищет брата — это контаминация с другим эпическим сюжетом):

Пошла искать я брата,
Искала в Суоми, искала на островах,
Обыскала обе (половины) Москвы,
Обыскала уголок Кронштадта,
Местечко маленькое в Питере.
(НПИ, 254)

Теперь рассмотрим несколько вариантов на ижорском языке. Топонимы выделены жирным шрифтом. Вначале приведем оригинал, перевод которого представлен выше:

Etsin **Soomet**, etsin **saaaret**,
Etsin **Moskovat** molloomat,
etsin kolkan **Kronstatiilt**,
piinen paikan **Petteriilt**
(НПИ 2. Soikkola, Voloitsa)

Далее — некоторые тексты из SKVR (в скобках — том, номер песни, место записи):

Etsin **Suomet**, etsin **saaret**,
45 **Moskovat** molemmin puolin,
Kahen puolen **Kaprioo**;
Turut etsin tunnustellen.
(SKVR III1 1162. Soikkola, Viistinä)

40 Etsin **saaret**, etsin **Suomet**,
Etsin **Turut** tunnustellen,
Viroin välit valkissellen,
Moskovan molemmin puolin,
Kahen puolin **Kaprioo**,
(SKVR III1 1163. Soikkola, Väärnoja)

Etsin **Suomet**, e[tsin] **saaret**,
Moskovat molommin puolin.
35 Kahen puolen **Kaprioo**,
Pienen **Petterin** selältä,
Kaijan Kaarassan selillä,
Uuen linnan uulitsoilla.
(SKVR III1 1164. Soikkola, Volotsa)

Etsin viikon vellojain,
Kuukavven Kalervuttain,
Etsin **Suomet**, etsin **Saaret**,
Etsin **Petterin** perukset,
Kahen puolen **Kaprioa**,
Moskovat molemmin puolin.
(SKVR IV2 1846. Hevaa, Koski)

Mäni velloni **Virroo**,
Kalervoni Kaarostaa,
Mokomani **Moskovaa**.
Etsin **Suomet**, etsin **saaret**,
Moskovan molemmin puolin,
Kahen puolen **Kapriota**,
Kahen **Kaprion** väliltä,
Harkkolan molemmat haarat,
(SKVR III1 613. Narvusi, Kurkola)

Mänin velloo etsimää:
Virot etsin velloani,
Virot etsin, vad valitsin,
Virot etsin viisin raksuin,
Vait vaites kaheksin raksuin.
(SKVR III1 615. Narvusi, Kulla)

Mänin velloo etsimmää:
Hulkuin **Suomet**, hulkuin **saaret**,
Moskovan molemmin puolin,
Kahen puolen **Kapriota**;
Vello oli **Venähen** maalla.
(SKVR III1 619. Narvusi, Pärspää)

Топонимы, выделенные в наших примерах, и их наличие в текстах эпических песен на трех ижорских диалектах сведены в Таблицу 1. Мы привели известные нам ижорские топонимы к номинативу единственного числа, так как в текстах они встречаются главным образом в генитиве единственного или множественного числа.

Небольшой объем материала и неравномерность распределения текстов по диалектам не позволяет нам получить достоверную статистическую информацию о топонимах в приведенных примерах, однако, их достаточно, на наш взгляд, чтобы можно было представить общую картину. Кроме того, приведенный отрывок сам по себе является часто повторяющейся частью текста, которую можно найти и в других

Таблица 1. **Топонимы и их представленность в ижорских диалектах по текстам эпических песен на сюжет «Большой дуб», где 1 обозначает наличие данного топонима, а 0 — его отсутствие в сойкинском, хэвасском и нижнелужском диалектах ижорского языка (по данным корпуса SKVR)**

Топоним (ижорский/русский)	Сойкинский	Хэвасский	Нижнелужский
Suomi '*Финляндия*'	1	1	1
saaret '*острова*'	1	1	1
Moskova '*Москва*'	1	1	1
Kaprio '*Копорье*'	1	1	1
Petteri '*Петербург*'	1	1	0
Viro '*Эстония*'	1	0	1
Turku '*Турку*'	1	0	0
Uuen linnan '*Новгород*'	1	0	0
Kronstatti '*Кронштадт*'	1	0	0
Harkkola '*Гарколово*'	0	0	1
Venähen '*Россия*'	0	0	1
Venähen '*Россия*'	0	0	1

эпических песнях, где встречаются те же топонимы. Мы эти примеры здесь приводить не будем.

В заключение мы можем сделать следующие выводы:

1. Представленные в текстах на всех трех диалектах топонимы, по-видимому, отражают известность и важность достаточно удаленных от Ингерманландии объектов: Финляндии и Москвы.

2. Апеллятив 'острова' нельзя соотнести с каким-то конкретным объектом, но его позиция во всех текстах всегда рядом с топонимом 'Финляндия' позволяет предположить, что речь идет об островах Финского залива в их совокупности, поэтому можно условно тоже считать его значимым топонимом.

3. Значительными региональными объектами являются Копорье, Петербург и Эстония. Копорье являлся одним из исторических центров Ингерманландии и приходским центром в лютеранской церковной жизни.

4. Исполнителям эпических песен и представителям ижоры были известны и исторические торговые центры: Турку и Новгород, хотя они, видимо, и не играли значительной роли.

5. Кронштадт появился в записях ижорских эпических песен только в XX веке. Топоним отражает новые приоритеты в значимых объектах Ингерманландии. В более ранних записях народных песен он не встречается.

6. Некоторые объекты, например, деревню Гарколово сложно объяснить в списке топонимов в эпических песнях. Деревня, насколько нам известно, не играла важной роли в экономической или политической жизни Ингерманландии. Она находится на Сойкинском полуострове, а записана в тексте на нижне-лужском диалекте. Можно предположить, что упомянута она случайно. В других текстах народных песен этот топоним не встречается.

7. Топоним *Venähen* 'Россия' (ген., ед. ч.) несложно объяснить с точки зрения его известности и значения, но его форма (генетив множественного числа) нехарактерна для ижорского языка и также больше нигде в текстах данного корпуса не встречается.

Литература

1. НПИ: Народные песни Ингерманландии. (1974) Ленинград.
2. *Николаев И. С.* (2011) Проблемы морфологического аннотирования корпуса ижорских народных песен. // Труды международной конференции «Корпусная лингвистика-2011», с. 266–269.
3. *Николаев И. С.* Корпус текстов на ижорском языке как основа лингвистической экспертной системы. // Труды международной конференции «Корпусная лингвистика-2017», с. 276–281.
4. SKVR: Suomen kansan vanhat runot (1908–1934) Helsinki. URL: skvr.fi (11.03.2019)

References

1. Folk songs of Ingermanland [Narodnyje pesni Ingermanlandii] (1974) Leningrad.
2. Nikolaev I. S. (2011) Problemy morfologicheskogo annotirovanija korpusa izhorskih narodnyh pesen [Problems of annotatiion of Izhorian folk songs corpus]. Trudy mezhdunarodnoj konferentsii «Korpusnaja lingvistika-2011» [Proceedings of International Conference «Corpus Linguistics-2011]. St.Petersburg.
3. Nikolaev I. S. (2017) Korpus tekstov na izhorskom jazyke kak osnova lingvisticheskoj ekspertnoj sistemy [Izhorian language text corpus as a basis of linguistic expert system]. Trudy mezhdunarodnoj konferentsii «Korpusnaja lingvistika-2017» [Proceedings of International Conference «Corpus Linguistics-2017]. St.Petersburg.
4. SKVR: Suomen kansan vanhat runot [Old songs of Finnish people] (1908–1934) Helsinki. URL: skvr.fi (11.03.2019)

Николаев Илья Сергеевич
Санкт-Петербургский государственный университет (Россия)
Ilya Nikolaev
St. Petersburg State University (Russia)
E-mail: i.s.nikolaev@spbu.ru

С. О. Савчук
S. O. Savchuk

ИНСТРУМЕНТАРИЙ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА В ДИАХРОНИЧЕСКИХ ИССЛЕДОВАНИЯХ[1]

ON USING THE RUSSIAN NATIONAL CORPUS IN DIACHRONIC STUDIES

Аннотация. Национальный корпус русского языка включает тексты разных речевых сфер и разных исторических эпох, поэтому на материале корпуса можно проводить диахронические и сопоставительные исследования. В статье на примере конкретной синтаксической конструкции будет показано, как корпусной инструментарий помогает выявить микроизменения в реализации правил координации подлежащего и сказуемого на протяжении разных периодов, проследить конкуренцию грамматических вариантов и определить тенденции развития.
Ключевые слова. Национальный корпус русского языка, диахронические исследования, грамматические варианты, согласование подлежащего и сказуемого, относительное предложение.

Abstract. The Russian National Corpus includes texts of different speech spheres and different historical epochs, so the material of the corpus can be used for diachronic and comparative studies. The article shows how the corpus tools help to identify micro-changes in the implementation of the rules of subject and predicate coordination for different periods, to trace the competition of grammatical variants and to determine the trends of development.
Keywords. The Russian National Corpus, diachronic studies, grammatical variants, subject and predicate coordination, relative clause.

Национальный корпус русского языка включает тексты разных речевых сфер и разных исторических эпох — от древнерусского периода до XXI века. Поэтому на материале корпуса можно проводить диахронические исследования. Поскольку бóльшую часть корпуса составляют тексты на литературном языке, эти исследования будут фиксировать прежде всего динамику литературной нормы. Однако в корпусе есть значительное количество некодифицированной речи — в составе корпусов устной речи и электронной коммуникации. В статье на конкретном примере будет показано, как корпус помогает выявить микроизменения в соотношении грамматических вариантов на протяжении разных периодов и определить динамику развития.

[1] Исследование выполнено при финансовой поддержке РФФИ, проект № 17-29-09154 и программы Президиума РАН «Памятники материальной и духовной культуры в современной информационной среде».

Правила, регулирующие согласование подлежащего со сказуемым, считаются одними из самых сложных, поскольку здесь переплетаются две тенденции — согласование по смыслу и согласование по форме.

Особый стилистический интерес вызывает координация сказуемого с подлежащим, выраженным некоторыми местоимениями. Согласно общему правилу при подлежащем-местоимении *кто, кто-то, кто-нибудь, кто-либо, кое-кто* глагол в сказуемом ставится в форме ед. ч., независимо от того, о каком количестве лиц идет речь: *Кто-нибудь знал об этом?* Однако если предложение с подлежащим-местоимением входит в состав сложного предложения в качестве придаточного и местоимение *кто* относится к другому местоимению, имеющему форму мн. ч.: *те, кто…, такие, кто…, вы* (*они*), *кто…, все, кто…*, в этом случае возможны варианты — глагол-сказуемое в форме ед. и мн. числа *…всем, кто* ***высунул*** *наружу головы и* ***слушают*** *мой прочный голос* (Фед.).

Эта проблема не является специфической именно для русского языка. Варианты согласования существуют и в других славянских и неславянских языках, при этом правила, регулирующие выбор варианта, могут отличаться от тех, которые действуют в русской грамматике. Приведем несколько примеров.

Белорусский: *Ён нешта расказваў, а* ***тыя, хто стаяў*** *вакол яго, гучна смяяліся. [Іван Чыгрынаў. Апраўданне крыві (1977)]. Он что-то рассказывал, а* ***те, кто стояли*** *вокруг, громко хохотали. [Иван Чигринов. Оправдание крови (пер. Инна Сергеева, 1978)].*

Польский: *Произошли многие изменения в жизни* ***тех, кто пострадал*** *от Воланда и его присных<…> [М. А. Булгаков. Мастер и Маргарита (1929–1940)]. W życiu* ***tych, którzy ucierpieli*** *przez Wolanda i jego kumpli <…> [Michaił Bułhakow. Mistrz i Małgorzata (пер. Irena Lewandowska, Witold Dąbrowski, 1969)].*

Болгарский: *Много промени станаха в живота на* ***онези, които бяха пострадали*** *от Воланд и неговите хора <…> [Михаил Булгаков. Майстора и Маргарита (пер. Лиляна Минкова, 1989)].*

Английский: *British Airways, Go With* ***Those Who Know****. Really good material even for* ***those who knows*** *fairly about tcp/ip [Интернет].*

Вопрос о согласовании форм сказуемого с подлежащим, выраженным местоимением *кто,* подробно рассматривается в работе [Ахапкина 2016]. Автор анализирует различные конструкции с подлежащим *кто*, сравнивает нормативные рекомендации с точки зрения

допустимости / недопустимости вариантов, а также реализацию этих рекомендаций в кодифицированном и некодифицированном узусе. В данной статье мы сосредоточимся на одной конструкции — с глаголом-сказуемым при подлежащем *кто* в относительном придаточном предложении, вершиной которого является местоимение во мн. ч.

Схематически правило можно выразить следующим образом:

Spro/Apro sg, ***кто*** **Vsg**	Пусть ответит **тот**, *кто* зна***ет*** ответ на этот вопрос
Spro/Apro pl, ***кто*** **Vsg/pl**	Пусть ответят **те**, *кто* зна*ет* ответ на этот вопрос Пусть ответят **те**, *кто* зна**ют** ответ на этот вопрос
Spro/Apro pl, ***которые*** **Vpl**	Пусть ответят **те**, котор**ые** зна**ют** ответ на этот вопрос

В случае, когда подлежащее придаточного предложения выражено местоимением *кто*, а вершина выражена формой мн. ч., возможны варианты выбора формы числа сказуемого. В этом случае, если предпочитается согласование по форме, сказуемое придаточного предложения будет выражено в форме ед. ч.; при согласовании по смыслу сказуемое в придаточном выражено формой множ. ч. В работе [Холодилова 2014] рассматривается поведение этой конструкции на материале НКРЯ и устанавливается статистическая зависимость выбора формы числа предикативного слова от формы числа и падежа вершины относительного предложения и позиционного типа придаточного [Холодилова 2014].

Нас будет интересовать соотношение вариантов ед. и множ. числа глагола в динамике на протяжении трех столетий, которое мы рассмотрим на материале корпусов текстов, относящихся к разным историческим периодам. Для этого проанализируем в подкорпусах текстов XIX в., 1-й пол. XX в. и 2-й пол. XX в. — нач. XXI в. конструкцию из трех компонентов вида:

1) Spro/Apro pl (,) 2) кто nom 3) V indic sg/pl (за исключением ‘есть’).

Частотность конструкции в речи в целом увеличивается, при этом обнаруживают рост оба варианта формы сказуемого — глагол в единственном числе (согласование по форме, предписываемое нормой) и глагол во множественном числе (согласование по смыслу). Для варианта с глаголом в форме ед. ч. это наращивание резко ускоряется во

Таблица 1. **Варианты согласования сказуемого с подлежащим *кто* по периодам**

	XVIII в. 5,2 млн		XIX в. 54,3 млн		1-я пол. XX в. 73 млн		XX в. —XXI в. 156,2 млн	
	всего	ipm	всего	ipm	всего	ipm	всего	ipm
V sg	18	3,5	914	16,8	3450	47,2	17431	111,6
V pl	0		58	1,1	253	3,5	742	4,8
Vsg/ Vpl	—		15,8	15,3	13,6	13,5	22,9	23,3

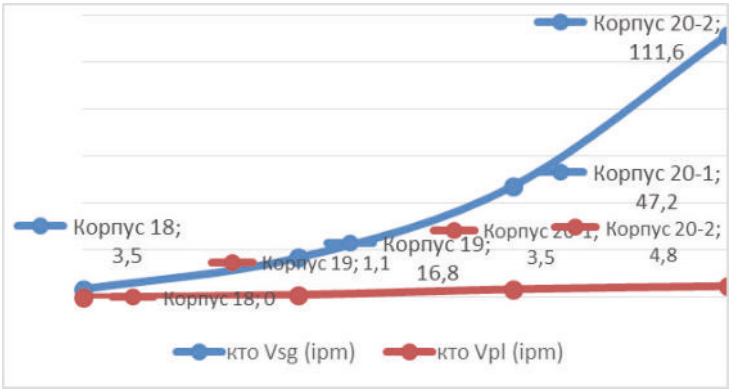


Рис. 1. **Соотношение вариантов согласования сказуемого с подлежащим (ipm) по периодам**

второй половине XX в., для варианта с глаголом в форме мн. ч. увеличение происходит более чем в 3 раза в первой половине XX в., а затем в текстах 2-й половины XX в. темп снижается.

Причина резкого роста частотности местоименно-соотносительных предложений с союзным словом *кто* во 2-й половине XX в. становится отчасти ясной при сравнении их с синонимичными придаточными с союзными словами *кой и который*.

> *Басня сия служит нравоучением для* ***тех****,* ***которые*** *воображают, что все создано для них. [Д. И. Фонвизин. Гордая свинья (1788)]. Басня сия надлежит до* ***тех****,* ***кои*** *знают все, кроме себя самих. [Д. И. Фонвизин. Бобр судьею (1788)]. Басня доказывает, сколь безрассудно смеяться над* ***тем****,* ***кто*** *имеет неверную жену. [Д. И. Фонвизин. Телята и олень (1788)].*

Как видим, в XVIII и XIX вв. в относительных придаточных предложениях в качестве относительного слова чаще использовались слова

кой и *который*, чем *кто*. К началу XX в. слово *кой* выходит из употребления, слово *который* в рассматриваемой позиции снижает употребительность, в то время как частотность конструкции с местоимением *кто*, напротив, растет.

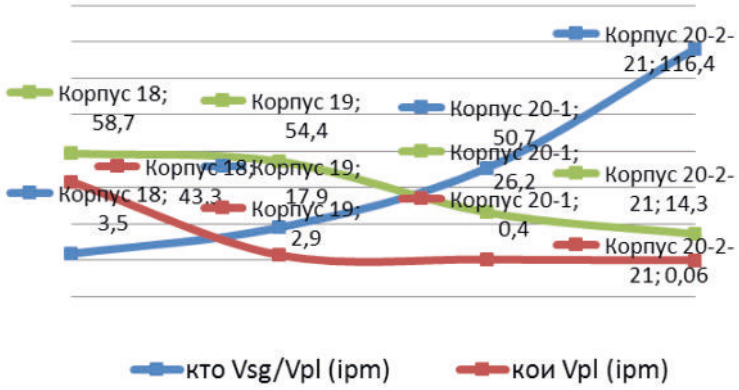


Рис. 2. **Относительные предложения с разными относительными словами (ipm) по периодам**

Зависит ли выбор вариантов от каких-то стилистических факторов? Отмечается, что координация по форме (выбор глагола в форме ед. ч.) предпочтительнее в книжных стилях, но в разговорном стиле в наше время все более закрепляется координация по смыслу (глагол в форме мн. ч.) и ее воспринимают писатели и журналисты [Голуб 1997, 380–381]. Проверим эти наблюдения на материале разных корпусов, содержащих современные тексты, относящиеся к разным функциональным сферам: публицистике (газетный и региональный газетный корпус), поэзии (поэтический корпус и корпус «наивной» поэзии) и устной речи (корпус устной речи, который помимо записей спонтанной разговорной речи и речи кино включает записи устной научной и политической речи). Соотношение вариантов изучаемых форм представлено на Рис. 3. Здесь же для сравнения приведено соотношение синонимических конструкций с союзными словами *кой* и *который*.

Как видим, приведенные выше наблюдения подтверждаются лишь отчасти. В устной речи действительно доля вариантов множ. числа значительно выше, чем в письменных текстах (15% от общего числа в сравнении с 2% в газетном корпусе). Однако газетный корпус

(включающий тексты начиная с 2000-х годов) демонстрирует отрыв вариантов с согласованием по форме, предписываемых нормативными рекомендациями, причем в региональном корпусе этот отрыв даже больше, в чем проявляется более строгое следование нормативным правилам.

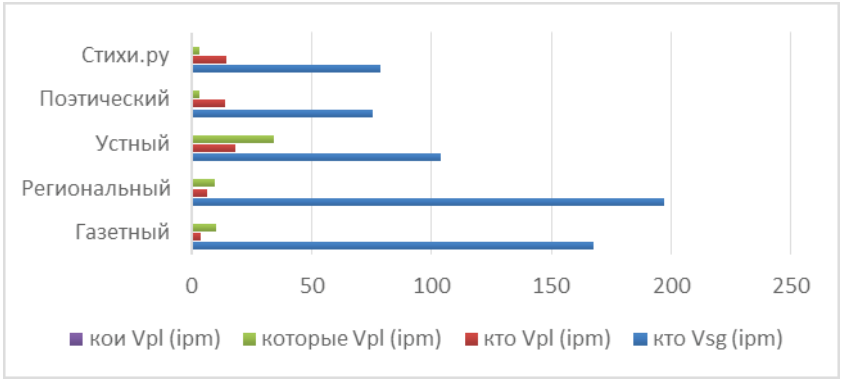


Рис. 3. **Соотношение вариантов согласования сказуемого с подлежащим (ipm) в разных корпусах**

Что касается поэзии, то обращают на себя внимание практически одинаковые невысокие частотные показатели соотношения вариантов как в профессиональной, так и в непрофессиональной, «наивной» поэзии (Стихи.ру). Это можно объяснить тем, что рассматриваемые конструкции, малочастотные в поэтических корпусах, нехарактерны для поэтической речи и являются по преимуществу прозаическими.

В целом данные разных корпусов свидетельствуют о том, что действие стилистического фактора проявляется в распределении вариантов согласования сказуемого с подлежащим, и это влияние может стать предметом специального исследования.

Литература

1. *Ахапкина Я. А.* (2016), Системные ошибки в нестандартных текстах: кто пришли: мн. число предиката при подлежащем «кто». Труды Института русского языка им. В. В. Виноградова. № 10. М., с. 25–36.
2. *Голуб И. Б.* (1997), Стилистика русского языка. М.
3. *Граудина Л. К., Ицкович В. А., Катлинская Л. П.* (2001), Грамматическая правильность русской речи. Стилистический словарь вариантов. 2-е изд. М.

4. *Розенталь Д. Э.* (2001), Справочник по русскому языку. Практическая стилистика. М.: Оникс 21 век: Мир и образование.
5. *Русская* грамматика. Т. 2. Синтаксис (1980). М., с. 244–245.
6. *Холодилова М. А.* (2017), Относительные придаточные. Материалы к корпусной грамматике русского языка. Выпуск II. Синтаксические конструкции и грамматические категории. В. А. Плунгян, Н. М. Стойнова (ред.). СПб., с. 205–279.

References

1. *Ahapkina Ya. A.* (2016), Sistemnye oshibki v nestandartnyh tekstah: kto prishli: mn. chislo predikata pri podlezhashchem "kto" [System mistakes in non-standard texts: *kto prishli* — plural form of predicate with subject KTO] // Trudy Instituta russkogo jazyka im. V. V. Vinogradova [Proceedings of the V. V. Vinogradov Russian Language Institute]. 10. 2016, pp. 25–36.
2. *Golub I. B.* (1997), Stilistika russkogo jazyka [Stylistics of the Russian language]. Moscow, Ajris-press Publ., 1997, pp. 380–381.
3. *Graudina L. K., Ickovich V. A., Katlinskaya L. P.* (2001) Grammaticheskaja pravil'nost' russkoj rechi. Stilisticheskij slovar' variantov [Grammatical correctness of the Russian speech. Stylistic dictionary of variants]. 2nd ed. Moscow, Nauka Publ., 2001.
4. *Holodilova M. A.* (2017), Materialy k korpusnoj grammatike russkogo jazyka. II. Sintaksicheskie konstrukcii i grammaticheskie kategorii [Materials for corpus grammar of the Russian language. Issue II. Syntactic constructions and grammatical categories]. V. A. Plungyan, N. M. Stojnova (eds). St.-Petersburg, Nestor-Istorija Publ., 2017, pp. 205–279.
5. *Rozental' D. E.* (2001), Spravochnik po russkomu jazyku. Prakticheskaja stilistika [Reference book on Russian language. Practical style]. Moscow, Oniks 21 vek: Mir i obrazovanie Publ.
6. *Russkaja grammatika* [Russian grammar. V. 2. Syntax]. Moscow, Nauka Publ., 1980. pp. 244–245.

Савчук Светлана Олеговна
Институт русского языка им. В. В. Виноградова РАН
Svetlana Savchuk
V. V. Vinogradov Russian Language Institute, RAS
E-mail: savsvetlana@mail.ru

РЕЧЕВЫЕ И МУЛЬТИМОДАЛЬНЫЕ КОРПУСЫ

SPEECH AND MULTIMODAL CORPORA

Н. В. Богданова-Бегларян

N. V. Bogdanova-Beglarian

КОРОЛЯ ДЕЛАЕТ СВИТА: О ДОПОЛНИТЕЛЬНЫХ УСЛОВИЯХ ПРАГМАТИКАЛИЗАЦИИ ЯЗЫКОВЫХ ЕДИНИЦ В ПОВСЕДНЕВНОЙ РЕЧИ[1]

KOROL'A DELAET SVITA (*FOLLOWERS MAKE LEADERS*): ADDITIONAL CONDITIONS FOR PRAGMATICALIZATION OF LANGUAGE UNITS IN EVERYDAY SPEECH

Аннотация. Корпус, как известно, отличает от простой коллекции текстов наличие специальной разметки, или аннотации. Большие корпуса требуют автоматизации такой разметки, для чего создано немало программ. Однако существует разновидность разметки корпусных данных, которая полезна во многих аспектах, но пока совсем не поддается автоматизации, и даже ручное аннотирование сталкивается с целым рядом трудно разрешимых проблем. Речь идет о вычленении в устном тексте прагматических маркеров (ПМ), которое затруднено, в частности, тем обстоятельством, что все ПМ внешне ничем не отличаются от значимых речевых единиц и лишь в контексте реализуют свой новый ста-тус, появляющийся, как правило, в результате процесса прагматикализа-ции (*как его* (*её, их*), *это, это самое, типа того что, я не знаю* и под.). Особое внимание в статье уделено единицам, попавшим в класс ПМ из разряда вводных слов и словосочетаний: *так сказать, что называется, как говорится, собственно* (*говоря*) и под. Хезитационная сущность таких единиц «высокого порядка», уже прошедших в языке этап грамма-тикализации и зафиксированных в своем новом качестве вводных единиц словарями и грамматиками, не лежит на поверхности, но выявляется в ходе контекстного анализа, чему часто способствуют другие ПМ тако-го же типа: *короля*, что называется, *делает свита*…

Ключевые слова: повседневная речь, звуковой корпус, прагматический маркер, прагматикализация.

Abstract. The corpus, as we know, differs from a simple collection of texts by the presence of special markup, or annotation. Large corps require automation of such markup, for which many programs

[1] Исследование выполнено при финансовой поддержке гранта РНФ «Система прагматических маркеров русской повседневной речи» (проект № 18-18-00242).

have been created. However, there is a variety of case data markup, which is useful in many aspects, but so far not at all amenable to automation, and even manual annotation encounters a number of difficult-to-solve problems. The article discusses the problems of isolating pragmatic markers (PM) in the oral text. Such isolation is difficult, in particular, by the fact that all PM do not outwardly differ from meaningful speech units and only in context do they realize their new status, which appears, as a rule, as a result of the pragmatisation process (*kak jego (jejo. ikh*), *eto, eto samoe, tipa togo chto, ja ne znaju* etc.). Particular attention is paid to the units that fall into the PM class from the category of introductory words and phrases: *tak skazat', chto nazyvaets'a, kak govorits'a, sobstvenno* (*govor'a*) and others. The hezitational essence of such "high order" units does not lie on the surface, but is revealed during contextual analysis, which is often promoted by other PMs of the same type: *korl'a delaet svita* (*followers make leaders*)…
Keywords: everyday speech, speech corpus, pragmatic marker, pragmatikalization.

1. Введение

Неотъемлемыми элементами повседневной речи, помимо речевых (значимых) единиц, отражающих ее содержательную сторону, являются разнообразные условно-речевые единицы, совсем не связанные с содержанием, но вербализующие сам процесс ре-чепорождения. Речь идет о *прагматических маркерах* (ПМ) устной речи, которые в известной степени сближаются с хорошо разработанными в лингвистике *дискурсивными маркерами* (ДМ) (*немного, с трудом, в общем*) (см. подробнее об их различиях: [*Bogdanova-Beglarian, Filyasova* 2018; *Богданова-Бегларян* 2019]), но при этом имеют ряд специфических особенностей. Так же, как ДМ, прагматические маркеры помогают говорящему строить дискурс, но выступают исключительно в устной речи (или ее письменной стилизации), практически лишены лексического и отчасти грамматического значения, а главное — порождаются не осознанно, как ДМ, а на уровне речевых автоматизмов, ср.[2]:

(1) *борщик кушай // вот молодец *Н // *П и* () ***это самое*** *… *П а что ещё врач-то сказал вам ?*

(2) *говорит / смолен… смоленские говорит* ***эти как его*** (…) *приз лучшему когда получаешь / дают* ***эти*** */ кольцо / с бриллиантом*;

(3) *и она как на нас налетела !* ***вот там ты-ты-ты-ты-ты-ты*** */ да мы алкаши* ***там*** */* ***ну*** *что-то* ***там*** *такое / я не помню.*

Существует типология ПМ, разработанная на корпусном материале и включающая такие их разновидности, как поисковые хезитативы (*этот, как его* (*её, их*), *это самое*), метакоммуникативы (*знаешь,*

[2] О конвенциях дискурсивной транскрипции корпусных материалов см.: [*Русский язык…* 2016: 242–243].

(*я*) *не знаю, боюсь что*), рефлексивы (*или как это? скажем так*), маркеры-ксенопоказатели (*такой/ая/ие, типа того* (*что*), *вроде того* (*что*), *вот*), аппроксиматоры разного типа (*то-сё, пятое-десятое,* (*и*) *все дела,* (*и*) *всё такое* (*прочее*), *бла-бла-бла*) и некот. др. [*Богданова-Бегларян* 2014; *Bogdanova-Beglarian et al.* 2018a]. В настоящее время эта типология легла в основу методики аннотирования ПМ на больших объемах корпусного материала и находится в стадии апробации и уточнения [*Bogdanova-Beglarian et al.* 2018b]. Подобное аннотирование, равно как и просто поиск и идентификация ПМ в устном тексте, сталкивается с целым рядом проблем.

2. Корпус ОРД как источник материала для исследования

Источником материала для настоящего исследования стал корпус повседневной русской речи «Один речевой день» (ОРД) [см.: *Русский язык…* 2016; *Bogdanova-Beglarian et al.* 2016 a, b, 2017]. Корпус создан на филологическом факультете СПбГУ, это наиболее представительный на сегодняшний день ресурс для анализа русского устного дискурса, который в настоящее время активно обрабатывается. Все записи для ОРД проведены в максимально «естественных» условиях, с использованием методики непрерывной 24-часовой записи, и содержат по преимуществу разговоры из частной жизни информантов. Количественные показатели корпуса таковы: 1250 часов звучания, более 2800 коммуникативных макроэпизодов, записи речи 128 информантов, а также более 1000 их коммуникантов. Корпус лингвистически аннотирован (лемматизация, морфологическая и синтаксическая разметка и др.). Стремление получить максимально полный инвентарь тех функциональных единиц, которые использует человек в ходе речепорождения, и стремление описать их функционирование в речи разных говорящих и в разных коммуникативных ситуациях поставило перед разработчиками корпуса задачу провести еще один тип разметки — аннотирование в корпусе прагматических маркеров.

3. О проблемах выделения прагматических маркеров в устном тексте

Помимо главной проблемы разграничения в устном дискурсе ПМ и ДМ, существуют и другие. Так, практически все ПМ оказываются полифункциональными, то есть выполняют в тексте сразу несколько разных функций, а также отчетливо «тяготеют» друг к другу, выстра-

иваясь порой в протяженные хезитационные цепочки, в которых достаточно трудно отграничить один маркер от другого, ср.:

(4) ***ну там*** [1] (…) *сильно дешевле не было* / *потому что я* () *здесь* ***как бы*** [2] / *они всё равно ехали* ([1] — два СТАРТ./ХЕЗ., [2] — АППР./ХЕЗ.);

(5) *вот* // **П* ***так*** / ***щас-щас-щас-щас*** (ХЕЗ.);

(6) *там*(:) *они* ***это*** / ***как его ?*** *ближе*(:) *к спрессованы* / *или как-то там* (*э*) *было* (два ХЕЗ.);

(7) *колёса раскрутились* / *и свой в резину выскочил* // **В этот* / *о* / ***вот мол типа*** [1] *этот* / *Шумахер*(:)*$* / ***там ну этот*** [2] / *Якоб_Мюллер-то$* ([1] — два КСЕН.+ КСЕН./АППР.; [2] — три ХЕЗ.).

Говорящему словно бы мало одного маркера для вербализации преодоления возникших затруднений, и он непроизвольно «нагромождает» в тексте целые цепочки ПМ (чаще всего хезитативных) (и тем самым дает себе больше времени на преодоление заминки, «речевого сбоя»). Это, с одной стороны, свидетельствует о наличии (формировании?) системных (синонимических) отношений в классе ПМ, а с другой, ставит перед исследователями задачу разделения этих цепочек на отдельные единицы и выделения как основного, базового, вида каждого ПМ, так и его структурных вариантов.

Наконец, аннотирование ПМ затруднено и тем обстоятельством, что все они внешне ничем не отличаются от значимых речевых единиц и лишь в контексте реализуют свой новый статус[3], появляющийся, как правило, в результате процесса прагматикализации, ср.:

(8) ***там*** *мне кажется ближе* (наречие места);

(9) *всё равно вся эта утилизация короче она* ***там*** *максимум давала гарантию* ***там*** *на 50 лет* (ПМ);

(10) ***я не знаю*** / *отправила она его или нет* (главное предложение в составе сложноподчиненного);

(11) *или… или какой-то немецкий ?* ***ну я не знаю*** / *Бранденбургские_ворота$* / *что-то такое* (ПМ).

---

[3] Сходные выводы о дискурсивных словах/маркерах находим у К. Л. Киселевой и Д. Пайара: ДМ часто «сливаются» с контекстом, их анализ требует более «длинных» контекстов; наряду с дискурсивными, у этих слов есть и другие, недискурсивные, употребления; их «можно изучать только через их употребление» [*Дискурсивные слова…* 1998: 8–10].

4. О дополнительных условиях прагматикализации языковых единиц в повседневной речи

Именно в процессе *прагматикализации* в устном дискурсе появляются ПМ. Этот процесс часто протекает параллельно с *грамммматикализацией*, что хорошо видно, например, на матери-але вводных единиц. Так, форма *скажем* «отрывается» от гла-гольной парадигмы и начинает употребляться как вводное слово со значением ‘допустим, например’ — это грамматикализация. Да-лее та же форма, «обрастая» структурными компонентами, теряет и значение вводности (‘выражение отношений перечисления, указание на приемы и способы оформления мыслей, их связь и последовательность их изложения’) и начинает употребляться как ПМ — чаще всего *хезитатив* или *рефлексив*, ср.:

(12) *там сложная публика* / ***скажем так*** (РЕФЛ.);

(13) *вот есть* / *вещи такие* / *вот* / *ну* / *у людей хобби например* / *да ?* **П *В ну*(:) / ***там скажем*** / **П ну* / *не знаю* / *паяет что-то* (ХЕЗ.).

В примере (12) ПМ *скажем так* можно интерпретировать как рефлексив, маркер эвфемизации предшествующего выражения (*сложная публика*) (см. о рефлексивах в таком понимании [*Богданова-Бегларян* 2015]). Хезитационный же характер маркера *там скажем* в примере (13) «поддерживается» общим «хезитационным контекстом»: пауза **П*, вздох **В* (паралингвистический элемент вполне хезитативного характера), растяжка гласного в *ну*(:) как хезитационное явление — ***до*** рассматриваемого ПМ, и снова пауза **П* и еще два вербальных хезитатива (*ну* / *не знаю*) — ***после*** ПМ. Ср. еще несколько примеров такого же типа (допол-нительные ПМ в контекстах подчеркнуты):

(14) *но <u>вот как-то</u> ещё <u>как-то</u> / я первый раз **как говорится** / я <u>так</u> писала какие-то свои <u>там</u> // *П <u>чисто такие</u> / визуальные посмотрела <u>там</u> // мне так показалось*;

(15) *<u>ну</u> (…) со мной / <u>как бы</u> () **как говорится** / намного легче в том плане / что я могу поворчать / но я всё равно это сделаю*;

(16) *хорошо / <u>значит (э) / короче говоря</u> / если я сейчас закрою / то он / **так сказать** / подвигать(?) уже не будет*;

(17) *<u>вот но (…) в этом **собственно** (…) как сказать / в этом … *П</u> загадка России*;

(18) *я **собственно** мне непонятно <u>только(:) про(:) / скажем / нет ну(:)</u> / несколько человек и Зоя_Араратовна% непонятно / Архангельская% и Васильков%*;

(19) *вот такой тортик / *П там четыре пирожных //*П угу // @ кусочек мяты ! @ вкусный / он* (…) *в пластмассовой / прозрачной* (…) *упако**…** коробочке **такой**.*

Во всех приведенных контекстах в роли ПМ — как правило, *вербального хезитатива* — выступает единица, которую ни при каких условиях нельзя отнести к разряду «слов-паразитов» и хезитационная сущность которой выявляется исключительно в контексте, чему в значительной степени способствует «хезитационное окружение»: *короля делает* (*играет*) *свита…*

Важно отметить также, что при отсутствии такой «свиты» и словарного значения в употреблениях подобных единиц «высоко-го порядка» можно говорить об их немотивированном использо-вании говорящим — исключительно в целях украшения своей ре-чи («для красного словца»), что позволяет отнести их в класс ПМ *декоративов* (в рабочем обиходе — «никчемутивов»), ср.:

(20) *ну // в принципе / я сейчас смотрю / потому что / **как говорится** / он всё-таки составлял два месяца назад*;

(21) *ну здорово / ну всё будет зависеть от моего **так сказать** нового графика.*

5. Некоторые выводы

Проведенный анализ с отчетливостью показал, что прагматические маркеры русской речи, среди прочих своих специфических характеристик, имеют еще и ту особенность, что в ходе прагматикализации «опираются» на общий «хезитационный контекст» фразы. Это лишний раз подчеркивает необходимость анализа при выявлении ПМ широкого контекста, непременной ручной доработки результатов автоматического аннотирования ПМ в корпусном материале, а также учета этой особенности при лексикографическом «портретировании» подобных единиц.

Литература

1. *Богданова-Бегларян Н. В.* (2014) Прагматемы в устной повседневной речи: определение понятия и общая типология // Вестник Пермского ун-та. Российская и зарубежная филология, № 3 (27), с. 7–20.
2. *Богданова-Бегларян Н. В.* (2015) Рефлексив в системе дискурсивных единиц русской устной речи // Мир русского слова, № 3, с. 11–17.

3. *Богданова-Бегларян Н. В.* (2019) Об одной из проблем современной коллоквиалистики (в поисках терминов для новых единиц) // Языковые категории и единицы: синтагматический аспект. XIII Международная научная конференция, посвященная 90-летию профессора Августа Борисовича Копелиовича и 100-летию педагогического образования во Владимирской области, 24–26 сентября 2019 г. Владимир (в печ.).
4. *Дискурсивные слова* русского языка: Опыт контекстно-семантического описания (1998). *К. Л. Киселева, Д. Пайар* (*ред.*). М.: Метатекст, 280 с.
5. *Русский язык* повседневного общения: особенности функционирования в разных социальных группах. Коллективная монография (2016). *Н. В. Богданова-Бегларян* (ред.). СПб.: ЛАЙКА, 244 с.
6. *Bogdanova-Beglarian, N. V., Filyasova, Yu. A.* (2018) Discourse vs. Pragmatic Markers: A Contrastive Terminological Study. 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2018, SGEM2018 Vienna ART Conference Proceedings, 19-21 March, 2018. Vol. 5, Issue 3.1, pp. 123-130.
7. *Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Baeva, E., Martynenko, G., Ryko, A.* (2016a) Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. *SPECOM 2016,* Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 659–666.
8. *Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G.* (2016b) An Exploratory Study on Sociolinguistic Variation of Spoken Russian. *SPECOM 2016.* Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 100–107.
9. *Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G.* (2017) Linguistic Features and Sociolinguistic Variability in Everyday Spoken Russian. *SPECOM 2017.* Lecture Notes in Artificial Intelligence, LNAI, vol. 10458. Springer, Switzerland, pp. 503–511.
10. *Bogdanova-Beglarian, N., Baeva, E., Blinova, O., Martynenko, G., Sherstinova, T.* (2018a) Towards a Description of Pragm-atic Markers in Russian Everyday Speech. SPECOM 2018. Lecture Notes in Computer Science, vol. 11096. Springer, Cham, pp. 42–48.
11. *Bogdanova-Beglarian, N., Blinova, O., Martynenko, G., Sherstinova. T., Zaides, K.* (2018b) Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks. Proceedings of the FRUCT'23. Bologna, Italy, 13–16 November 2018 (eds.). FRUCT Oy, Fin-land, pp. 69–77.

References

1. *Bogdanova-Beglarian, N. V.* (2014) Pragmatemy v ustnoj povsednevnoj rechi: opredelenie pon'atia i obshchaja tipologia [Pragmatic Items in Everyday Speech: Definition of the Concept and General Typology]. Vestnik Permskogo un-ta. Rossijskaja i zarubezhnaja filologia [Perm University Herald. Russian and Foreign Philology]. № 3 (27), p. 7–20.
2. *Bogdanova-Beglarian, N. V.* (2015) Refleksiv v sisteme diskursivnykh jedinic russkoj ustnoj rechi [Reflexive in System of Discursive Items of Russian Oral Speech]. Mir russkogo slova [The World of a Russian Word]. № 3, p. 11–17.

3. *Bogdanova-Beglarian, N. V.* (2019) Ob odnoj iz problem sovremennoj kollokvialistiki (v poiskakh terminov dl'a novykh jedinic) [On One Problem of Modern Colloquialism (in Search of Terms for New Units)]. Jazykovye kategorii i jedinicy: sintagmaticheskiy aspekt. XIII Mezhdun. nauch. konf. [Language Categories and Units: a Syntagmatic Aspect. XIII International Scientific Conference], Vladimir (In Print).
4. *Diskursivnye slova* russkogo jazyka: Opyt kontekstno-semanticheskogo opisania [*Discursive Words of the Russian Language: Experience of Context-Semantic Description*] (*1998*). *Kiseleva, K. L., Payar, D.* (*eds.*). Moscow, 280 p.
5. *Russkij jazyk povsednevnogo obshchenia: osobennosti funkcionirovania v raznykh social'nykh gruppakh* [Everyday Russian Language: Functioning Features in Different Social Groups] (2016). *Bogdanova-Beglarian, N. V.* (ed.). Collective Monograph, St. Petersburg, 244 p.
6. *Bogdanova-Beglarian, N. V., Filyasova, Yu. A.* (2018) Discourse vs. Pragmatic Markers: A Contrastive Terminological Study. 5th Intern. Multidisciplinary Scientific Conf. on Social Sciences and Arts SGEM 2018, Vienna ART Conference Proceedings, 19–21 March, 2018. Vol. 5, Issue 3.1, pp. 123–130.
7. *Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Baeva, E., Martynenko, G., Ryko, A.* (2016a) Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. *SPECOM 2016,* Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 659–666.
8. *Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G.* (2016b) An Exploratory Study on Sociolinguistic Variation of Spoken Russian. *SPECOM 2016.* Lecture Notes in Artificial Intelligence, LNAI, vol. 9811. Springer, Switzerland, pp. 100–107.
9. *Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G.* (2017) Linguistic Features and Sociolinguistic Varia-bility in Everyday Spoken Russian. *SPECOM 2017.* Lecture Notes in Artificial Intelligence, LNAI, vol. 10458. Springer, Switzerland, pp. 503–511.
10. *Bogdanova-Beglarian, N., Baeva, E., Blinova, O., Martynenko, G., Sherstinova. T.* (2018a) Towards a Description of Pragmatic Markers in Russian Everyday Speech. Speech and Computer. SPECOM 2018. Lecture Notes in Computer Science, vol. 11096. Springer, Cham, pp. 42–48.
11. *Bogdanova-Beglarian, N., Blinova, O., Martynenko, G., Sherstinova. T., Zaides, K.* (2018b) Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks. Proceedings of the FRUCT'23. Bologna, Italy, 13–16 November 2018. FRUCT Oy, Finland, pp. 69–77.

Богданова-Бегларян Наталья Викторовна
Санкт-Петербургский государственный университет (Россия)
Bogdanova-Beglarian Natalia
Saint Petersburg State University (Russia)
E-mail: n.bogdanova@spbu.ru

А. В. Венцов, И. И. Коробейникова, Е. И. Риехакайнен
A. V. Ventsov, I. I. Korobeynikova, E. I. Riekhakaynen

АЛГОРИТМ ВОССТАНОВЛЕНИЯ РЕДУЦИРОВАННЫХ СЛОВОФОРМ В СПОНТАННОЙ РЕЧИ[1]

ALGORITHM OF REDUCED WORD FORMS RESTORATION IN SPONTANEOUS SPEECH

Аннотация. В статье описывается реализованный на языке Python алгоритм распознавания редуцированных словоформ в русской устной речи. Алгоритм учитывает имеющиеся на сегодняшний день данные о том, каким образом происходит распознавание неполного речевого сигнала человеком в процессе естественного общения, и включает в себя, помимо основного тела программы, модуль извлечения и предобработки морфологической информации потенциальных вариантов словоформ и правила обработки словоформ и построения синтаксических групп. Словарь словоформ и их реализаций, с которым работает программа, формируется на основе Корпуса транскрибированных русских устных текстов (http://narusco.ru/search/trn-search.php).
Ключевые слова: спонтанная речь, редуцированные словоформы, автоматическое распознавание речи, русский язык, язык программирования Python.

Abstract. In the paper, we introduce an algorithm of reduced word forms restoration for automatic speech recognition which takes into account the results of psycholinguistic experiments on spoken word recognition and includes, besides the main block, the module that retrieves the morphological information about candidates for recognition and the rules for word form processing and syntactic grouping. The source for the list of word forms and their realizations used by the algorithm is the Corpus of Transcribed Oral Russian Texts (http://narusco.ru/search/trn-search.php).
Keywords: spontaneous speech, reduced word forms, automatic speech recognition, Russian, Python programming language.

1. Проблема распознавания естественной устной речи

За последние пятьдесят лет был достигнут значительный прорыв в улучшении качества автоматических систем распознавания речи (см., например, [Ронжин, Ли 2007; Тампель 2015]). Прогресс в данной области обеспечивался не только повышением качества техники, но и изменениями в подходах к распознаванию устной речи, последнее из которых ознаменовалось началом использования глубоких нейронных сетей [Maas et al 2015; Zhang et al 2017]. Несмотря на это задача распознавания слитной речи все еще остается актуальной и нерешенной.

[1] Исследование выполняется при поддержке гранта РФФИ № 19-012-00629 «Алгоритмы восстановления редуцированных форм: роль системы языка».

Одной из основных проблем при этом остается фонетическая редукция словоформ в естественной речи [Кипяткова и др. 2013: 6]. Возможным способом решения данной проблемы и, как следствие, усовершенствования существующих автоматических систем распознавания речи является алгоритмизация стратегий, используемых носителем языка при восприятии фонетически неполного речевого сигнала.

Экспериментальные исследования свидетельствуют в пользу реалистичности гипотезы о том, что единицей перцептивного словаря носителя русского языка является словоформа [Венцов и др. 2003]. Однако статус редуцированных реализаций в ментальном лексиконе слушающего остается неясным. Ключевым фактором, обеспечивающим распознавание редуцированных единиц, признается контекст [Brouwer et al. 2013; Риехакайнен 2016], который позволяет слушающему восстановить словоформу с учетом сохранившихся перцептивно значимых фонетических элементов (для русского языка это прежде всего согласные). Таким образом, в распознавании редуцированных единиц при восприятии естественной речи участвует информация с разных уровней языка.

Вместе с тем нельзя исключать, что в перцептивном словаре носителя языка представлены все реализации, которые он когда-либо слышал. Привлекательность такого подхода становится очевидной при попытке моделирования распознавания речи как сегментации через идентификацию. Программа, реализующая данный подход при сегментации русских беспробельных текстов (в орфографии и транскрипции), работает с очень высокой надежностью при условии, что в словаре, к которому она обращается, представлены все словоформы, которые могут встретиться в тексте [Венцов и др. 2003]. Для того чтобы подобная задача была решена применительно к естественной устной речи, необходимо иметь словарь, включающий все возможные редуцированные реализации. Создание подобного словаря стало возможным в последние годы благодаря появлению корпусов устной речи. Исследование, которое будет описано в следующем разделе, осуществляется на материале Корпуса транскрибированных русских устных текстов, в котором все записи снабжены орфографической расшифровкой и акустико-фонетической транскрипцией (см. подробнее в [Nigmatulina et al. 2016]).

2. Моделирование процесса распознавания редуцированных словоформ

2.1. Материал

Основной задачей настоящего исследования являлось составление программной реализации алгоритма выбора единственно верного варианта интерпретации словоформы в межпаузальном интервале, соответствующего тому, которым предположительно пользуется слушающий в процессе восприятия речи. Для тестирования алгоритма из корпуса были выбраны пять клауз с именными группами, в которых редукции подвергается и существительное, и согласованные с ним зависимые слова. Были отобраны только те клаузы, внутри которых отсутствовали какие-либо паузы.

В словаре, имитирующем словарь слушающего, каждой словоформе в орфографии соответствуют все возможные варианты ее произнесения, которые встретились в Корпусе транскрибированных русских устных текстов.

2.2. Описание технической реализации

Для реализации алгоритма на языке программирования Python были написаны три модуля, один из которых использовался для извлечения и предобработки морфологической информации потенциальных вариантов словоформ, другой являлся основным телом программы, а третий содержал правила обработки словоформ и построения синтаксических групп.

На вход программе передается словоформа в транскрипции, для которой из словаря извлекаются все орфографические варианты, соответствующие данной транскрипции. Затем с помощью морфологического модуля для каждого из вариантов определяются его морфологические характеристики. В тех случаях, когда возникала морфологическая омонимия, для словоформы сохранялись несколько вариантов морфологического описания.

Дальнейшая обработка словоформ сводится к проверке наличия согласования между вариантами текущей словоформы и вариантами предшествующих словоформ (см. Рис. 1). Вариант текущего слова сопоставляется с ранее обработанными вариантами предыдущих словоформ в соответствии с заданными правилами[2]. При совпадении

[2] Правила опираются на частеречную принадлежность сравниваемых слов. Так, например, если вариант предшествующего слова является прилагательным, а теку-

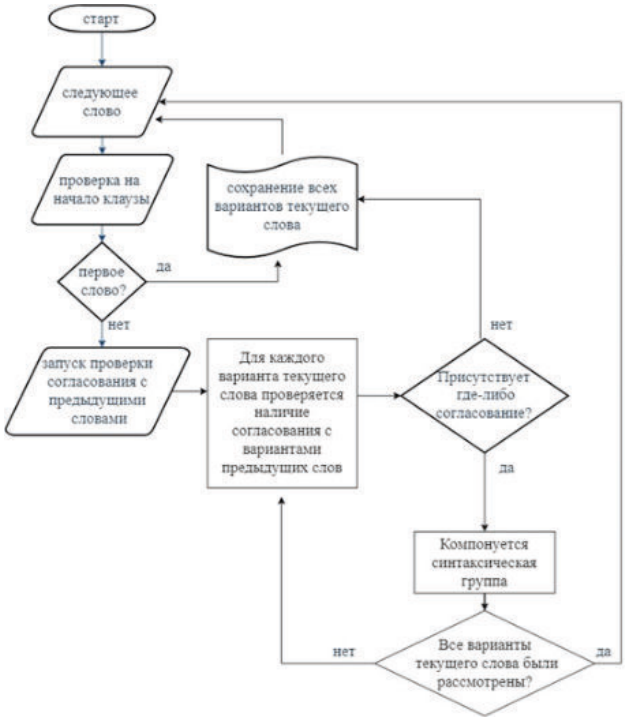


Рис. 1. **Схема работы алгоритма распознавания редуцированных словоформ**

морфологических характеристик у текущей и предшествующей словоформ компонуется именная, глагольная или местоименная группа. Как только все варианты текущего слова будут проверены, скомпонованные группы сохраняются, в то время как те варианты словоформы, которые не согласовались с вариантами предыдущих словоформ, удаляются. Алгоритм переходит к обработке вариантов следующего слова.

При проверке работы программы на отобранном материале было выявлено, что большая часть редуцированных словоформ успешно восстанавливается при опоре на контекст. Например, для клаузы а *некоторые школы придётся поднимать родителям, родительским комитетам* [a n'e:kte Skol pQd'o:c p@dn'e:maIt' reId'it'l'e:m reId'it'@sk'e:m

щее слово — существительное, то для проверки наличия согласования словоформы будут проверяться на совпадение рода, числа и падежа.

k@m'it'e:t][3], где несколько потенциально возможных вариантов имели словоформы [n'e:kte] (*некоторые, некоторая, некоторое, некто*), [Skol] (*школа, школу, школы*), [reId'it'l'e:m] (*родителям, родителем*), [k@m'it'e:t] (*комитетам, комитет, комитету, комитета, комитеты*), алгоритм вернул следующие скомпонованные группы:

'PossibleNP': ['не+которая', 'шко+ла']
'PossibleNP': ['не+которые', 'шко+лы']
'PossibleNP': ['не+которые', 'шко+лы']
'PossibleVP': ['придё+тся', 'поднима+ть']
'PossibleNP': ['роди+тельским', 'комите+там']

Для словоформы [reId'it'l'e:m] однозначность установить не удалось. Алгоритм вернул оба возможных варианта. Также можно заметить, что выбор падежа именных групп не был осуществлен, поскольку отсутствует информация о валентностной структуре глагола.

2.3. Проблемы и перспективы

Тестирование работоспособности алгоритма на ограниченном корпусном материале показало принципиальную возможность его использования, а также выявило ряд проблем, на решение которых должна быть направлена дальнейшая работа. В частности, при возникновении конкуренции в равной степени возможных вариантов необходимо учитывать частотное распределение единиц в речи. Возможным решением данной проблемы может стать внедрение n-граммных моделей или составление частотного словаря на основе используемого материала, что требует предварительного анализа больших объемов естественной устной речи. Кроме того, значительную трудность представляла обработка глаголов. Для учета моделей управления глагола необходима готовая формализованная база глаголов и их валентностных структур.

Таким образом, в настоящем исследовании была осуществлена попытка моделирования процесса восстановления редуцированных единиц в русской речи, которым пользуется носитель языка при распознавании речи. Дальнейшая работа будет нацелена как на улучшение качества технической реализации (дополнение правил восстановления единиц, внедрение учета частотных характеристик словоформ),

[3] Здесь и далее используются принципы транскрипции, принятые в Корпусе транскрибированных русских устных текстов (см. http://narusco.ru/transkrip.htm).

так и на расширение алгоритма в направлении синтаксической и семантической обработки дискурсивных единиц.

Литература

1. *Венцов А. В., Касевич В. Б., Ягунова Е. В.* (2003), Корпус русского языка и восприятие речи, Научно-техническая информация. Сер. 2: Информационные процессы и системы, 6, с. 25–32.
2. *Кипяткова И. С., Ронжин А. Л., Карпов А. А.* (2013), Автоматическая обработка разговорной русской речи. СПб.
3. *Риехакайнен Е. И.* (2016), Восприятие русской устной речи: контекст + частотность. СПб.
4. *Ронжин А. Л., Ли И. В.* (2007), Автоматическое распознавание русской речи, Вестник РАН, 77 (2), с. 133–138.
5. *Тампель И. Б.* (2015), Автоматическое распознавание речи — основные этапы за 50 лет, Научно-технический вестник информационных технологий, механики и оптики, 15 (6), с. 957–968.
6. *Brouwer S., Mitterer H., Huettig F.* (2013), Discourse context and the recognition of reduced and canonical spoken words, Applied Psycholinguistics, 34, pp. 519–539.
7. *Maas A. L, Xie Z., Jurafsky D., Ng A. Y*, (2015), Lexicon-free conversational speech recognition with neuralnetworks, inNAACL HLT 2015, The 2015 Conference ofthe North American Chapter of the Association for Compu-tational Linguistics: Human Language Technologies, Denver,Colorado, USA, May 31 — June 5, 2015, 2015, pp. 345–354
8. *Nigmatulina Ju., Rajeva O., Riechakajnen E., Slepokurova N., Vencov A.* (2016), How to Study Spoken Word Recognition: Evidence from Russian, Slavic Languages in Psycholinguistics: Chances and Challenges for Empirical and Experimental Research, Tübinger Beiträge zur Linguistik, 554, pp. 175–190.
9. *Zhang Y., Pezeshki M., Brakel P., Zhang S., Bengio CLY., Courville A.* (2017),Towards end-to-end speech recognition with deep convolutional neural networks, arXiv preprint arXiv:1701.02720.

References

1. *Ventsov A. V., Kasevitch V. B., Yagunova E. V.* (2003), Korpus russkogo jazyka i vosprijatie rechi [Russian Language Corpus and Spoken Word Recognition], Nauchno-tekhnicheskaja informatsija. [Scientific and Engineering Information], 2: Informatsionnye protsessy i sistemy [Information Processes and Systems], 6, pp. 25–32.
2. *Kipjatkova I. S., Ronzhin A. L., Karpov A. A.* (2013), Avtomaticheskaja obrabotka razgovornoj russkoj rechi [Automatic Processing of Conversational Russian]. St. Petersburg.
3. *Riekhakaynen E. I.* (2016), Vosprijatie russkoj ustnoj rechi: kontekst + chastotnosť [Recognition of Russian speech: context + frequency]. St. Petersburg.

4. *Ronzhin A. L., Li I. V.* (2007), Avtomaticheskoje raspoznavanije russkoj rechi [Automatic Recognition of Russian Speech], Vestnik RAN [Bulletin of the Russian Academy of Sciences], 77 (2), pp. 133–138.
5. *Tampel' I. B.* (2015), Avtomaticheskoje raspoznavanie rechi — osnovnye etapy za 50 let [Automatic Speech Recognition: the Main Stages of 50 Years], Nauchno-tekhnicheskij vestnik informatsionnykh tekhnologij, mekhaniki i optiki [Scientific and Engineering Bulletin for Informational Technologies, Mechanics and Optics], 15 (6), pp. 957–968.
6. *Brouwer S., Mitterer H., Huettig F.* (2013), Discourse context and the recognition of reduced and canonical spoken words, Applied Psycholinguistics, 34, pp. 519–539.
7. *Maas A. L, Xie Z., Jurafsky D., Ng A. Y*, (2015), Lexicon-free conversational speech recognition with neuralnetworks, inNAACL HLT 2015, The 2015 Conference ofthe North American Chapter of the Association for Compu-tational Linguistics: Human Language Technologies, Denver,Colorado, USA, May 31 — June 5, 2015, 2015, pp. 345–354.
8. *Nigmatulina Ju., Rajeva O., Riechakajnen E., Slepokurova N., Vencov A.* (2016), How to Study Spoken Word Recognition: Evidence from Russian, Slavic Languages in Psycholinguistics: Chances and Challenges for Empirical and Experimental Research, Tübinger Beiträge zur Linguistik, 554, pp. 175–190.
9. *Zhang Y., Pezeshki M., Brakel P., Zhang S., Bengio C. L. Y., Courville A.* (2017),Towards end-to-end speech recognition with deep convolutional neural networks, arXiv preprint arXiv:1701.02720.

Венцов Анатолий Владимирович
Санкт-Петербургский государственный университет (Россия)
Ventsov Anatoly
Saint Petersburg State University (Russia)
E-mail: av.ventsov@gmail.com

Коробейникова Ирина Игоревна
Санкт-Петербургский государственный университет (Россия)
Korobeynikova Irina
Saint Petersburg State University (Russia)
E-mail: korobeinikova.ir@mail.ru

Риехакайнен Елена Игоревна
Санкт-Петербургский государственный университет (Россия)
Riekhakaynen Elena
Saint Petersburg State University (Russia)
E-mail: e.riehakajnen@spbu.ru

К. Д. Зайдес
K. D. Zaides

ОБ УНИФИКАЦИИ РАЗМЕТКИ КОРПУСА «СБАЛАНСИРОВАННАЯ АННОТИРОВАННАЯ ТЕКСТОТЕКА»[1]

ON THE STANDARDIZATION OF THE CORPUS "BALANCED ANNOTATED TEXT COLLECTION"

Аннотация. Доклад посвящен процессу и результатам унификации разметки корпуса «Сбалансированная аннотированная текстотека». Данный корпус состоит из нескольких отдельных блоков, репрезентирующих устную речь представителей разных социальных и психологических групп. Для дальнейших лингвистических исследований, а также в целях сравнения данных, полученных на материале иных корпусов, необходимо было унифицировать систему разметки корпуса. На текущем этапе производилась замена основных знаков транскрипции, отмечающих особые явления, свойственные устной спонтанной речи (обрывы, паузы хезитации и т. п.). Полученный в результате массив текстов отражает иной, более современный, подход в аннотации лингвистических корпусов.
Ключевые слова: корпусная лингвистика, тег, корпусная разметка, устная речь, спонтанная речь, аннотирование, монолог, Сбалансированная аннотированная текстотека.

Abstract. The report describes the process and the results of unification of annotation of the corpus «Balanced Annotated Texts» (BAT). BAT consists of several blocks that represent oral speech of members of different social and psychological groups. For further linguistic research, it is needed to unify the system of corpus annotation. At this stage, the replacement of main tags (used for marking the speech disfluencies) in transcription has been made. As a result, in the received corpus the modern approach to the corpus annotation is reflected.
Keywords: corpus linguistics, tag, corpus annotation, oral speech, spontaneous speech, monologue, Balanced Annotated Text Collection.

1. Корпус «Сбалансированная аннотированная текстотека»: история формирования и структура

Корпус «Сбалансированная аннотированная текстотека» (САТ) является масштабной базой данных по устной спонтанной монологической речи носителей русского языка и изучающих русский язык. Корпус формировался с 1997 года, когда началась запись первого его блока, включающего речь женщин-медиков разного возраста и медицинской специализации (от медсестер до кандидатов медицинских наук), всего — 150 монологов. Информанты читали и пересказывали

[1] Статья подготовлена при финансовой поддержке гранта РНФ (проект № 18-18-00242 «Система прагматических маркеров русской повседневной речи»).

сюжетный и несюжетный тексты, описывали сюжетное и несюжетное изображения, рассказывали о том, как проводят свободное время; таким образом, от каждого говорящего было записано 7 монологов, построенных по различным коммуникативным сценариям. После записи полученные монологи расшифровывались в орфографическом виде; в расшифровках с помощью специальных помет отмечались некоторые характерные явления устной спонтанной речи.

После записи первого блока САТ, корпус пополнился записями речи юристов, разного пола, возраста, специализации (201 монолог). В дальнейшем для корпуса были записаны следующие блоки: речь преподавателей РКИ (32 монолога); речь преподавателей-философов (12 монологов); речь смешанной профессиональной группы (сбалансированной по иным социальным характеристикам) (12 монологов); речь студентов (филологов и нефилологов) (172 монолога); речь «компьютерщиков» (28 монологов).

Помимо этого, в корпусе появился новый блок — русская интерферированная речь иностранцев, т. е. монологи носителей других языков, как на русском, так и на их родном языке, посвященные одной и той же теме («Ваши впечатления о Петербурге»): 50 монологов от носителей китайского языка и по 16 монологов от голландцев, американцев и франкофонов. Разметка явлений спонтанности в этом блоке в основном соответствовала тем принципам и тегам, которые были приняты при записи другого корпуса спонтанной речи, созданного в СПбГУ, — «Один речевой день» (ОРД). Именно этот тип разметки должен был стать единым, унифицированным для двух корпусов речи, в целях их дальнейшего исследования и планомерной систематизации материала.

Тексты, входящие в корпус САТ, впоследствии издавались в виде отдельных сборников [Русская спонтанная речь… 2008, 2010, 2011, 2018]; планируется издание и других блоков корпуса. Подробнее о корпусе и исследованиях, выполненных на его материале, см.: [Звуковой корпус… 2013].

На данный момент корпус содержит 705 монологов разного типа, среди них 72 чтения текста, 224 пересказа текста, 254 описания изображения и 155 свободных рассказов, записанных от более чем 200 информантов. Однако характеризующийся полнотой и сбалансированностью корпус не был унифицирован по пометам, которые используются для маркирования различных спонтанных явлений устной речи. Вследствие этого была предпринята необходимая унификация

основных тегов корпуса, которые использовались аннотаторами при расшифровке не согласованно, так как корпус записывался с перерывами.

Кроме того, перед началом работы по приведению расшифровок к единому виду, корпус необходимо было собрать в единую базу данных, поскольку до этого он существовал в виде разрозненных расшифровок и прилагаемых аудио-файлов, опубликованных в сборниках или ученических работах. С этой целью была создана таблица, в которую были внесены как сами монологи, так и информация о текстах и говорящих, что послужило первичной метаразметкой корпуса. Таким образом, в текущей базе хранятся сам текст монолога и сведения о его типе, а также номер, пол, возраст, профессиональная принадлежность информанта, определенный исследователями УРК говорящего и данные о его психологическом типе (экстраверсия/интроверсия, уровень нейротизма и тип темперамента, определенный на основании первых двух характеристик). Фрагмент полученной таблицы представлен на рис. 1.

	A	B	C	D	E	F	G	H	I	J	K
99	человек ловит рыбу / в озере / Описание С		18	м		25	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	холерик
100	однажды / рыбак пошёл () на о Описание С		19	м		30	СПб	среднее	юрист (ОВД)	низкий	амбиверт	сангвиник-флегматик
101	в одной деревне / проживал ст Описание С		20	м		35	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	сангвиник
102	один рыбак / пошёл ловить ры Описание С		21	ж		28	СПб	среднее	юрист (ОВД)	низкий	амбиверт	сангвиник-флегматик
103	один мужчина / решил пойти н Описание С		22	ж		22	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	сангвиник
104	мужчина / пожилой / неприятн Описание С		23	ж		35	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	холерик
105	плоды рекламы // (м-м) на дан Описание С		24	м		36	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	сангвиник
106	(э-э) престарелый господин / к Описание С		25	м		37	СПб	высшее (к.ю.н.)	юрист (преподаватель	высокий	амбиверт	флегматик
107	на первой картинке / изображ Описание С		26	м		29	СПб	высшее (к.ю.н.)	юрист (преподаватель	высокий	типичный экстраверт	сангвиник
108	плоды рекламы / деликатес по Описание С		27	м		43	СПб	высшее (к.ю.н.)	юрист (преподаватель	высокий	типичный интроверт	флегматик
109	значит / (э-э) суда () по картинк Описание С		28	ж		29	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	сангвиник
110	человек (э-э) поймал / (э-э) оч Описание С		29	м		27	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	сангвиник
111	рыбалка // (м-м) рыбак пытает Описание С		30	ж		48	СПб	высшее (к.ю.н.)	юрист (преподаватель	высокий	амбиверт (но неискренна в отв	флегматик (но неискренна в отве
112	значит / (э-э) на самой верней Описание С		31	ж		29	СПб	высшее (к.ю.н.)	юрист (преподаватель	высокий	амбиверт	сангвиник
113	итак передо мной рисунок / по Описание С		32	м		32	СПб	высшее (к.ю.н.)	юрист (преподаватель	высокий	типичный экстраверт	сангвиник
114	(э-э) на картине () изображен ь Описание С		33	ж		22	СПб	среднее	юрист (ОВД)	низкий	амбиверт	меланхолик
115	один мужичок / пошёл *К поло Описание С		34	ж		35	СПб	среднее	юрист (ОВД)	низкий	типичный экстраверт	холерик
116	*К значит / я смотрел / рисунок Описание С		35	м		37	СПб	среднее	юрист (ОВД)	низкий	амбиверт	сангвиник

Рис. 1. **Фрагмент таблицы, содержащей размеченные тексты САТ**

2. Процесс, принципы и проблемы унификации корпуса «Сбалансированная аннотированная текстотека»

Перед началом работы по унификации помет было определено, что для маркирования одних и тех же явлений разными аннотаторами использовались разные знаки расшифровки. Так, например, для обозначения незаполненной паузы хезитации использовались следующие

символы: ∫, ⌊, (), (…), <…>. Символы ∫, ⌊ обозначали короткие паузы хезитации, заминки, остальные — длинные незаполненные хезитационные паузы. В силу того что блоки корпуса создавались разными собирателями для разных исследовательских целей, с долгими временными интервалами, несогласованность обозначений в разметке затрагивала как собственно фонетические явления (по-разному маркировались паузы, растяжки звуков, обрывы фраз), так и паралингвистические характеристики речи.

В процессе унификации текстов корпуса был принят следующий список помет (сходный с тем, что используется в корпусе ОРД [Asinovsky et al. 2009]) — специальных тегов, отмечающих значимые для дальнейших исследований особые черты устной спонтанной речи (табл. 1).

Таблица 1. Список **помет для унификации корпуса**

Помета	Явление
/	короткая пауза
//	длительная пауза
()	короткая пауза хезитации, заминка
(…)	длительная пауза хезитации
(а-а), (м-м), (э-э) и т. п.	заполненные паузы хезитации
*С	смех
*О	вздох
*К	кашель
*Н	неразборчивая речь
*Ц	цыканье языком
…	обрыв слова (ставится без пробела)
…	обрыв фразы (ставится с пробелом)
(:)	растяжка звука
[речь экспериментатора]	специфическая для монологической речи помета

При обработке текстов корпуса осуществлялась автоматизированная замена помет на унифицированный их набор, приведенный выше. Это позволит в дальнейшем работать с двумя корпусами устной

речи — САТ и ОРД, — пользуясь единой системой разметки, и полноценно сравнивать полученные данные.

На первом этапе унификации знаков разметки в таблицах, содержащих тексты САТ и структурную и метаразметку, были автоматически заменены те элементы, которые могли быть подвергнуты автоматической замене в рамках программы Excel, а именно: короткие и длительные паузы хезитации, а также заполненные паузы хезитации, которые в исходных текстах не всегда заключались в скобки. В некоторых случаях заполнение паузы хезитации отмечалось с помощью более чем 2 букв, например, *э-э-э*. Для таких примеров пришлось сначала заменить их на эквивалентные, но заключенные в скобки, а затем удалить из них лишнюю букву автоматической заменой на выражение без нее.

На втором этапе осуществлялось приведение к единому виду знаков расшифровки, передающих паралингвистические явления, которые не могли быть заменены автоматически. Эти знаки, отражающие смех, кашель, вздох, а также неразборчивый фрагмент расшифровки, заключались в исходных текстах в разного вида скобки: <>, (), [], — и должны были быть в итоге заменены на кириллические буквы с астериском: *С, *К, *О, *Н. Для учета всех вариантов написания рассматриваемых помет в текстах корпуса, был создан список всех возможных вариантов символов, нуждающихся в замене. Для осуществления замены было создано регулярное выражение, написанное на языке программирования Python. Данный язык программирования был избран в связи с наличием удобной библиотеки Pandas, предназначенной для эффективного оперирования с таблицами. После тестирования корректности работы выражения и осуществления замен, в результате был получен унифицированный по пометам, отмечающим паралингвистические явления, корпус.

Однако после унификации всех рассмотренных помет оставался один тег, для замены которого требуется учет исключений, существующих в русском языке. В корпусе САТ для обозначения растяжек (продлений) звуков в речи использовалось дефисное написание удлиняемого звука, например:

- *понял что его заманили в ветеринар**н-н**ую лечебницу*;
- *Зина! // держи его! мерзавца! за шиворот! / **с-с**котина!*

В корпусе ОРД традиционно, с начала его формирования, растяжки звука отмечались знаком (:), и именно к этому эталону должна была быть приближена разметка унифицируемого корпуса. При автомати-

ческой замене любых кириллических символов такого рода, разделенных дефисом, не будут учтены следующие исключения:

1) заполненные паузы хезитации: (*м-м*), (*э-э*) и т. п.;
2) стыки как разных, так и однородных по типу звуков, графически разделяемых дефисом (*како**й-т**о*, *ко**т-т**о поел?* и т. д.).

Ручная замена данных символов методом сплошного прочтения и замены исключена в силу большого объема входящих в корпус текстов и относительной частотности случаев растяжки звука.

Предполагается, что исключение тех случаев, когда замена написания с дефисом на соответствующий знак растяжки (:) происходить не должна, может достигаться следующим решением. Во-первых, необходимо учитывать левый и правый соседний символы, не производя замену тогда, когда через 1 символ от дефиса есть открывающая или закрывающая скобки. Таким образом решается первая из озвученных проблем.

Полного решения второй проблемы — разграничения тех случаев, где графический дефис не отражает растяжку звука, — пока не найдено. Поскольку растягиваемый звук отражается на письме одной и той же буквой, при разработке алгоритма следует сразу исключить из рассмотрения те варианты написания слов, где слева и справа от дефиса стоят разные по качеству (месту и способу образования) звуки. Однако ситуация осложняется тем, что расшифровщики иногда отмечали растяжку «фонетически», оставляя на первом месте ту букву, которая должна быть написана в данном слове, но после дефиса ставили не букву, а, фактически, растянутый звук, ср.:

- *рассвет в такие дни* / ***не-э*** () *пылает* // *заревом*.

Кроме того, даже если оставить эти случаи для последующей ручной коррекции автоматически унифицированной разметки корпуса, все равно необходимо сделать еще одно исключение. Так, иногда в русском языке некоторые частицы, отделяющиеся от слова, к которому они присоединяются, дефисом, начинаются с той же буквы, на которую заканчивается слово:

- *вот как он оде**т-т**о* / *неплохо* / *ну вот*.

Отметим, что растягивание глухих согласных в русском языке затруднено, однако вполне характерно для устной речи. Конечно, чаще всего в речи встречаются удлиняемые гласные и сонорные согласные,

которые не смешиваются с дефисными написаниями некоторых частиц. Таким образом, перед последним этапом разметки необходимо задать список слов-исключений, тем самым решив проблему тех замен, которые будут произведены по ошибке.

Литература

1. Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 1. Чтение. Пересказ. Описание (2013) / Отв. ред. *Н. В. Богданова-Бегларян*. СПб.
2. Русская спонтанная речь. Монологи-описания. Тексты. Лексические материалы (2011) / Сост. *В. В. Куканова* / Отв. ред. и автор предисловия *Н. В. Богданова*. СПб.
3. Русская спонтанная речь. Монологи-репродуктивы. Тексты. Лексические материалы (2010) / Сост. *В. В. Куканова* / Отв. ред. и автор предисловия *Н. В. Богданова*. СПб.
4. Русская спонтанная речь. Спонтанные монологи разных типов. Тексты. Лексические материалы (CD) (2018) / Сост. *Н. В. Богданова-Бегларян*, *И. С. Бродт* / Отв. ред. *М. Краузе* // Бюллетень Фонетического Фонда. Бохум, Германия.
5. Русская спонтанная речь. Свободные монологи-рассказы на заданную тему. Тексты. Лексические материалы (2008) / Сост. *В. В. Куканова* / Отв. ред. и автор предисловия *Н. В. Богданова*. СПб.
6. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: creation principles and annotation. V. Matoušek, P. Mautner (eds.), TSD 2009, LNAI, vol. 57292009, Berlin-Heidelberg, Springer publ., pp. 250–257.

References

1. Zvukovoj korpus kak material dlya analiza russkoj rechi. Kollektivnaya monografiya. Chast' 1. Chtenie. Pereskaz. Opisanie [The Sound Corpus as a Material for the Analysis of Russian Speech. Collective Monograth. Part I. Reading. Retelling. Description] (2013) / Ed. *N. V. Bogdanova-Beglaryan*. Saint Petersburg.
2. Russkaya spontannaya rech'. Monologi-opisaniya. Teksty. Leksicheskie materialy [Russian Spontaneous Speech. Descriptive Monologues. Texts. Lexical Materials] (2011) / Comp. V. V. Kukanova / Ed. and author of preface N. V. Bogdanova. Saint Petersburg.
3. Russkaya spontannaya rech'. Monologi-reproduktivy. Teksty. Leksicheskie materialy [Russian Spontaneous Speech. Reproductive Monologues. Texts. Lexical Materials] (2010) / Comp. V. V. Kukanova / Ed. and author of preface N. V. Bogdanova. Saint Petersburg.
4. Russkaya spontannaya rech'. Spontannye monologi raznykh tipov. Teksty. Leksicheskie materialy (CD) [Spontaneous Monologues of Different Types. Texts. Lexical Materials] (2017) / Comp. N. V. Bogdanova-Beglaryan, I. S. Brodt / Ed. M. Krauze // Byulleten' Foneticheskogo Fonda [Bulletin of Phonetic Fond]. Bokhum, Germany.

5. Russkaya spontannaya rech'. Svobodnye monologi-rasskazy na zadannuyu temu. Teksty. Leksicheskie materialy [Monologues-narratives on Given Subject. Texts. Lexical Materials] (2008) / Comp. V. V. Kukanova / Ed. and author of preface N. V. Bogdanova. Saint Petersburg.
6. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: creation principles and annotation. V. Matoušek, P. Mautner (eds.), TSD 2009, LNAI, vol. 57292009, Berlin-Heidelberg, Springer publ., pp. 250–257.

Зайдес Кристина Денисовна
Санкт-Петербургский государственный университет (Россия)
Zaides Kristina
Saint Petersburg State University (Russia)
E-mail: kristina.zaides@student.spbu.ru

А. А. Зинина, А. А. Котов, Н. А. Аринкин,
Л. Я. Зайдельман, М. М. Цфасман

A. Zinina, A. Kotov, N. Arinkin, L. Zaydelman, M. Tsfasman

НАПРАВЛЕНИЯ КОММУНИКАТИВНЫХ ДЕЙСТВИЙ В МУЛЬТИМОДАЛЬНОМ КОРПУСЕ REC[1]

DIRECTIONS OF COMMUNICATIVE ACTIONS IN MULTIMODAL CORPUS REC

Аннотация. В статье описан новый блок разметки мультимодального корпуса REC (Russian Emotional Corpus) – разметка коммуникативно-значимых движений глаз (2 видео из подкорпуса эмоциональных интервью, общей продолжительностью 127 минут). Выделено 1875 аннотаций для движений глаз и 1415 аннотаций функциональной разметки. В работе показано, что направления взглядов и жестов важны в естественном общении: как в диалоге, так и при решении пространственных задач. На роботе Ф-2 смоделировано коммуникативное поведение в ситуации совместного с пользователем решения пространственной задачи (танграм) с использованием ориентированных жестов и направления взгляда. Взаимодействие пользователя с роботом было также записано для анализа коммуникативных ответов человека на ориентированные жесты робота (31 видеозапись, общей продолжительностью 310 минут).
Ключевые слова. мультимодальная коммуникация, ориентированные жесты, движения глаз, человеко-машинное взаимодействие.

Abstract. The article describes a new block in the marking of multimodal corpus REC (Russian Emotional Corpus): the marking of communicatively significant eye movements (2 videos from the emotional interviews subcorpus with a total duration of 127 minutes). There are 1875 annotations for eye movements and 1415 annotations of functional marking. The paper shows that eye gaze direction and pointing gestures are important in natural communication: both in dialogue and in spatial problem solving. We have modeled the communicative behavior of F-2 robot to solve spatial problems (tangram) together with the user. Oriented gestures and eye gaze directions are used in the robot behavior. Human – robot interaction was recorded to analyze the person's communicative responses to the robot's pointing gestures (31 videos with a total duration of 310 minutes).
Keyword. multimodal communication, oriented gestures, eye movements, human-machine interaction.

Введение

Интерфейс робота-компаньона должен быть «естественным» и простым для пользователя. Поэтому коммуникативное поведение такого робота должно быть максимально приближено к коммуникативному поведению человека. Кроме этого, поведение робота должно быть сложным и разнообразным, чтобы поддерживать более длительное и комфортное взаимодействие человека с роботом.

[1] Исследование выполнено при поддержке РФФИ, проект № 16-29-09601 офи_м.

В естественном коммуникативном взаимодействии собеседники используют сразу несколько каналов передачи информации и помимо естественной речи применяют целый комплекс невербальных средств: мимику, жесты, движения головы и тела, а также направление взгляда. Направление взгляда говорящего фиксируется слушающим, поэтому помимо зрительной функции обладает ещё и большой коммуникативной значимостью. В работе [Kibrik, Fedorova, 2018] рассматривается распределение внимания участников диалога: показывается, что внимание рассказчика в большей степени направлено на лицо говорящего и в меньшей степени — на его руки. Во многих работах, посвященных исследованию реальной коммуникации, отмечаются индвидуальные различия в поведении собеседников. Например, некоторые люди смотрят преимущественно в глаза, некоторые — преимущественно на рот собеседника, в то время как другие в различной степени распределяют взгляд между глазами и ртом [Kanan et al., 2015]. Более того, в эксперименте [Rogers et al., 2018] показывается, что субъективное восприятие зрительного контакта является продуктом взаимного взгляда на лицо (континуум между глазами и ртом собеседника), а не фактического взаимного зрительного контакта. Кроме этого, участники диалога часто субъективно переоценивают степень взаимного зрительного контакта [Gamer, Hecht, 2007]. Таким образом, если направление взгляда является существенным элементом коммуникации, то моделирование этого процесса важно как для анализа диалога, так и для создания привлекательных роботов, производящих ощущение зрительного контакта. Как было показано в [Häring, et. al., 2012] направленный взгляд робота повышает эффективность взаимодействия между роботом и пользователем.

Исследователи также выявили, что использование роботом ориентированных жестов, сопровождающих речь, способствует пониманию пространственной информации, увеличивает скорость и точность выполнения пространственных задач [Cabibihan, et. al., 2009]. В работе [Salem et al., 2012] показано, что испытуемые оценивают робота более позитивно, когда робот сопровождает речь жестами, даже если они семантически не соответствуют речи.

В лаборатории нейрокогнитивных технологий Курчатовского института мы разрабатываем робота Ф-2: персонального робота-компаньона, который способен поддерживать коммуникацию с человеком с помощью речи, жестов и мимики [Kotov et al, 2019]. Поведение робота-компаньона мы моделируем на основе анализа поведения людей

в реальных коммуникативных ситуациях на основе мультимодального корпуса REC (Russian Emotional Corpus). Корпус содержит размеченные в программе ELAN видеозаписи реальных эмоциональных диалогов (n = 815). В корпусе вручную размечаются речевые высказывания участников диалога. Для информанта размечаются движения его глаз, рук и губ. Для жеста или элемента мимики также отмечается коммуникативная функция в соответствии с типологией, представленной в [Котов, Зинина, 2015].

1. Исследование ориентированных жестов в корпусе

1.1. Направление взгляда

Стандартная разметка движений глаз учитывает значимые изменения направления линии взгляда (взгляд вбок, взгляд вверх), расширение глаз, прищуры, подмигивания, закрытие глаз. Также в корпусе размечены поднятие бровей и движения носом. Такая разметка позволяет исследовать ключевые коммуникативные особенности информантов: например, глазодвигательное поведение во время фрустрации, радости, апелляции и др. Выделять паттерны типичные для начала или окончания разговора. Однако этой разметки недостаточно для моделирования сложного коммуникативного поведения робота-компаньона. Поэтому 2 видео из подкорпуса эмоциональных интервью (127 минут) были размечены с помощью дополненной разметки. Применявшаяся ранее в корпусе разметка направления взгляда была дополнена 7-ю новыми тегами: «*справа*» (от говорящего), «*слева*», «*вверх*», «*вниз*», «*к собеседнику*», «*к объекту*», «*закрыты*». Кроме этого, общий инвентарь коммуникативных функций был дополнен функциями, специфическими для глазодвигательного поведения: «*размышление*», «*говорение*», «*фразовое ударение*», «*перечисление*», «*шутка*», «*слушание*», «*внимание на реакцию собеседника*», «*подражание направлению взгляда собеседника*», «*смягчение антисоциальной ситуации*», «*иконический взгляд*», «*сопровождение жеста*», «*взгляд на объект разговора*». В общей сложности выделено 1875 аннотаций для движений глаз и 1415 аннотаций функциональной разметки. Результаты разметки представлены в таблице 1.

Исходя из полученных данных можно заключить, что информанты 36 % всего времени находились в размышлении: отводили глаза в левую (38 %) или в правую (23 %) стороны. Во время разговора (20 % от всех движений глаз) информанты в 100 % случаев поддерживали зри-

Таблица 1. **Функциональная разметка движений глаз X**

Функция	Направление							
	к собеседнику	вверх	слева	вниз	справа	закрыты	к объекту	сумма
размышление	0	107	197	92	117	0	0	513
иконический взгляд	29	20	39	32	51	18	10	199
перечисление	1	3	11	4	5	0	0	24
шутка	2	3	5	1	7	0	0	18
смягчение антисоциальной ситуации	0	1	12	13	17	0	0	43
говорение	286	0	0	0	0	0	0	286
фразовое ударение	140	0	0	0	0	0	0	140
слушание	98	0	0	0	0	0	0	98
внимание на реакцию собеседника	68	0	0	0	0	0	0	68
подражание направлению взгляда собеседника	0	0	0	0	0	0	12	12
сопровождение жеста	0	0	0	7	3	0	0	10
взгляд на объект разговора	0	0	0	0	4	0	0	4
1415								

тельный контакт. Во время перечисления испытуемые в 46 % случаев смотрят налево, в 21 % — направо, в 17 % — вниз и в 13 % — вверх. Во время шутки взгляд респондентов также может быть направлен налево в 28 % случаев, направо — в 41 % случаев, вверх — в 17 % и вниз — в 6 %. Только иконический взгляд может выполняться в разных направлениях, к этому классу мы также отнесли демонстративное закрытие глаз (9 % случаев). Взгляд информанта в 100 % случаев направлен на собеседника при говорении, фразовом ударении, слушании, а также при обращении внимания на реакцию собеседника.

На основе разметки была разработана компьютерная модель на языке Python, регулирующая глазодвигательное поведение робота.

В этой модели были смоделированы коммуникативные состояния, соответствующие функциям, перечисленным выше. Эффективность этой модели в ситуации человеко-машинного взаимодействия была доказана в экспериментальном исследовании [Цфасман и др., 2018]. Полученные результаты позволяют судить о вкладе направления взгляда в формирование положительного впечатления от робота, а также доказывают эффективность разработанной модели глазодвигательного поведения в ситуации человеко-машинного взаимодействия.

1.2. Указательные жесты

Базовая разметка движений рук включает разметку на четырех слоях: это активный и пассивный органы, способ выполнения движения и траектория. В корпусе встречается 1165 случаев, когда информанты указывают на объект, на себя или на оппонента. Как правило, указательные жесты выполняются указательным пальцем (59,1 %) или ладонью/кистью руки (20,3 %) (Рис. 1)

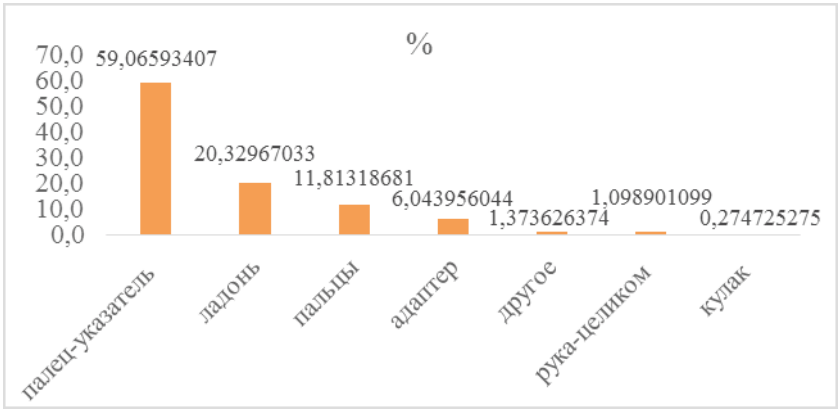


Рис. 1. **Активные органы, с помощью которых выполняются указательные жесты (в процентах)**

Указательные жесты обслуживают разные коммуникативные функции, описанные в [Котов, Зинина, 2015]. Как видно из Рис. 2, указательные жесты используются для апелляции (41,2 %) и сопровождают эмфазу (29,4 %), задействуются при ожидании обратной связи (8,8 %) и операциях с референтами (8,8 %).

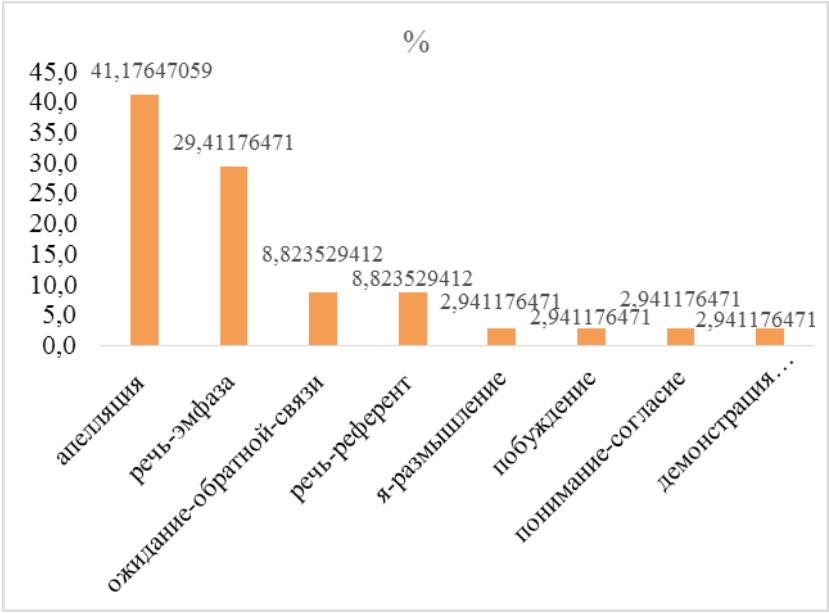


Рис. 2. **Коммуникативные функции указательных жестов**

1.3. Экспериментальное исследование указательных жестов робота

Для оценки влияния указательных жестов на формирование у пользователя положительного впечатления от робота был проведен эксперимент, в котором робот помогает испытуемому собирать головоломку (танграм) [Зинина и др., 2019]. В одном экспериментальном условии робот указывает человеку на необходимую фигуру головоломки (выполняет ориентированный жест), в другом — не указывает (выполняет неориентированные жесты). Экспериментальное взаимодействие робота и человека фиксировалось на две видеокамеры: первая записывала экспериментальную ситуацию сбоку, вторая записывала игровое поле и робота сверху. Взаимодействие пользователя с роботом было записано для анализа коммуникативных ответов человека на ориентированные жесты робота (31 испытуемый, 310 минут).

Разницу между экспериментальными условиями заметили 48,4 % от всей выборки, из которых подавляющее большинство (73,3 %) предпочло робота, демонстрирующего указательные жесты. Даже

если испытуемые не заметили разницы между двумя стратегиями жестового поведения робота (51,6 % от всей выборки отметили в анкете, что не увидели разницы в поведении), они, тем не менее, следовали указательным жестам робота и выбирали именно ту фигуру, на которую указал робот, в 78,5 % случаев. Следовательно, можно говорить об имплицитном влиянии ориентированных жестов робота на поведение пользователя, даже если испытуемые четко не осознавали этого влияния. Поскольку оценки человека являются имплицитными и в половине случаев не проявляются в самоотчёте, реакции человека на поведение робота в эксперименте удобно изучать корпусными методами, записывая и размечая поведение испытуемых.

Выводы

Руководствуясь преимущественно прикладной задачей, мы развиваем корпус REC, при этом особое внимание уделяем ориентированным жестам и функциональному значению взглядов, ведь это ключевые особенности, способствующие повышению привлекательности робота для пользователя. Видеозаписи экспериментального исследования открывают возможности для изучения стратегий взаимодействия человека с роботом, например, реакций человека на одобрение роботом его действий, или, наоборот, на сообщение робота об ошибке человека. Анализ видеозаписей также позволяет исследовать разные стратегии взаимодействия человека с роботом, что важно как для теоретического анализа, так и для практического применения.

Литература

1. *Cabibihan JJ., So W. C., Nazar M., Ge S. S.* (2009), Pointing Gestures for a Robot Mediated Communication Interface // Intelligent Robotics and Applications. ICIRA 2009. Lecture Notes in Computer Science, Vol. 5928.
2. *Gamer M., Hecht H.* (2007), Are you looking at me? Measuring the cone of gaze. Journal of Experimental Psychology, Vol. 33, pp. 705–715.
3. *Häring M., Eichberg J., André E.* (2012), Studies on Grounding with Gaze and Pointing Gestures in Human-Robot-Interaction // Social Robotics. ICSR 2012. Lecture Notes in Computer Science, Vol. 7621.
4. *Kanan C., Bseiso D. N. F., Ray N. A., Hsiao J. H., Cottrell G. W.* (2015), Humans have idiosyncratic and task-specific scanpaths for judging faces // Vision Research, Vol. 108, pp. 67–76.
5. *Kibrik A. A., Fedorova O. V.* (2018), Language production and comprehension in face-to-face multichannel communication // Компьютерная лингвистика и интеллектуальные технологии. Вып. 17 (24), с. 305–316.

6. *Kotov A. A., Arinkin N. A., Zaydelman L. Y., Zinina A. A.* (2019), Linguistic Approaches to Robotics: From Text Analysis to the Synthesis of Behavior, Language, Music and Computing // LMAC 2017. Communications in Computer and Information Science, Vol. 943, pp. 207–214.
7. *Rogers S. L., Speelman C. P., Guidetti O., Longmuir M.* (2018), Using dual eye tracking to uncover personal gaze patterns during social interaction // Scientific Reports, Vol. 8, Article number: 4271
8. *Salem M., Kopp S., Wachsmuth I.* (2012), Generation and Evaluation of Communicative Robot Gesture // International Journal of Social Robotics, Vol. 4(2), pp. 201–2017.
9. *Зинина А. А., Аринкин Н. А., Зайдельман Л. Я., Котов А. А.* (2019), Роль ориентированных жестов при коммуникации робота с человеком // Компьютерная лингвистика и интеллектуальные технологии. В печати.
10. *Котов А. А., Зинина А. А.* (2015), Функциональный анализ невербального коммуникативного поведения. Компьютерная лингвистика и интеллектуальные технологии. Вып. 14. Т. 1. М.: РГГУ, с. 299–310.
11. *Цфасман М. М., Аринкин Н. А., Зайдельман Л. Я., Зинина А. А., Котов А. А.* (2018), Разработка глазодвигательной коммуникативной системы робота Ф-2 на основе мультимодального корпуса. М.: Институт психологии РАН, с. 1328–1330.

References

1. *Cabibihan JJ., So W. C., Nazar M., Ge S. S.* (2009), Pointing Gestures for a Robot Mediated Communication Interface // Intelligent Robotics and Applications. ICIRA 2009. Lecture Notes in Computer Science, Vol. 5928.
2. *Gamer M., Hecht H.* (2007), Are you looking at me? Measuring the cone of gaze. Journal of Experimental Psychology, Vol. 33, pp. 705–715.
3. *Häring M., Eichberg J., André E.* (2012), Studies on Grounding with Gaze and Pointing Gestures in Human-Robot-Interaction // Social Robotics. ICSR 2012. Lecture Notes in Computer Science, Vol. 7621.
4. *Kanan C., Bseiso D. N. F., Ray N. A., Hsiao J. H., Cottrell G. W.* (2015), Humans have idiosyncratic and task-specific scanpaths for judging faces // Vision Research, Vol. 108, pp. 67–76.
5. *Kibrik A. A., Fedorova O. V.* (2018), Language production and comprehension in face-to-face multichannel communication // Компьютерная лингвистика и интеллектуальные технологии. Вып. 17 (24), с. 305–316.
6. *Kotov A. A., Arinkin N. A., Zaydelman L. Y., Zinina A. A.* (2019), Linguistic Approaches to Robotics: From Text Analysis to the Synthesis of Behavior, Language, Music and Computing // LMAC 2017. Communications in Computer and Information Science, Vol. 943, pp. 207–214.
7. *Rogers S. L., Speelman C. P., Guidetti O., Longmuir M.* (2018), Using dual eye tracking to uncover personal gaze patterns during social interaction // Scientific Reports, Vol. 8, Article number: 4271
8. *Salem M., Kopp S., Wachsmuth I.* (2012), Generation and Evaluation of Communicative Robot Gesture // International Journal of Social Robotics, Vol. 4(2), pp. 201–2017.

9. *Zinina A. A., Arinkin N. A., Zajdel`man L. Ya., Kotov A. A.* (2019), Rol` orientirovanny`x zhestov pri kommunikacii robota s chelovekom [The role of oriented gestures during robot's communication to a human] // Komp'juternaja lingvistika i intellektual'nye tehnologii [Computer linguistics and intellectual technologies]. In press.
10. *Kotov A., Zinina A.* (2015), Funkcional'nyj analiz neverbal'nogo kommunikativnogo povedenija [Functional analysis of nonverbal communicative behavior]. Komp'juternaja lingvistika i intellektual'nye tehnologii [Computer linguistics and intellectual technologies] Issue 14. Vol. 1. Moscow.: RSUH, pp. 299–310.
11. *Tsfasman M. M., Arinkin N. A., Zajdel`man L. Ya., Zinina A. A., Kotov A. A.* (2018), Razrabotka glazodvigatel`noj kommunikativnoj sistemy` robota F-2 na osnove mul`timodal`nogo korpusa [Developing eye gaze system for the F-2 robot based on a multimodal corpus]. M.: Institute of Psychology of RAS, pp. 1328–1330.

Зинина Анна Александрвона
Курчатовский исннтитут, Москва, Россия
Zinina Anna
Kurhcatov Insititute, Moscow, Russia
E-mail: Zinina_aa@nrcki.ru

Е. И. Риехакайнен
E. I. Riekhakaynen

МЕТОДИКА СОЗДАНИЯ КОРПУСА ДЛЯ ИЗУЧЕНИЯ РЕДУЦИРОВАННЫХ РЕАЛИЗАЦИЙ В ДЕТСКОЙ РЕЧИ[1]

HOW TO DEVELOP A CORPUS FOR STUDYING REDUCED REALIZATIONS IN CHILDREN'S SPEECH?

Аннотация. В статье описываются принципы сбора данных для корпуса устной речи русскоязычных детей 3–6 лет, который будет использоваться для изучения фонетической редукции словоформ в детской речи. Корпус состоит из трех частей: 1) записи, полученные в ходе лонгитюдного исследования с участием нескольких детей 3–5 лет; 2) записи, полученные в рамках эксперимента с участием 71 ребенка 4–6 лет; 3) фрагменты разработанных ранее корпусов детской речи. В программе Praat ведется сплошная орфографическая и выборочная фонетическая расшифровка записей.
Ключевые слова: детская речь, редуцированные словоформы, корпус устной речи, русский язык, лонгитюдное исследование.

Abstract. The paper introduces a corpus of oral speech of Russian-speaking children aged from 3 to 6 years. The corpus will be used to study phonetic reduction of word forms in children's speech. The corpus includes: 1) the records collected during a longitudinal study of several children between 3 and 5 years old; 2) the records collected in the experiment with 71 children between 4 and 6 years old; 3) fragments of several children's speech corpora that were developed earlier. The Praat program is used for the complete orthographic and selective phonetic annotation of the data.
Keywords: children speech, reduced word forms, spoken corpus, Russian, longitudinal study.

1. Введение

Несмотря на многочисленные фонетические и психолингвистические исследования, направленные на изучение явления фонетической редукции в различных языках, статус редуцированных словоформ в ментальном лексиконе говорящего и слушающего и механизмы их обработки до сих пор не ясны. В последнее время возрос интерес к вопросу усвоения редуцированных единиц, ответ на который предполагает выяснение того, в каком количестве и каким образом редуцированные реализации попадают в ментальный лексикон носителя языка и как происходит овладение механизмами редуцирования при усвоении языка.

[1] Исследование выполняется при поддержке гранта Президента РФ для молодых кандидатов наук № МК-6776.2018.6.

Чаще всего для ответа на поставленные вопросы исследователи проводят эксперименты в области изучения иностранных языков (second language acquisition) или изучают речь детей школьного возраста [Barth 2015; Tuomainen et al. 2015]. Как кажется, одной из основных причин игнорирования данных речи информантов более младшего возраста может быть отсутствие подходящего материала для исследования и сложность его получения. Однако именно речь детей 3–6 лет представляет наибольший интерес для изучения редукции, потому что в этот период происходит формирование развернутой спонтанной речи и можно предполагать появление механизмов редукции, которые свойственны взрослым носителям языка.

В работе описывается методика записи и обработки детской речи, в результате применения которой в настоящий момент создается корпус устной речи русскоязычных детей 3–6 лет.

2. Методика сбора данных

Наиболее адекватным и часто используемым методом сбора данных для изучения редукции в речи взрослых является запись речи, максимально приближенной к спонтанной (см., например, [Ernestus 2000; Raeva, Riekhakaynen 2016; Стойка 2017 и др.]). На сегодняшний день есть несколько общедоступных корпусов русской устной речи, в которых имеется фонетическая транскрипция записей (как в Корпусе транскрибированных русских устных текстов — http://narusco.ru/search/trn-search.php) или есть возможность прослушать и скачать звуковые файлы для последующей фонетической расшифровки (как в Мультимедийном подкорпусе Национального корпуса русского языка — http://ruscorpora.ru/search-murco.html или в проекте «Рассказы о сновидениях и другие корпуса звучащей речи» — http://spokencorpora.ru/). Наиболее представительным корпусом русской неподготовленной устной речи является корпус «Один речевой день» (см., например, [Asinovsky et al. 2009]), однако он пока не является общедоступным.

В качестве примеров корпусов русской детской речи большого объема можно отметить корпусы «INFANT.RU», «CHILD.RU» и «EmoChildRu» [Ляксо и др. 2017], но они также не являются общедоступными. Недавно появился доступ к корпусу «Кондуит» (http://konduitcorpus.ru), но только к орфографическим расшифровкам. Таким образом, любой исследователь, который начинает заниматься из-

учением звучащей детской речи на материале русского языка, так или иначе сталкивается с необходимостью сбора материала, т. е. составления собственного рабочего корпуса.

На первом этапе исследования по результатам пилотных записей одного ребенка (девочка, монолингв, возраст на момент проведения записей: 4 г. 9 мес. 0 дн. — 4 г. 11 мес. 7 дн.) было решено, что в домашних условиях родители детей могут осуществлять запись с помощью установленной на их телефоны бесплатно распространяемой программы «Диктофон» (https://play.google.com/store/apps/details?id=com.appstar.audiorecorder; параметры записи: 44100 Гц; 16 бит). С точки зрения изучения редукции наиболее перспективным, судя по предварительному слуховому анализу полученных данных, является рассмотрение реализаций высокочастотных словоформ и формул вежливости, а также сохранности окончаний существительных и согласованных с ними зависимых слов.

С июня 2018 года ведется лонгитюдное исследование — запись 4 детей. Родители записывают речь детей самостоятельно с интервалом около одного раза в месяц и высылают файлы с записями исследователям. Одна сессия предполагает около 30 минут записи в течение одного дня или за два следующих друг за другом дня. Для того чтобы собрать тематически более или менее однородный материал, родителям было дано задание записывать разговоры с детьми о том, что произошло за день (что ребенок делал в детском саду, дома и т. п.), а также общение ребенка с родителями во время совместных игр и за столом.

Вторым способом сбора материала стал эксперимент, направленный на получение данных о том, как дети произносят высокочастотные словоформы и формулы вежливости. В исследовании принял участие 71 ребенок из средних и старших групп двух детских садов Санкт-Петербурга[2]. Эксперимент проводился с каждым ребенком отдельно и длился в среднем от 5 до 10 минут. Основная часть эксперимента представляла собой ролевую игру «Магазин», в которой ребенок выступил и в роли продавца, и в роли покупателя. Кроме того, экспериментатор задавал ребенку вопросы о том, что ребенок делает в течение дня в детском саду и о любимых игрушках. Реплики экспериментатора были составлены так, что провоцировали ребенка на употребление ряда высокочастотных единиц (например, *только,*

[2] За проведение этого эксперимента автор выражает благодарность студентке СПбГУ Полине Вадимовне Шаньгиной.

когда, потому что, если, сейчас и др.). При составлении вопросов мы ориентировались на список высокочастотных единиц, для которых известно, что они часто подвергаются редукции в речи взрослых носителей русского языка (см, например, [Raeva, Riekhakaynen 2016]).

Родители всех участников исследования подписывали письменное согласие на запись речи их детей.

3. Структура корпуса и перспективы его развития

В соответствии с описанными выше принципами получения материала корпус на данный момент включает себя две основные части: 1) лонгитюдные записи 4 детей (возраст самого младшего на момент начала записи — 3 года 1 месяц; возраст самой старшей — 5 лет 6 месяцев); более 8 часов записей разговоров детей с их родственниками; 2) около 10 часов записей разговоров, полученных в ходе проведения эксперимента в детских садах, —71 файл по количеству детей, принявших участие в эксперименте.

Все звуковые файлы представлены в формате WAV. В настоящее время ведется многоуровневое аннотирование в программе Praat (http://www.fon.hum.uva.nl/praat/), поскольку поисковую систему в корпусе планируется организовать так, как это сделано в Фонетически транскрибированном корпусе эстонской спонтанной речи (https://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech). Файлы с расшифровками включают в себя следующие уровни: уровень диктора, на котором отмечается, кто говорит — ребенок или его собеседник; уровень орфографической расшифровки; уровень фонетической транскрипции.

На уровне орфографической расшифровки каждый интервал соответствует отдельному слову (или т.н. «составному слову», например: *потому что, то есть и* др.), все слова (в том числе те, которые встретились в редуцированном виде) записываются в нормативной орфографии. Редукция отмечается только на уровне фонетической транскрипции, на котором осуществляется позвуковая разметка. Подобный подход обусловлен тем, что в дальнейшем на материале этих записей планируется составить орфографически-транскрипционный словарь, в котором каждой орфографической единице будут соответствовать все варианты ее произнесения.

Акустико-фонетическая расшифровка ведется по принципам, разработанным для Корпуса транскрибированных русских устных

текстов: различаются не только все фонемы русского языка, но и перцептивные варианты гласных, ударение не размечается (используемые символы транскрипции представлены здесь: http://narusco.ru/transkrip.htm).

На данный момент лонгитюдные записи расшифровываются полностью и в орфографии, и в транскрипции. Для записей, полученных в ходе эксперимента, пока проводится только выборочная фонетическая аннотация: в ходе сплошного прослушивания (вручную) выбираются и транскрибируются только интересующие нас высокочастотные словоформы и формулы вежливости. Параллельно с расшифровкой создается база звуковых файлов, в которой представлены вырезанные из исходных файлов реализации анализируемых высокочастотных словоформ. В названии каждого файла содержится не только указание на то, какая словоформа в нем представлена, но и условное обозначение информанта, что позволяет изучать, как один и тот же ребенок произносит одну и ту же словоформу, если она представлена в его речи несколькими реализациями.

Третью часть корпуса, как ожидается, составят расшифровки фрагментов других корпусов детской речи, к которым нам удастся получить доступ. На настоящий момент в эту часть корпуса входят фрагменты корпуса «Кондуит», звуковые записи и орфографические расшифровки из которого были предоставлены нам его составителем П. М. Эйсмонт. Орфографические расшифровки из этого корпуса проверены и перенесены в файлы многоуровневой разметки, которые используются в Praat, ведется акустико-фонетическая расшифровка данных записей.

Полноценный анализ собранного материала будет возможен только после завершения расшифровки, но даже предварительные результаты и наблюдения, сделанные в ходе расшифровки и транскрибирования, свидетельствуют о том, в речи детей 3-6 лет уже присутствуют редуцированные реализации, часть из которых (в первую очередь те, в которых редукции подвергаются один-два элемента: например, [stoka] для *столько*, [kada] для *когда* и т. п.) совпадает с теми, которые являются наиболее частотными в речи взрослых носителей русского языка.

Литература

1. *Ляксо Е. Е., Фролова О. В., Григорьева А. С., Остроухов А. В.* (2017), Корпуса детской речи «INFANT.RU», «CHILD.RU», «EmoChildRu» на материале русского языка и использование в исследованиях речевого онтогенеза, Теоретическая и прикладная лингвистика, 3 (1), с. 28–58.
2. *Стойка Д. А.* (2017), Редуцированные формы русской речи: лингвистический и экстралингвистический аспекты: дис. … канд. филол. наук. СПб.
3. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation, V. Matoušek & P. Mautner (eds.), Text, Speech and Dialogue (Lecture Notes in Computer Science 5729). Berlin, Heidelberg: Springer, pp. 250–257.
4. *Barth D. G.* (2015), To have and to be: Function word reduction in child speech, child directed speech and inter-adult speech, PhD thesis. The University of Oregon Graduate School.
5. *Ernestus M.* (2000), Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface, PhD Thesis, Utrecht.
6. *Raeva O., Riekhakaynen E.* (2016), Frequent Word Forms in Spontaneous Russian: Realization and Recognition, Linguistica Lettica, 24, pp. 122–139.
7. *Tuomainen O., Lee Ch., Granlund S., Hazan V.* (2015), Phonetic Reduction in Spontaneous Speech by Children Aged 9–14 Years, Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK. Paper number 0412.1-5 retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0412.pdf

References

8. *Lyakso E. E., Frolova O. V., Grigorjeva A. S., Ostroukhov A. V.* (2017), Korpusa detskoj rechi «INFANT.RU», «CHILD.RU», «EmoChildRu» na materiale russkogo jazyka i ispol'zovanie v issledovanijakh rechevogo ontogeneza [Children's Speech Corpora «INFANT.RU», «CHILD.RU», «EmoChildRu» on Russian and their Application for Studying First Language Acquisition], Teoreticheskaja i prikladnaja lingvistika [Theoretical and Applied Linguistics], 3 (1), pp. 28–58.
9. *Stojka D. A.* (2017), Redutsirovannye formy russkoj rechi: lingvisticheskij I ekstralingvisticheskij aspekty [Reduced Forms in Russian Speech: Linguistic and Extralinguistic Aspects], PhD Thesis, St. Petersburg.
10. *Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T.* (2009), The ORD speech corpus of Russian everyday communication "One Speaker's Day": creation principles and annotation, V. Matoušek & P. Mautner (eds.), Text, Speech and Dialogue (Lecture Notes in Computer Science 5729). Berlin, Heidelberg: Springer, pp. 250–257.
11. *Barth D. G.* (2015), To have and to be: Function word reduction in child speech, child directed speech and inter-adult speech, PhD thesis. The University of Oregon Graduate School.

12. *Ernestus M.* (2000), Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface, PhD Thesis, Utrecht.
13. *Raeva O., Riekhakaynen E.* (2016), Frequent Word Forms in Spontaneous Russian: Realization and Recognition, Linguistica Lettica, 24, pp. 122–139.
14. *Tuomainen O., Lee Ch., Granlund S., Hazan V.* (2015), Phonetic Reduction in Spontaneous Speech by Children Aged 9–14 Years, Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK. Paper number 0412.1-5 retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0412.pdf

---

Риехакайнен Елена Игоревна
Санкт-Петербургский государственный университет (Россия)
Riekhakaynen Elena
Saint Petersburg State University (Russia)
E-mail: e.riehakajnen@spbu.ru

Т. В. Тимкин
T. V. Timkin

ПРИМЕНЕНИЕ КОРПУСНОГО ПОДХОДА ПРИ ИССЛЕДОВАНИИ ФОНЕТИКИ СУРГУТСКОГО ДИАЛЕКТА ХАНТЫЙСКОГО ЯЗЫКА[1]

THE CORPUS APPROACH IN THE RESEARCH OF THE PHONETICS OF THE SURGUT KHANTY LANGUAGE

Аннотация. В докладе описаны лингвистические и технические аспекты создания фонетического корпуса для исследования сургутского диалекта хантыйского языка с применением корпусной системы Emu SDMS. Корпус включает аннотированные аудиозаписи речи информантов, поддерживает поиск по разметкам и вычисление акустических данных для дальнейшей обработки в программе R. В качестве примера использования корпуса приводятся данные о длительности сургутских гласных первого слога.
Ключевые слова: экспериментальная фонетика, фонология, корпусная лингвистика, Praat, Emu-SDMS, хантыйский язык, сургутский диалект.

Annotation. The report considers linguistic and technical aspects of the building of the phonetic corpus for the research of the Surgut dialect of the Khanty language. Corpus system Emu Speech Database Management System is used. The corpus contains annotated audio files of speakers' speech, provides search within annotation and acoustic data evaluation for further processing in R. The data on first syllable vowel duration are given as an example of corpus application.
Keywords: experimental phonetics, phonology, corpus linguistics, Praat, Emu-SDMS, Khanty language, Surgut dialect.

Введение

Исследование фонетики языков коренных народов Сибири является не только актуальной проблемой ареального и типологического языкознания, но и представляет практическую необходимость. Без точных экспериментальных данных невозможно решение вопроса об оптимальной транскрипции, что, в свою очередь, связано с проблемами совершенствования орфографии, издания текстов и лексикографической фиксации.

Подробное экспериментальное описание хантыйского консонантизма и вокализма выполнено в работах [Верте, 2003], [Куркина, 2000] на материале казымского диалекта. Однако сургутский диалект, име-

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 19-012-00388)

ющий значительные отличия в фонологической системе, до сих пор не подвергался инструментальному фонетическому анализу. Фонология диалекта была описана без привлечения экспериментальных данных в ряде источников по хантыйскому языку (например, [Терешкин, 1981]). В монографии [Чепреги, 2017] помимо фонологии диалекта рассмотрены отдельные детали произношения. В серии работ [Уртегешев, Кошкарева, 2017а], [Уртегешев, Кошкарева, 2017б] представлено наиболее подробное на сегодня описание сургутской артикуляционной базы идиома, однако эти работы, выполненные с использованием субъективно-слуховых методик, не снимают необходимость экспериментального исследования.

Важность многоаспектного исследования фонетики, отсутствие непротиворечивых исходных данных, ограниченность ресурсов для полевой работы приводят нас к выбору корпусного исследования в качестве основного метода. Корпусный подход в фонетике подразумевает такую процедуру исследования, при которой аудиозаписи подвергаются комплексной разметке, а отбор сегментов для исследования отдельного явления осуществляется методами машинного поиска, причем система должна поддерживать учет метаданных, иерархической организации сегментов и синтагматики.

Речевые корпусы разрабатываются с середины 1980-х гг. в первую очередь для прикладных задач синтеза и распознавания речи, но также являются одним из ключевых инструментов современной экспериментальной фонетики: “In some respects corpus methods complement laboratory-based experimental methods in phonology,and for some fields of inquiry corpus materials are essential” [Cole, Hasegawa-Johnson, 2012: 431]. Создание подобных корпусов для аборигенных языков Сибири является актуальной задачей.

В настоящей работе мы осветим опыт применения корпусных технологий к исследованию фонетики сургутского диалекта хантыйского языка, распространенного преимущественно на территории Сургутского района ХМАО и насчитывающего почти три тысячи носителей [Чепреги, 2016].

Корпус включает полевые материалы Института филологии СО РАН, собранные в ходе экспедиций 2014–2018 гг. На данный момент объем корпуса составляет около двухсот лексем, каждая из которых записана в троекратном прочтении минимум от трех информантов — носителей исследуемого диалекта. На данный момент в базу включены

данные от шести информантов. Общее число записанных фонетических слов составляет более 1500.

Основной инструмент работы с корпусом — фонетическая корпусная система Emu Speech Database Management System (продукт Мюнхенского университета), включающая корпус-менеджер Emu-WebApp и пакет расширений EmuR для статистического обработчика R [Harrington, 2010], [Winkelmann, Harrington, Jänsch, 2017]. Разметка производится в программе Praat [Boersma, Weenink, 2018]. Обработка данных осуществляется базовыми средствами R и средствами EmuR. Для визуализации данных используются средства пакетов EmuR, ggplot2, ggrepel для R.

1. Лингвистические аспекты построения корпуса

1.1. Основная единица разметки

Основная проблема при атрибуции фонетических единиц заключается в том, что нет единого свода объективных акустических признаков, которые могли бы охватить все разнообразие голосов дикторов и были бы универсальны для всех языков. Мы видим проблему также в том, что значительное свободное варьирование признаков, комбинаторные изменения, различия между дикторами создают континуум тембров, внутри которого крайне трудно провести границы. Разброс данных, многообразие оттенков, релевантность которых заранее неизвестна, затрудняют последовательное применение этапов анализа «фонетическая транскрипция — инвентаризация фонетических средств — анализ оппозиций и дистрибуции звуков — выявление системы фонем».

При данном состоянии изученности материала мы вынуждены опираться не на абсолютные значения акустических параметров звуков, а на их соотношения, а также прибегать к поэтапному уточнению разметки.

Разметчик производит предварительную атрибуцию звука на основе субъективного слухового анализа, учитывая также орфографическую запись и традиционную фонологическую транскрипцию. На данном этапе мы должны предполагать, что разметка может не соответствовать объективным данным. Такая разметка характеризует, таким образом, не звуки речи, а фонетические токены — предварительно приписанные фонетические атрибуты, которые нуждаются в верификации. Одинаково размеченные сегменты организуют кластеры, для

которых собирается информация о средних значениях и варьировании акустических параметров. На основании этих данных выявляются ошибки и неточности разметки, и атрибуция сегмента может быть изменена, после чего сводные данные пересчитываются. В ходе такой постепенной коррекции мы добиваемся следующего: периферийные пересечения кластеров должны быть минимизированы; варьирование звуков внутри кластера может быть признано свободным; звуки, для которых возможно установить позиционное варьирование, имеют различную разметку. После этого мы можем считать, что разметка является уже не условным токеном, а фонетической транскрипцией, фиксирующей аллофон.

Для разметки используются знаки IPA. Система транскрипции сургутских фонем была выработана на этапе до составления корпуса путем предварительного анализа спектрограмм.

1.2. Обозначение акцентных и суперсегментных признаков

Ударение в хантыйском языке, как правило, падает на первый, всегда корневой слог, который характеризуется повышенной длительностью и интенсивностью [Чепреги, 2017: 37]. Безударные гласные характеризуются значительной редукцией, их фонологический статус не вполне ясен. Помимо этого, не выяснено, как распределяются признаки долготы и интенсивности в синтагмах, состоящих из нескольких фонетических слов; неясно, являются ли клитиками личные местоимения в притяжательных конструкциях.

По этой причине при разметке ударение обозначается обязательно, причем используется отдельно разметка «основное ударение» для изолированных словоформ и «побочное ударение» для обозначения ударных слогов в синтагмах.

Акцентная характеристика не выносится в отдельный уровень разметки, а объединена с фонетической транскрипцией.

Качество и длительность звуков в хантыйском языке находятся в зависимости от слоговой структуры словоформы. Тем не менее, специальной суперсегментной разметки на данном этапе не проводится, поскольку наиболее релевантное разграничение ударного и безударного слога обеспечивается обязательной постановкой ударения в транскрипции. Другие аспекты слоговой структуры, такие как количество слогов, противопоставление серединного и конечного слога, противопоставление открытого и закрытого слога на данном возможно учитывать только ручной фильтрацией поисковой выдачи.

2. Технические аспекты построения корпуса

2.1. Разметка и загрузка файлов в корпус

Корпус-менеджер Emu-WebApp позволяет работать с разметками аудио, однако основным инструментом разметки файлов была выбрана программа Praat. В первую очередь, это связано с тем, что Praat имеет режим работы с длинными аудио, что позволяет размечать записи длительных полевых сессий с информантами без предварительной нарезки файлов на фрагменты, соответствующие отдельным словоформам.

На этапе предварительной аннотации возможна параллельная работа нескольких разметчиков над базой, при этом каждый разметчик работает над своим аудиофайлом.

Для хранения метаданных выделено две дорожки разметки, содержащих код информанта, перевод слова и его орфографическую запись. Границы разметки на этих уровнях служат метками для автоматического нарезания файлов на фрагменты.

Для того чтобы избежать возможных конфликтов с кодировкой, знаки IPA вводятся с использованием кодов Backslash Trigraphs, которые не включают символов Unicode. Перед непосредственной загрузкой файлов в корпус дорожка фонетической разметки дублируется, и одна из копий автоматически конвертируется в кодировку Unicode с использованием скрипта Praat. Таким образом, IPA-разметки хранятся в базе в двух экземплярах: в кодировке Unicode и с использованием Backslash Trigraphs.

Нарезка аудио и разметок на фрагменты, соответствующие отдельным словоформам и экспорт файлов производится автоматически при помощи скрипта Praat.

Конвертирование данных в формат системы Emu-SDMS производится средствами EmuR. Система поддерживает возможность загрузки новых данных в существующий корпус и объединения корпусов.

Для получения акустических данных (спектрограмм, формантных частот, частоты основного тона и т. д.) используются средства EmuR. Вычисленные данные сохраняются в корпусе в файлах особого формата и доступны для дальнейшей статистической обработки.

2.2. Организация поиска в корпусе

Для поиска используется функция query из пакета EmuR, позволяющая находить сегменты с заданной разметкой. Функция поддержи-

вает поиск на нескольких дорожках, и, поскольку код информанта на этапе разметки записывается в отдельную дорожку, мы можем производить поиск с учетом информанта. Также функция query позволяет производить поиск с учетом соседних сегментов, благодаря чему мы можем учитывать фонетическую позицию.

При поиске по транскрипциям необходимо учитывать отдельные фонетические признаки, такие как долгота или краткость гласного, огубленность, принадлежность звука к гласным или согласным и т. д. Это реализуется с применением инструмента label groups из пакета EmuR, который позволяет сгруппировать разметки (например, объединить в одну группу все сегменты, соответствующие гласным) и запрашивать при поиске не отдельные звуки, а целые группы.

По результатам запроса функции EmuR возвращают акустические данные как фреймы данных, с которыми далее можно работать средствами R для статистического анализа и визуализации.

Если анализ данных показывает необходимость корректировки разметок, обратный экспорт в Praat уже не производится, разметки исправляются через интерфейс Emu-WebApp.

3. Пример использования системы: длительность гласных первого слога в моносиллабах и бисиллабах

В исследованиях сургутского диалекта (например, [Терешкин, 1981]) описывается релевантность противопоставления долгих и кратких гласных первого слога. В монографии [Чепреги, 2017], однако, указывается необходимость выделения также сверхкратких гласных. В работах [Уртегешев, Кошкарева, 2017а], [Уртегешев, Кошкарева, 2017б] транскрибируются полудолгие гласные и прерывистые гласные, состоящие из двух артикуляционных фаз.

Рассмотрим в качестве примера применения корпуса задачу вычисления длительностей ударных гласных звуков в базовом стословном подкорпусе, содержащем моносиллабы и бисиллабы.

Корпус загружается для обработки как объект R при помощи функции load_emuDB.

С помощью поисковой функции query и механизма группировки разметок найдем все гласные, размеченные как ударные. При помощи функции dur вычислим абсолютную длительность найденных сегментов в миллисекундах. Запросим при помощи функции requery_hier код информанта для каждого сегмента. Воспользуемся функцией re-

query_seq, чтобы узнать для каждого сегмента разметку последующего сегмента. К данным о последующей разметке применим ряд условных конструкций, которые позволяют определить слоговой тип словоформы. Атрибуты слогового типа для каждого сегмента сохраним в отдельном векторе.

Полученные данные объедим в один фрейм, содержащий для каждого сегмента разметку, длительность в миллисекундах, фонологическую характеристику долготы или краткости, код информанта и тип слоговой структуры.

Для визуализации фрейма воспользуемся средствами ggplot2.

На рисунке 1 показаны графики разброса длительности гласных звуков, встречающихся в первом слоге, по данным одного информанта, в речи которого зафиксировано наиболее отчетливое произнесение долгих гласных. Каждая точка показывает отдельное произнесение звука информантом. В левой панели графика собраны фонологически долгие гласные, в правой — фонологически краткие. Цвет точки обозначает слоговой тип словоформы: красные точки соответствуют бисиллабам с закрытым первым слогом, синие — бисиллабам с открытым первым слогом, зеленые — односложным формам.

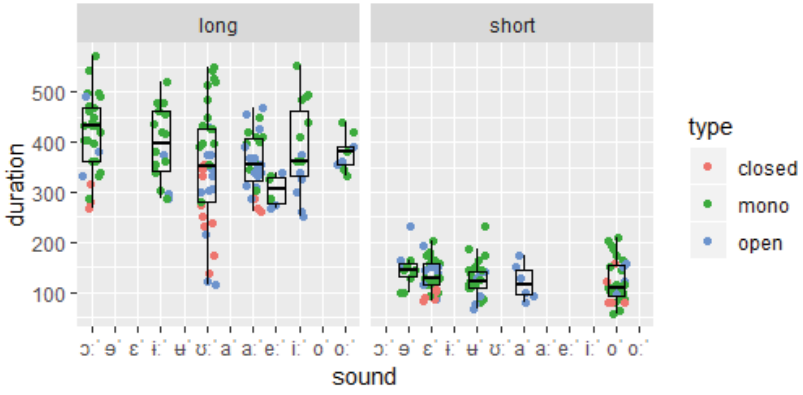


Рис. 1. **Разброс значения абслютной длительности гласных звуков по данным первого информанта**

Как показывают графики, в сургутском диалекте хантыйского языка гласные первого слога значительно варьируются по длительности. Наиболее длинные реализации фонем встречаются в моносиллабах;

в бисиллабах гласные сокращаются, причем сокращение значительно более выражено в закрытых слогах. Варьирование гласных приводит к пересечению диапазонов реализации фонем. Однако, если рассматривать отдельно словоформы каждого слогового типа, противопоставление долгих и кратких остается достаточно четким. На рисунке 2 показано распределение длительности долгих и кратких гласных в различных слоговых типах.

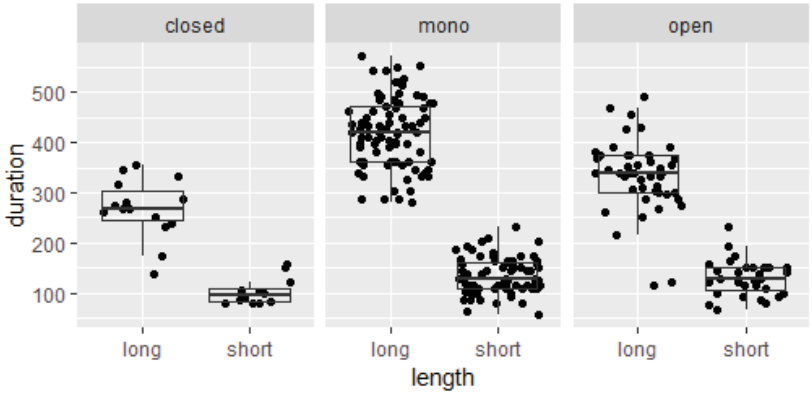


Рис. 2. **Разброс значения абсолютной длительности гласных звуков в различных слоговых структурах по данным первого информанта**

Другие информанты склонны к более краткому произнесению долгих согласных (менее 400 мс), из-за чего долгие гласные в позиции закрытого слога сближаются с краткими. В многосложных формах наблюдается тенденция к сокращению долгих гласных и их сближению с краткими.

Эти данные подтверждают точку зрения о двух, но не трех степенях фонологической долготы в сургутском диалекте, выделение сверхкраткости в первом слоге неоправданно.

Фиксируемая в работах [Уртегешев, Кошкарева, 2017а], [Уртегешев, Кошкраева, 2017б] полудолгота, судя по всему, является вариативной реализацией фонологической краткости в открытом слоге и соответствует нашим данным о растяжении кратких. Описанные в указанных работах прерывистость и вариативная дифтонгизация, очевидно, соотносятся с большей длительностью. Выделяемые в данных работах долгие непрерывистые гласные в моносиллабах находят соответствие

как кратким, так и долгим гласным в наших измерениях, однако прерывистые — только долгим. Вероятно, выделяемая слуховым анализом прерывистость влияет на восприятие длительности и участвует в различении долгих и кратких гласных.

4. Перспективы работы над корпусом

На данный момент корпус позволяет судить о варьировании количественных и качественных признаков отдельных фонем; мы можем верифицировать и корректировать фонетические транскрипции. На данном этапе не все случаи варьирования звуков и пересечения зон варьирования находят убедительное объяснение.

Представляется целесообразным расширять корпус, вводя в него новые записи.

Литература

1. *Верте Л. А.* (2003) Консонантизм хантыйского языка (экспериментальное исследование). Новосибирск: Сибирский хронограф.
2. *Куркина Г. Г.* (2000) Вокализм хантыйского языка (экспериментальное исследование). Новосибирск: Сибирский хронограф.
3. *Терешкин Н. И.* (1981) Словарь восточнохантыйских диалектов. Л.: Наука.
4. *Уртегешев Н. С., Кошкарева Н. Б.* (2017) Система долгих гласных звуков первого слога в сургутском диалекте хантыйского языка // Вестник угроведения. Т. 7, № 3 (30), 2017. С. 74–97.
5. *Уртешешев Н. С., Кошкарева Н. Б.* (2017) Система кратких гласных звуков первого слога в сургутском диалекте хантыйского языка // Вестник угроведения. Т. 7, № 4 (31), 2017. С. 70–85.
6. *Чепреги М.* (2017) Сургутский диалект хантыйского языка. Ханты-Мансийск.
7. *Boersma P., Weenink D.* (2018) Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/
8. *Cole J., Hasegawa-Johnson M.* (2012) Corpus phonology with speech resources // The Oxford handbook of laboratory phonology. Oxford University Press. P. 431–440.
9. *Harrington J.* (2010) Phonetic Analysis of Speech Corpora. Wiley-Blackwell.
10. *Winkelmann R., Harrington J., Jänsch K.* EMU-SDMS: Advanced speech database management and analysis in R // Computer Speech & Language. Vol. 45, P. 392–410.

Reference

1. *Boersma P., Weenink D.* (2018) Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/
2. *Chepregi M.* (2017) Surgutskij dialekt khantyiskogo jazyka [The Surgut dialect of the Khanty language], Khanty-Mansijsk.

3. *Cole J., Hasegawa-Johnson M.* (2010) Corpus phonology with speech resources // The Oxford handbook of laboratory phonology. Oxford University Press. P. 431–440.
4. *Harrington J.* (2010) Phonetic Analysis of Speech Corpora. Wiley-Blackwell.
5. *Kurkina G. G.* (2000) Vokalizm khantyjskogo jazyka (eksperimental'noe issledovanie) [The vowel system of the Khanty language (An experimental research)], Novosibirsk: Sibirski khronograf.
6. *Tereshkin N. I.* (1981) Slovar' vostochnokhantyjskikh dialektov [The dictionary of East Khanty dialects]. Leningrad: Nauka.
7. *Urtegeshev N. S., Koshkareva N. B.* (2017) Sistema dolgikh glasnykh zvukov pervogo sloga v surgutskom dialekte khantyjskogo jazyka [The system of the long vowels of the first syllable in Surgut Khanty] // Vestnik ugrovedenija [Bulletin of Ugric studies]. V. 3, № 3 (30), 2017. P. 74–97.
8. *Urtegeshev N. S., Koshkareva N. B.* (2017) Sistema kratkikh glasnykh zvukov pervogo sloga v surgutskom dialekte khantyjskogo jazyka [The system of the short vowels of the first syllable in Surgut Khanty] // Vestnik ugrovedenija [Bulletin of Ugric studies]. V. 7, № 4 (31), 2017. P. 70–85.
9. *Verte L. A.* (2003) Konsonantizm khantyjskogo jazyka (eksperimental'noe issledovanie) [The consonant system of the Khanty language (An experimental research)], Novosibirsk: Sibirski khronograf.
10. *Winkelmann R., Harrington J., Jänsch K.* EMU-SDMS: Advanced speech database management and analysis in R // Computer Speech & Language. Vol. 45, P. 392–410.

Тимкин Тимофей Владимирович
Институт филологии Сибирского отделения РАН (Новосибирск, Россия)
Timkin Timofej

Т. Ю. Шерстинова
T. Yu. Sherstinova

О ПОДГОТОВКЕ К ВЕБ-ПУБЛИКАЦИИ КОРПУСА ПОВСЕДНЕВНОЙ РУССКОЙ РЕЧИ «ОДИН РЕЧЕВОЙ ДЕНЬ»: АНОНИМИЗАЦИЯ ТЕКСТОВ И ВЫБОРОЧНОЕ КОДИРОВАНИЕ ЛЕКСИКИ[1]

ON THE PREPARATION FOR WEB-PUBLICATION OF "ONE DAY OF SPEECH" CORPUS OF EVERYDAY SPOKEN RUSSIAN: TEXTS ANONIMIZATION AND SELECTED WORDS ENCODING

Аннотация. Речевой корпус «Один речевой день» (ОРД) является на сегодняшний день крупнейшим лингвистическим ресурсом, предназначенным для исследования русского языка повседневного общения. Несмотря на высокий научный потенциал материалов корпуса, эффективность его использования до сих пор существенно ограничена фактом закрытости ресурса для широкого круга пользователей, что вызвано частным характером большинства звукозаписей повседневной речи. Компромиссным решением представляется веб-публикация анонимизированных текстовых расшифровок корпуса ОРД. В статье рассматриваются основные сложности, возникающие при подготовке текстовок корпуса ОРД к онлайн публикации, связанные с анонимизацией текстов и их «цензурной» редактурой, и намечаются пути их решения.
Ключевые слова. Русский язык, повседневная устная речь, речевой корпус, интернет-ресурс, онлайн публикация, анонимизация текстов, кодирование лексики.

Abstract. Speech corpus "One Day of Speech" (ORD corpus) is the largest linguistic resource designed for studies of everyday spoken Russian. Despite the high scientific potential of ORD data, the effectiveness of its use is still significantly limited by the fact that the resource is not accessible for a wide range of online users, which is caused by the private nature of the most of its audio recordings. The most suitable option appears to be the web publication of selected anonymized text transcripts. The article outlines the main difficulties that arise during the preparation of ORD texts to web publication, including texts anonymization and their "censorship" editing, and discusses the ways to solve these problems.
Keywords. Russian language, everyday spoken speech, speech corpus, Internet resource, online publication, texts anonymization, word coding.

1. Введение

Речевой корпус «Один речевой день» (ОРД) является на сегодняшний день крупнейшим лингвистическим ресурсом, предназначенным для исследования русского языка повседневного общения [Bogdanova-Beglarian et al. 2016]. Корпус содержит более 1400 часов аудиозаписей,

[1] Работа выполнена при поддержке Российского Научного Фонда «Система прагматических маркеров русской повседневной речи» (проект № 18-18-00242).

выполненных в условиях естественной речевой коммуникации; расшифровки получены для 530 макроэпизодов и насчитывают 1 млн. словоупотреблений. Корпус разрабатывается для изучения устной русской речи и речевого поведения человека, для проведения исследований в области антропологии, речевой коммуникации, устного дискурса, социолингвистики, психолингвистики, когнитивной лингвистики и других смежных дисциплин, а также для решения ряда прикладных задач — в частности, для поддержки систем автоматического мониторинга речи, голосового поиска, систем синтеза и распознавания речи, искусственного интеллекта, разработок голосовых диалоговых систем при общении человека с компьютером/роботом, для преподавания русского языка как иностранного, для проведения лингвистической и судебной экспертизы по аудиозаписям речевой коммуникации и т. п. В настоящее время на базе данного корпуса проводится многоаспектное исследование прагматических маркеров устной русской речи [Bogdanova-Beglarian et al. 2018].

Несмотря на высокий научный потенциал материалов корпуса, эффективность его использования до сих пор существенно ограничена фактом закрытости ресурса для широкого круга пользователей, что вызвано частным характером большинства звукозаписей повседневной речи, отнюдь не предназначенных для публичного доступа. Поэтому оригинальные звукозаписи корпуса ОРД не являются и, по-видимому, в обозримом будущем не смогут быть свободно распространяемым контентом.

Компромиссным решением представляется публикация в сети Интеренет анонимизированных текстовых расшифровок корпуса ОРД. В статье рассматриваются основные сложности, возникающие при подготовке текстовок корпуса к веб-публикации, связанные с анонимизацией текстов и их «цензурной» редактурой, и намечаются пути их решения.

2. Подготовка текстов корпуса ОРД к веб-публикации

2.1. Анонимизация персональных данных

Звукозаписи, представляющие собой текстовый массив корпуса ОРД, содержат преимущественно разговоры из частной жизни информантов. Анонимность участников звукозаписи (при открытости их социологических характеристик) — один из ключевых моментов методики сбора данных, который позволял участникам эксперимен-

та проводить свой речевой день наиболее естественно. Поэтому наиболее важным требованием к публикации материалов коллекции онлайн должно стать обеспечение сохранения анонимности авторства речевого материала.

Для исключения истинной или ложной атрибуции говорящего по акустическим свойствам его голоса и манере речи, сами звукозаписи повседневного общения, по-видимому, не могут быть опубликованы в свободном доступе. Что касается текстов расшифровок этих звукозаписей, то они могут быть обнародованы только при условии полной анонимизации личных имен, фамилий, прозвищ (их изменения на другие, вымышленные), а также исключения из текстов, представленных на сайте, любой другой информации, которая может повлечь раскрытие личности говорящего (номера телефон или паспорта, места работы и пр. информации, которая может быть озвучена в процессе «речевого дня»). Кроме того, по-видимому, текстовки не всех макроэпизодов речевой коммуникации могут быть опубликованы по этическим или иным соображениям. Отсюда, возникают следующие задачи:

1) определение типов эпизодов, которые не могут быть представлены на сайте ни в каком виде. К решению этой задачи будут привлечены квалифицированные юристы;
2) отбор коммуникативных эпизодов, которые могут быть опубликованы после их анонимизации;
3) осуществление анонимизации текстов: замена всей личной информации (в первую очередь имен и фамилий) на иные, но предпочтительно состоящие из того же количества слогов и с сохранением ритмической структуры (позиции ударения). Например, *Катя* → *Маша*%, *Юра Иванов* → *Даня*% *Королев*%). Анони-мизированные данные маркируются в транскриптах знаком %;
4) исключение из текстов расшифровок другой личной информации о говорящих или ее анонимизация.

2.2. Кодирование непечатной лексики

Характерной особенностью частного неформального речевого общения является активное использование субстандартной и непечатной лексики, особенно в речи отдельных социальных групп (преимущественно, в речи молодежи [Химик 2000] и в «мужских» разговорах [Потапова, Потапов 2006]). Этим определяются сложности представления текстов подобных эпизодов в открытом онлайн доступе.

Действительно, с одной стороны, «из песни слова не выкинешь», поэтому для изучения повседневной речевой коммуникации важно, чтобы исследовательский материал передавал по возможности все ее особенности. Игнорировать целый пласт неформальной речевой коммуникации, содержащий мат, невозможно вследствие высокой частоты этого явления, косвенным отражением чего можно считать высокий ранг отдельных непечатных слов в верхней зоне частотного словаря звукозаписей повседневного общения [Шерстинова 2016]. Более того, наблюдения над повседневной речью показывают, что непечатные слова зачастую выполняют не только свои специфические функции (экспрессивную, «протестную», «эпатирующую», маркирующую статус говорящего и др.), но и ряд других важных функций, присущих «стандартной» лексике — ритмообразующую [Богданова-Бегларян и др. 2013], делимитативную (дискурс-структурирующую) [Богданова-Бегларян 2014] и пр.

Согласно поправке № 53 от 2005 г. к Федеральному Закону «О государственном языке РФ», с 2014 г. введен ряд запретов на применение ненормативной лексики, в частности — на ее публикацию в печати и онлайн. При этом Роскомнадзор оставляет издателям возможность маскировки мата звездочками, если вводится цитата или это необходимо для сохранения художественного смысла [Рекомендации…]. Поскольку до сих пор отсутствует юридически утвержденный полный список нецензурных выражений, Роскомнадзор рекомендует издателям опираться на опубликованные словари ненормативной лексики (в частности, [Квеселевич 2011; Мокиенко, Никитина 2004; и др.]).

Таким образом, важным моментом подготовки расшифровок повседневной речи к онлайн публикации является «цензурная» редактура. Поскольку для лингвистических исследований желательно, чтобы по тексту расшифровки можно было однозначно восстановить произнесенный текст, использование звездочек (*), даже при сохранении начальной буквы текста, не является оптимальным методом ввиду высокой вариативности непечатной лексики. Альтернативным вариантом видится разработка более сложной системы кодирования, учитывающей в том числе частеречную принадлежность маскируемого слова. Разработка такой системы кодирования, как и получение рабочего списка «непечатных» единиц, являются одними из приоритетных задач, реше-ние которых необходимо для создания сайта корпуса ОРД.

3. Планируемые возможности сайта

Сайт корпуса ОРД планируется к публикации в свободном доступе на сайте СПбГУ. К концу 2020 г. будет опубликован анонимизированный фрагмент корпуса в объеме 300 тыс. словоупотреблений, содержащий представительную выборку эпизодов повседневного речевого общения для разных социальных групп говорящих, включающих как профессиональную, так и бытовую коммуникацию. В онлайн режиме будут доступны лишь текстовые расшифровки звукозаписей. Свободная публикация аудиозаписей в настоящее время не планируется. Пользовательский интерфейс онлайн версии корпуса будет обеспечивать:

а) текстовый поиск по расшифровкам звукозаписей по заданному слову/подстроке, результатом которого будет список всех реплик опубликованной части корпуса с вхождением заданного слова;
б) каждая реплика, полученная по запросу, будет сопровождаться социологической информацией о говорящем (пол, возраст, профессия/род занятий, социальный статус и др.) и о типе коммуникативной ситуации (бытовой разговор, профессиональный разговор, учебный разговор, общение по типу «клиент-сервис»);
в) результаты поиска смогут быть развернуты до нескольких реплик, расширяющих контекст (однако, как и в большинстве лингвистических корпусов — НКРЯ, BNC и др., — полные тексты по запросу приводиться не будут);
г) станет возможным создание поискового подкорпуса с заданными социальными и психологическими характеристиками говорящих, что обеспечит фильтрацию результатов поиска;
д) наконец, онлайн версия корпуса будет поддерживать поиск по типам и функциям прагматических маркеров — частотным единицам устного дискурса с ослабленным или полностью стертым семантическим компонентом, осуществляющих в речи множество прагматических задач (делимитативную, дейктическую, ксенопоказательную, метакоммуникативную, ритмообразующую, хезитативную и др.) [Богданова-Бегларян 2014].

Публикация текстовых материалов ОРД на веб-сайте сделает их доступными для всех заинтересованных исследователей русской устной речи и повседневной коммуникации.

Литература

1. *Богданова-Бегларян Н. В.* (2014), Прагматемы в устной повседневной речи: определение понятия и общая типология // Вестник Пермского университета. Российская и зарубежная филология. Вып. 3 (27). С. 7–20.
2. *Богданова-Бегларян Н. В., Кислощук А. И., Шерстинова Т. Ю.* (2013), О ритмообразующей функции дискурсивных единиц // Вестник Пермского университета. Российская и зарубежная филология. Вып. 2 (22). С. 7–17.
3. *Квеселевич Д. И.* (ред.) (2011), Самый полный словарь ненормативной лексики и фразеологических единиц: 20 000 слов. М.: Астрель: АСТ.
4. *Мокиенко В. М., Никитина Т. Г.* (ред.) (2004), Словарь русской брани (матизмы, обсценизмы, эвфемизмы). СПб.: Норинт.
5. *Потапова Р. К., Потапов В. В.* (2006), Язык, речь, личность. М.: Языки славянской культуры.
6. *Рекомендации* по применению Федерального закона от 05.04.2013 № 34-ФЗ «О внесении изменений в статью 4 Закона Российской Федерации «О средствах массовой информации»… // https://16.rkn.gov.ru/directions/mass-communications/control-nadzor-smi/spravka/p15638/ [Электронный ресурс] (дата обраще-ния — 05.05.2019)
7. *Химик В. В.* (2000), Поэтика низкого, или просторечие как культурный феномен. СПб.: Филологический ф-т СПбГУ.
8. *Шерстинова Т. Ю.* (2016), Наиболее употребительные слова повседневной русской речи (в гендерном аспекте и в зависимости от условий коммуникации) // Компьютерная лингвистика и интеллектуальные технологии, 15 (22), М.: РГГУ. С. 616–631.
9. *Bogdanova-Beglarian, N., Baeva, E., Blinova, O., Martynenko, G., Sherstinova, T.* (2018), Towards a Description of Pragmatic Markers in Russian Everyday Speech. In: SPECOM 2018, LNCS, vol. 11096. Springer, Cham. P. 42–48.

References

1. *Bogdanova-Beglarian, N.* (2014), Pragmatemy v ustnoj povsednevnoj rechi: opredelenie pon'atia i obshchaja tipologia [Pragmatems in Spoken Everyday Speech: Definition and General Typology]. In: Perm University Herald. Russian and Foreign Philology, 3 (27), 7–20.
2. *Bogdanova-Beglarian, N., Baeva, E., Blinova, O., Martynenko, G., Sherstinova, T.* (2018), Towards a Description of Pragmatic Markers in Russian Everyday Speech. In: SPECOM 2018, LNCS, vol. 11096. Springer, Cham, 42–48.
3. *Bogdanova-Beglarian, N. V., Kisloshchuk, A. I., Sherstinova, T. Iu.* (2013), O ritmoobrazujushchej funkcii diskursivnyx jedinic [On Rhythm-Forming Function of Discourse Markers]. In: Perm University Herald. Russian and Foreign Philology, 2 (22), 7–17.
4. *Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Baeva, E., Martynenko, G., Ryko, A.* (2016), Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. In: *SPECOM 2016.* LNAI. vol. 9811. Springer. Switzerland, 659–666.

5. *Khimik, V. V.* (2000), Poetika nizkogo, ili prostorechie kak kul'turnyj fenomen [Poetics of the Low, or Vernacular as a Cultural Phe-nomenon]. St. Petersburg: St. Petersburg State University.
6. *Kveselevich, D. I.* (ed.) (2011), Samyj polnyj slovar' nenormativnoj leksiki i frazeologicheskikh edinic. [The Most Comprehensive Vocabulary of Swear Words and Idiom]. Moscow: Astrel, AST.
7. *Mokienko, V. M., Nikitina, T. G.* (eds.) (2004), Slovar' russkoj brani (matizmy, obscenizmy, evfemizmy) [Dictionary of Russian Swear Words (Matism, Obscene, Euphemism)]. St. Petersburg: Norint.
8. *Potapova, R. K., Potapov, V. V.* (2006), Jazyk, rech', lichnost' [Language, Speech, Personality]. Moscow: Jazyki slavyanskoj kul'tury [Languages of Slavic Culture].
9. *Rekomendacii* po primeneniju Federal'nogo zakona ot 05.04. 2013 № 34-FZ «O vnesenii izmenenij v stat'ju 4 Zakona Rossijskoj Federacii «O sredstvakh massovoj informacii»… In: https://16.rkn.gov.ru/ directions/mass-communications/control-nadzor-smi/spravka/p15638/ [Electronic resource] (last accessed on 05.05.2019).
10. *Sherstinova, T.* (2016), The Most Frequent Words in Everyday Spoken Russian (in the Gender Dimension and Depending on Com-munication Settings). In: Komp'juternaja lingvistika i intellektual'nye tekhnologii, 15 (22), Moscow, 616–631.

Шерстинова Татьяна Юрьевна
Национальный исследовательский университет
«Высшая школа экономики» (Санкт-Петербург, Россия)
Санкт-Петербургский государственный университет (Россия)
Tatiana Sherstinova
National Research University Higher School of Economics
(St. Petersburg, Russia)
St. Petersburg State University (Russia)
E-mail: sherstinova@gmail.com

П. М. Эйсмонт
P. M. Eismont

МУЛЬТИМОДАЛЬНОСТЬ В КОРПУСЕ УСТНЫХ ДЕТСКИХ ТЕКСТОВ «КОНДУИТ»[1]

MULTIMODALITY IN THE CORPUS OF SPOKEN CHILD NARRATIVES "KONDUIT"

Аннотация. В статье описывается принцип организации мультимодального компонента Корпуса Неподготовленных Детских Устных Извлеченных Текстов «Кондуит». Корпус состоит из 213 устных текстов русскоязычных детей в возрасте 2;7 – 7;6 лет, представленных в виде аннотированных орфографических записей, а также аннотированных аудио и видеофайлов полученных рассказов. Корпус «Кондуит» на данный момент является единственным русскоязычным корпусом устных детских текстов, содержащим мультимодальные данные в открытом доступе.
Ключевые слова. Детская речь, корпус, Кондуит, мультимодальность, жесты, просодика.

Abstract. The paper describes the principle of organizing the multimodal component of the Corpus "Konduit" – the corpus of child oral unprepared elicited narratives. The corpus consists of 213 oral texts produced by Russian-speaking children aged 2, 7–7, 6 years old and presented in the form of annotated texts, as well as annotated and synchronized audio and video files. The "Konduit" Corpus is currently the only Russian-language corpus of oral children's texts with multimodal data in the public domain.
Keywords. Child language, corpus, Konduit, multimodality, gesture, prosody.

1. Постановка проблемы

Коммуникация мультимодальна. На рубеже 20 и 21 веков все большее число лингвистов пришло к мнению, что изучение языка, анализ живой речи невозможен без анализа всех прочих компонентов коммуникативной ситуации [Кибрик 2010]. Значительную часть информации человек передает при помощи невербальных средств, к которым в первую очередь относятся мимика, жестикуляция, позы, просодика и пр. Взаимодействие зрительного и слухового восприятия обеспечивает успешность коммуникации, а их расхождение или непонимание каких-либо невербальных средств может привести к коммуникативной неудаче даже при полном понимании вербального компонента (ср. например, незнание определенных жестов носителями разных языковых культур).

[1] Проект № 16-04-50114 «Усвоение семантико-синтаксической структуры русского глагола» выполнен при поддержке Российского фонда фундаментальных исследований.

За период речевого развития ребенку необходимо полностью овладеть набором невербальных средств коммуникации как с точки зрения их порождения, так и с точки зрения их восприятия. Проведенные М. Томаселло исследования показали, что именно невербальные средства коммуникации (жестикуляция и просодика) сыграли важную роль в эволюции человека и развитии языка, а также оказываются неотъемлемой частью процесса усвоения языка детьми [Томаселло 2011 (Tomasello 2008)]. Так, известно, что понимать и выражать различные эмоции при помощи мимических средств дети начинают уже в раннем младенческом возрасте, а вербализация таких ярких эмоций, как страх или радость появляется в возрасте 2 лет [Галкина 2016]. На ранних этапах развития речи жесты не просто дополняют вербальную информацию, а компенсируют ее несовершенство, принимая на себя в том числе и передачу грамматических характеристик высказывания. Так, например, по мнению некоторых авторов [Седов 2004, Сизова 2015], указующий жест ребенка, сопровождающий какую-либо вокализацию, отражает информационную структуру высказывания, составляя из вокализации и жеста пару «топик-коммент».

Осознание важности мультимодальной структуры коммуникации привело к появлению в последние годы разнообразных мультимодальных корпусов. Научно-исследовательская группа под руководством А. А. Кибрика работает над созданием корпусов устных рассказов — «Рассказы о сновидениях» [Кибрик, Подлесская 2009], «Веселые истории из жизни» (www.spokencorpora.ru), рассказы о грушах, в которых именно невербальному компоненту уделяется наибольшее внимание [Кибрик, Федорова 2018]. В Национальном корпусе русского языка развивается мультимодальный подкорпус «Мурко», содержащий фрагменты кинофильмов и театральных постановок, где также основное внимание уделяется именно жесту и его отражению в речи [Гришина 2017].

В то же время существующие корпусы детской речи практически не уделяют внимания проблеме мультимодальности. В самой известной базе детских данных CHILDES некоторые записи содержат аудио и видеофайлы, однако они никак не аннотированы и не проанализированы с точки зрения невербальных компонентов общения. Остальные базы русскоязычных детских данных находятся к закрытом доступе. Представляемый ниже корпус «Кондуит» является, таким образом, единственным русскоязычным корпусом устных детских текстов, содержащим мультимодальные данные в открытом доступе.

2. Мультимодальность в корпусе «Кондуит»

КОрпус Неподготовленных Детских Устных Извлеченных Текстов «Кондуит» (konduitcorpus.ru) был собран в 2014–2016 гг. при проведении серии экспериментов с 213 русскоязычными монолингвами в возрасте от 2;7 до 7;6 лет [Эйсмонт 2017]. Все дети, принявшие участие в экспериментах, посещали детские сады г. Санкт-Петербурга. При проведении экспериментов велась аудио и видеофиксация, и именно собранные аудио и видеоданные и составляют мультимодальный компонент данного корпуса. Все дети были разделены на 5 возрастных групп, для которых было разработано три экспериментальных дизайна. Дети самой младшей группы (в возрасте от 2;7 до 3;6 лет) участвовали в игре, где два помощника экспериментатора при помощи игрушек-бибабо совершали различные действия. Задача ребенка состояла в том, чтобы эти действия максимально подробно и точно описать. Дети второй возрастной группы (в возрасте от 3;7 до 4;6 лет) должны быть рассказать историю по книжке в картинках «Три котенка» (автор В. Сутеев), а дети трех старших возрастных групп (в возрасте от 4;7 до 5;6, от 5;7 до 6;6 и от 6;7 до 7;6) должны были рассказать историю по мультфильму, для чего был выбран четырехминутный фрагмент мультфильма «Как стать большим?» (Союзмультфильм, 1967). На примере последних и будет ниже представлен мультимодальный компонент корпуса.

К сожалению, в связи с требованиями законодательства РФ, а также в соответствии с этическими нормами проведения экспериментов с участием детей ни аудио, ни видеофайлы, содержащие запись голоса или представляющие лицо ребенка не могут быть представлены в сети Интернет в открытом доступе. В корпусе данные материалы представлены в следующем виде. При помощи программы субтитрирования Aegisub 3.2.2 был разработан шаблон для субтитрирования демонстрировавшегося детям фрагмента мультфильма, по которому все полученные тексты были покадрово соотнесены с описываемой ребенком в данный момент времени ситуацией (см. рис. 1). В свою очередь при помощи программы аудио и видеообработки ELAN производится аннотация речевого сигнала, его просодики и жестикуляции испытуемых. Полученные форматы (исходный мультфильм с наложенным в виде субтитров рассказом ребенка и сам рассказ в аннотированном виде) синхронизируются, что позволяет одновременно наблюдать как процесс восприятия (как быстро ребенок реагирует на различные дей-

ствия героев, какие из действий персонажей оказываются достаточно важными, чтобы найти своё отражение в устном рассказе и т. д.), так и процесс порождения (какие вербальные и невербальные средства использует для описания данного действия ребенок, имитирует ли он жесты или мимику персонажей и т. д.).

Рис. 1. **Кадр из мультфильма с наложенным субтитром (на данном кадре бабушка сокрушенно качает головой, однако ребенок не описывает само движение, а передает его содержание так, как он его воспринял)**

На протяжении всего мультфильма персонажи совершают различные жесты и демонстрируют чувства при помощи мимики: котенок опускает голову, когда бабушка его ругает; бабушка качает головой и разводит руками, когда видит беспорядок в комнате; котенок виляет хвостиком, когда видит бобров; бобры отрицательно качают головой, когда прогоняют котенка; котенок морщится, плачет, озадаченно округляет глазки и т. д. Все жесты и мимические движения представлены ярко и точно, режиссер мультфильма акцентирует на них внимание зрителя, однако далеко не все дети отражают этот невербальный компонент в своих рассказах.

Предлагаемый формат синхронизации мультфильма и реакции испытуемых позволяет также изучать процессы восприятия и наличие или отсутствие сопереживания героям у детей разного возраста. Так, например, мимика детей отражает мимику и переживания персонажа

даже в тех случаях, когда это никак не представлено в их рассказах. В то же время жесты воспроизводятся детьми в единичных случаях, и это больше свойственно детям младшего возраста. При этом воспроизводимый жест может сопровождаться яркой эмоциональной нагрузкой интонационной структуры высказывания, которую ребенок подбирает, основываясь исключительно на мимике и жестах персонажей, поскольку мультфильм демонстрировался испытуемым в беззвучном режиме.

Исследования процессов усвоения детьми невербальных компонентов общения и их сопоставление с ходом формирования нарративных навыков может стать важным шагом в изучении сложной мультимодальной структуры коммуникации.

Литература

1. *Галкина Е. В.* (2016) Вербализация эмоционального состояния страха в детском возрасте. Проблемы онтолингвистики — 2016. СПб., Листос. С. 838–864.
2. *Гришина Е. А.* (2017) Русская жестикуляция с лингвистической точки зрения. М., ЯСК.
3. *Кибрик А. А.* (2010) Мультимодальная лингвистика. Когнитивная лингвистика — IV. М., ИП РАН. С. 134–152.
4. *Кибрик А. А., Подлесская В. И.* (ред.) (2009) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М., ЯСК.
5. *Кибрик А. А., Федорова О. В.* Рассказы и разговоры о грушах: промежуточные итоги. VIII международная конференция по когнитивной науке. Светлогорск. С. 499–501.
6. *Седов К. Ф.* (2004) Дискурс и личность. М., Лабиринт.
7. *Сизова О. Б.* (2015) Жестовая и речевая коммуникация у детей я тяжелыми нарушениями речи. Психолингвистические аспекты изучения речевой деятельности. № 13. С. 154–167.
8. *Томаселло М.* (2011) Истоки человеческого общения. М., ЯСК.
9. *Эйсмонт П. М.* (2017) «Кондуит»: корпус устных детских текстов. Корпусная лингвистика — 2017: Труды международной конференции. С. 373–377.

References

1. *Galkina E. V.* (2016) Verbalizatiya emocionalnogo sostojanija straha v detskom vozraste [Verbal representation of the emotion FEAR in childhood]. Problemy ontolingvistiki — 2016 [Problems of ontholinguistics — 2016]. Saint-Petersburg, pp. 838–864.
2. *Grishina E. A.* (2017) Russkaya zhestikulyatsiya s lingvisticheskoy tochki zrenija [Russian gestures from the linguistic point of view]. Moscow.

3. *Kibrik A. A.* (*2010*) Multimodalnaya lingvistika [Multimodal linguistics]. Kognitivnaya lingvistika — IV [Cognitive Linguistics — IV]. Moscow, pp. 134–152.
4. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009) Rasskazy o snovideniyakh: Korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow.
5. *Kibrik A. A., Fedorova O. V.* Rasskazy i razgovory o grushah: promezhutochnye itogi [Stories and conversations about pears: preliminary results]. VIII mezhdunarodnaya konferentsiya po kognitivnoy nauke [VIII International Conference on Cognitive Science]. Svetlogorsk, pp. 499–501.
6. *Sedov K. F.* (2004) Diskurs i lichnost' [Discourse and personality]. Moscow.
7. *Sizova O. B.* (2015) Zhestovaya i rechevaya kommunikatsiya u detey s tyazhelymi narushenijami rechi [Gesture and verbal communication of children with severe speech disorders]. Psiholingvisticheskie aspekty izychenija recgevoy dejatelnosti [Psycholinguistic aspects of speech], 13, pp. 154–167.
8. *Tomasello M.* (2008) Origins of human communication. MIT press.
9. *Eismont P. M.* (*2017*) «Konduit»: korpus ustnyh detskih tekstov ["Konduit": corpus of child oral narratives]. Korpusnaya lingvistika — 2017: Trudy mezhdunarodnoj konferentsii [Corpus linguistics — 2017: Proceedings of international conference]. Saint-Petersburg, pp. 373–377.

Эйсмонт Полина Михайловна
Санкт-Петербургский государственный университет
аэрокосмического приборостроения (Россия)
Eismont Polina
Saint Petersburg State University of Aerospace Instrumentation (Russia)
E-mail: polina272@hotmail.com

КОРПУСЫ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

CORPORA OF LITERARY TEXTS

А. О. Гребенников, А. Н. Ассель

A. O. Grebennikov, A. N. Assel

БАЗА РУССКОГО РАССКАЗА XIX–XX ВЕКОВ. МОДЕЛИ АППРОКСИМАЦИИ[1]

XIX –XX CENTURIES' RUSSIAN SHORT STORIES CORPUS. APPROXIMATION MODELS

Аннотация. В продолжение предыдущих исследований сравниваются результаты использования функции Вейбулла и функции Хауштайна для аппроксимации зависимости объема словаря от объема выборки на материале из корпуса «База русского рассказа XIX –XX веков» (640 рассказов А. Т. Аверченко). Показано, что, в отличии от полученных ранее результатов, использование функции Хауштайна не всегда является предпочтительным для описания характера зависимости «текст – словарь». Выбор аппроксимирующей функции должен определяться характером данной зависимости.

Ключевые слова. Писательская лексикография, статистическое моделирование, стилометрия.

Abstract. Further the previous experiments the use of Weibull and Haustein functions for the approximation of the dependence between sample size and resulting vocabulary size is analyzed. The short stories by A. T. Averchenko was chosen as the material for the experiment (total volume is more than 500 000 tokens). Haustein function is not proved to be the preferable one for the approximation of the dependency that may result from the different character of vocabulary growth for the authors under investigation.

Keywords. Authors' lexicography, statistical modeling, stylometry.

[1] Работа выполнена при поддержке РФФИ, грант № 17-29-09173 «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

Выбор аппроксимирующей функции является одним из основных вопросов в стилеметрических исследованиях, в частности при моделировании зависимости объема словаря от объема выборки. Традиционно, прогностические результаты об объеме генеральной совокупности за пределами реального диапазона наблюдений представляются труднопроверяемыми из-за большого размера исходной генеральной совокупности. В продолжение предпринятого ранее исследования на материале частотного словаря рассказов А. П. Чехова [Гребенников 2017] были проанализированы 640 рассказов А. Т. Аверченко (общий объем выборки — свыше 500 000 словоупотреблений являющиеся частью проекта «База русского рассказа XIX–XX веков» [Мартыненко и др. 2018].

Для аппроксимации зависимости объема словаря от объема выборки были выбраны:

функция Вейбулла:

$$y = N_{max} - N_{max}e^{-cx^d}, \quad (1);$$

и функция Хауштайна:

$$y = \frac{N_{max}x^\gamma}{x^\gamma + q}, \quad (2)$$

где N — объем словаря, x — объем выборки, N_{max} — асимптотический объем словаря, c, d; γ, q — параметры распределения [Косарева и др. 2015; Haustein 1970].

Рассказы последовательно объединялись порциями по 10 в хронологическом порядке (предпочтенным по результатам предыдущих исследований [Гребенников 2017]) и для этих объединений составлялись ранговые частотные словари лексем. Полученные результаты нарастания объема словаря затем были аппроксимированы по авторизированной методике исследования, разработанной Г. Я. Мартыненко с использованием метода наименьших квадратов [Косарева и др. 2015; Гребенников1998]. Полученные результаты приводятся в табл. 1.

Также возможно рассмотреть прогностические величины, полученные при помощи исследуемых формул, за пределами реального диапазона наблюдений (табл. 2).

Таблица 1. **Результаты аппроксимации для кумулятивных совокупностей текстов**

Кол-во рассказов	Словоупотр.	Лексемы	Апп. по Вейбуллу (N_{max} = 176 347)	Апп. по Хауштайну (N_{max} = 284 566)
10	11 437	3 071	3 093	3 053
20	22 389	4 795	5 802	5 770
30	32 659	6 182	8 241	8 225
40	39 758	9 562	9 882	9 879
50	47 307	12 429	11 594	11 605
60	54 102	14 560	13 107	13 130
70	62 370	16 950	14 917	14 955
80	73 208	19 469	17 241	17 296
90	85 551	20 862	19 826	19 898
100	95 360	23 112	21 837	21 920
110	105 975	25 438	23 972	24 065
120	106 405	25 483	24 057	24 151
130	117 977	27 694	26 337	26 439
140	126 666	29 460	28 019	28 125
150	136 589	31 297	29 909	30 019
160	148 483	33 222	32 134	32 245
170	157 090	34 708	33 718	33 828
180	165 331	36 103	35 213	35 323
190	179 102	37 943	37 669	37 774
200	191 212	39 471	39 786	39 885
210	202 694	41 376	41 756	41 848
220	217 437	43 250	44 237	44 318
230	229 317	44 826	46 196	46 267
240	242 204	47 676	48 283	48 341
250	254 332	50 174	50 211	50 257
260	267 173	52 682	52 215	52 249

Продолжение табл. 1

Кол-во рассказов	Словоупотр.	Лексемы	Апп. по Вейбуллу (N_{max} = 176 347)	Апп. по Хаустайну (N_{max} = 284 566)
270	275 561	53 804	53 505	53 530
280	286 772	55 089	55 205	55 217
290	296 799	56 322	56 702	56 704
300	306 637	57 431	58 151	58 142
310	315 109	58 582	59 382	59 365
320	325 895	59 919	60 929	60 901
330	335 883	61 381	62 341	62 303
340	348 001	63 437	64 028	63 978
350	358 302	65 012	65 440	65 381
360	363 759	65 712	66 180	66 116
370	369 977	66 371	67 017	66 948
380	377 763	67 301	68 054	67 979
390	384 215	68 008	68 905	68 826
400	388 311	68 494	69 442	69 359
410	393 272	69 096	70 088	70 002
420	403 262	70 379	71 376	71 284
430	413 332	71 753	72 656	72 560
440	423 435	73 044	73 924	73 823
450	434 885	74 539	75 340	75 235
460	448 892	76 195	77 043	76 935
470	458 250	77 203	78 163	78 055
480	471 424	78 999	79 717	79 609
490	480 913	80 364	80 820	80 713
500	492 650	81 358	82 164	82 062
510	505 478	82 611	83 611	83 514
520	515 574	83 885	84 732	84 641
530	521 133	84 559	85 343	85 256

Окончание табл. 1

Кол-во рассказов	Словоупотр.	Лексемы	Апп. по Вейбуллу (N_{max} = 176 347)	Апп. по Хаустайну (N_{max} = 284 566)
540	525 814	85 132	85 854	85 770
550	532 029	86 128	86 528	86 449
560	539 631	87 098	87 345	87 273
570	546 062	88 024	88 030	87 964
580	552 065	88 895	88 664	88 605
590	556 887	89 487	89 170	89 117
600	563 019	90 290	89 809	89 763
610	570 317	91 332	90 562	90 526
620	577 435	91 635	91 290	91 264
630	586 523	92 573	92 210	92 198
640	590 387	92 974	92 598	92 593

Таблица 2. **Прогностические результаты об объеме словаря за пределами диапазона наблюдений**

Словоупотр.	Апп. по Вейбуллу	Апп. по Хаустайну
650 000	98 347	98 465
750 000	107 067	107 509
850 000	114 725	115 676
950 000	121 562	123 091
1 000 000	124 662	126 549

Из таблиц видно, что и на эмпирическом, и на теоретическом интервалах прогностические кривые весьма близки друг к другу, иногда с различием до всего лишь нескольких десятков единиц. При этом, в отличии от эксперимента со словарем Чехова, функция Вейбулла дает гораздо более реальные значения максимального объема словаря (176 347 с/у) и меньшие, хотя и не с такой значительной разницей, прогностические величины. Это, скорее всего, объясняется разным характером динамики нарастания объема словаря у Аверченко и Чехова (см. рис. 1 –2).

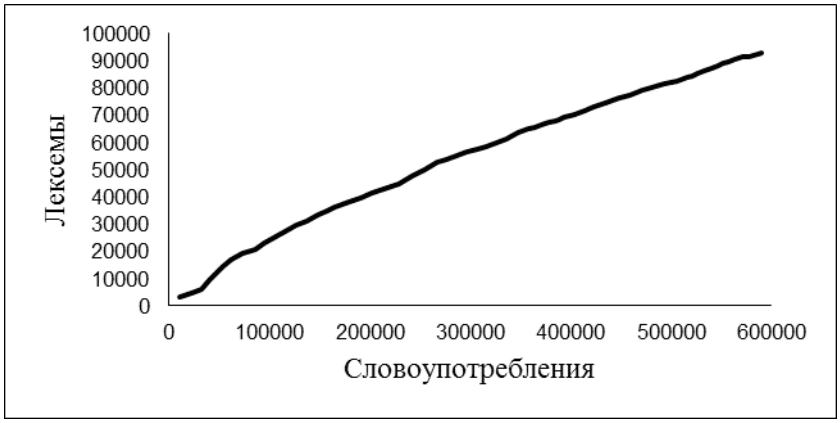


Рис. 1. **Динамика нарастания словаря А. Т. Аверченко**

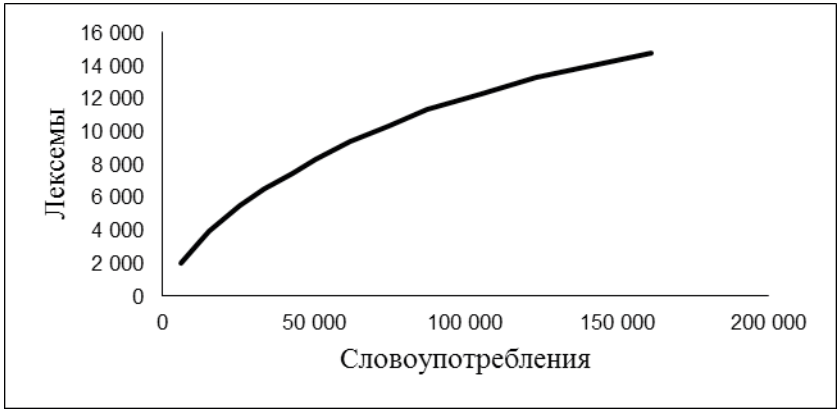


Рис. 2. **Динамика нарастания словаря А. П. Чехова**

Приведенные рисунки позволяют увидеть затухающую скорость нарастания объёма словаря со стремлением к некому верхнему пределу, начиная приблизительно со 150 000 с/у для словаря Чехова и неуклонное возрастание объема словаря с лишь слабым намеком на возможную стабилизацию словаря Аверченко. Причины этого лежат за пределами тематики данной статьи но, в самом общем приближении, вероятно вызваны господством рассказов крайне небольшого размера и чрезвычайно разнообразной тематики в творчестве Аверченко.

Таким образом, представляется возможным сделать вывод о том, что функция Хауштайна, показывающая гораздо более точные значения при прогнозировании словарей с явным и относительна быстрым стремлением к верхнему пределу, значительно уступает пока функции Вейбулла при аппроксимации словарей, демонстрирующих неуклонную тенденцию к нарастанию со слабой тенденцией к стабилизации.

Дальнейший содержательный и сопоставительный анализ с привлечением расширенного материала из «Базы русского рассказа XIX–XX веков» представляется перспективным.

Литература

1. *Гребенников А. О.* (1998), Исследование устойчивости лексико-статистических характеристик текста. Дис. … канд. филол. наук. СПб.
2. *Гребенников А. О.* (2017), К вопросу об аппроксимации зависимости объема словаря от объема выборки. Корпусная лингвистика–2017. Труды международной конференции. СПб, с. 151−156.
3. *Косарева Е. О., Мартыненко Г. Я.* (2015), Отношение текст — словарь в повседневной устной речи. Структурная и прикладная лингвистика: межвуз. сб., вып. 11. СПб, с. 220–228.
4. *Мартыненко Г. Я., Шерстинова Т. Ю., Попова Т. И., Мельник А. Г., Замирайлова Е. В.* (2018), О принципах создания корпуса русского рассказа первой трети XX века: Труды XV Международной конференции по компьютерной и когнитивной лингвистике «TEL 2018». Казань, с. 180–197.
5. *Haustein H.-D.* (1970), Prognoseferfahren in den sozialistischen Wiertschaft. Berlin.

References

1. *Grebennikov A. O.* (1998), Issledovanie ustoychivosti leksiko-statisticheskikh kharakteristik teksta [Validity of the Statistics for Fiction]. Dis. … kand. filol. nauk [PhD (Linguistcs) Thesis]. Saint Petersburg.
2. *Grebennikov A. O.* (2017), K voprosu ob approksimatsii zavisimosti ob"ema slovarya ot ob"ema vyborki [Approximation of the Sample Size–Vocabulary Size Dependence]. Korpusnaya lingvistika–2017. Trudy mezhdunarodnoj konferentsii [Corpus Linguititcs–2017. Proceedings of International Conference]. Saint Petersburg, pp. 151–156.
3. *Haustein H.-D.* (1970), Prognoseferfahren in den sozialistischen Wiertschaft. Berlin.
4. *Kosareva E. O., Martynenko G. Ya.* (2015), Otnoshenie tekst — slovar v povsednevnoy ustnoy rechi [The Type-Token Ratio in Everyday Spoken Russian]. Strukturnaya i prikladnaya lingvistika [Structural and Applied Linguistics], Vol. 11. Saint Petersburg, pp. 220–228.
5. *Martynenko* G. *Ya., Sherstinova T. Yu., Popova T. I., Melnik A. G., Zamirajlova E. V.* (2018), O printsipakh sozdaniya korpusa russkogo rasskaza pervoj treti XX veka [On

the Principles of Creation of the Russian Short Stories Corpus of the First Third of the XX Century]. Trudy XV Mezhdunarodnoj konferentsii po komp'yuternoj i kognitivnoj lingvistike «TEL 2018» [TEL 2018. Proceedings of International Conference on Computer and Cognitive Linguistics]. Kazan, pp. 180–197.

Гребенников Александр Олегович
Санкт-Петербургский государственный университет (Россия)
Alexander Grebennikov
Saint Petersburg State University (Russia)
E-mail: a.grebennikov@spbu.ru

Ассель Анастасия Никитична
Санкт-Петербургский государственный университет (Россия)
Anastasiya Assel
Saint Petersburg State University (Russia)
E-mail: dzigrobzoro69@gmail.com

О. А. Митрофанова
O. A. Mitrofanova

ИССЛЕДОВАНИЕ СТРУКТУРНОЙ ОРГАНИЗАЦИИ ХУДОЖЕСТВЕННОГО ПРОИЗВЕДЕНИЯ С ПОМОЩЬЮ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ: ОПЫТ РАБОТЫ С ТЕКСТОМ РОМАНА «МАСТЕР И МАРГАРИТА» М. А. БУЛГАКОВА

ANALYSIS OF FICTION TEXT STRUCTURE BY MEANS OF TOPIC MODELLING: CASE STUDY OF «MASTER AND MARGARITA» NOVEL BY M. A. BULGAKOV

Аннотация. В докладе обсуждаются особенности построения тематической модели текста романа «Мастер и Маргарита» М.А.Булгакова. Результаты, полученные при автоматической обработке текста рассматриваемого художественного произведения, согласуются с литературоведческими данными о сложной композиции романа.
Ключевые слова. тематическое моделирование, художественные тексты, русский язык.

Abstract. The talk discusses peculiarities of topic modeling for «Master and Margarita» novel by M.A.Bulgakov. Results achieved in course of text processing conform with literary data on compositional structure of the novel in question.
Keywords. Topic modeling, Fiction, Russian.

1. Постановка проблемы

Фундаментальные принципы исследования сюжета художественного произведения были предложены В. Я. Проппом [Пропп 2001]. Его представление о выводе сюжета волшебной сказки на основе комбинации функций персонажей легло в основу разнообразных логических моделей сюжета. Сюжетные тексты представляют последовательность действий через отношение импликации между категориями сюжета, определяемыми шаблонами поведения персонажей. Это отношение наблюдается в глубинной структуре текста и в ее трансформациях (ср. анализ жанра сказки, басни и т. д.) и может быть объяснено в терминах логического исчисления [Мартемьянов 2004]. С позиций генеративной лингвистики исследование композиции текста можно проводить с помощью сюжетных грамматик, реконструирующих процесс порождения сюжета текста на основе набора правил, описывающих допустимые цепочки микроситуаций [Олкер 1987].

Теоретические модели генерации сюжета получили практическое воплощение в прикладных программах автоматического анализа и ге-

нерации сказочных сюжетов [Гаазе-Раппапорт 1980; Рафаева 2014]. Современные исследования сюжета художественных произведений связаны с реконструкцией связей между персонажами с помощью подходов, основанных на правилах (ср. опыт применения правил в Томита-парсере для извлечения фактов об отношениях [Bocharov, Bodrova 2014]) и нейронных сетей [Iyyer et al. 2016].

Наряду с этим, есть основания считать, что автоматизированное описание сюжета художественного произведения и его отдельных компонентов возможно с привлечением вероятностного тематического моделирования, процедуры, которая, как правило, применяется в работе с новостными или научными текстами [Blei et al. 2003]. Тем не менее, тематическое моделирование было использовано в ходе анализа англоязычных поэтических текстов [Rhody 2012]; существует опыт построения тематических моделей русскоязычных художественных текстов [Митрофанова 2015; Mitrofanova, Sedova 2017].

В данной работе мы обсуждаем результаты экспериментов по тематическому моделированию текста романа «Мастер и Маргарита» М. А. Булгакова, одного из самых загадочных и противоречивых произведений художественной прозы 20 века. Как известно, роман создавался длительное время, предположительно в период между 1928 и 1940 годами, и за это время сюжет претерпел существенные изменения. Рукопись романа дорабатывалась уже после кончины М. А. Булгакова его вдовой Е. С. Булгаковой и была опубликована лишь в 1966–1967 годах. Особенность данного произведения заключается в том, что это «роман в романе», где противопоставляются два сюжетных плана: с одной стороны, историческое повествование об Иешуа и Понтии Пилате в Ершалаиме (далее роман-Е), с другой — описание разнообразных событий, происходящих с Мастером и Маргаритой, с Воландом и его свитой в современной автору Москве (далее роман-М). В литературной критике признано несколько объяснений композиционной структуры текста (ср. Вулис 1991; Соколов 2006]). Целью нашего исследования является определение соотношения двух сюжетных планов в тексте романа с помощью эмпирического метода, а именно, тематического моделирования.

2. Исследовательская методика

Тематическое моделирование является одной из стандартных процедур статистического исследования содержательной структуры кор-

пусов текстов [Blei et al. 2003]. Вероятностная тематическая модель соотносит документы корпуса с распределением на множестве тем, а темы — с распределением на множестве слов. Тематическое моделирование представляет собой разновидность нечеткой кластеризации: отдельное слово может быть с разными вероятностями отнесено к нескольким темам, и, аналогичным образом, отдельному тексту может быть поставлено в соответствие несколько тем. В исследовательской практике наиболее распространенена модель LDA (Latent Dirichlet Allocation (латентное размещение Дирихле) [Blei et al. 2003]. Вероятностная модель порождения данных в LDA описывается следующим образом: $p(w|d) = \Sigma p(t|d)\, p(w|t)$, где $p(w|d)$ — известная частота появления термина w в документе d, $p(w|t)$ — неизвестная вероятность появления термина w в теме t, $p(t|d)$ — неизвестная вероятность появления темы t в документе d. Построить тематическую модель текстовой коллекции D — значит найти множество тем T, распределения $p(w|t)$ для всех тем и $p(t|d)$ для всех документов. Процесс является итеративным и прекращается при достижении сходимости алгоритма.

Для извлечения тем использовалась реализация модели LDA в библиотеке scikit-learn [scikit-learn.org]. Для визуализации результатов тематического моделирования была применена библиотека pyLDAvis [https://github.com/bmabey/pyLDAvis], которая позволяет оценить полученные темы в интерактивном режиме. Текст романа извлечен из Библиотеки М. Мошкова, объем текста составляет ~ 120 тыс. с/у. Непосредственно перед проведением тематического моделирования была проведена предобработка текста, включающая лемматизацию и удаление стоп-слов. Тематическая модель строилась со следующими параметрами: 30 тем, по 30 первых слов в выдаче.

3. Результаты

Эксперименты позволили выявить следующие особенности текста романа «Мастер и Маргарита», отраженные в тематической модели.

1. В модели четко противопоставлены темы, связанные с сюжетами романа-Е и романа-М. Единственная тема, содержащая лексику из глав об Иешуа и Понтии Пилате из романа-Е — это тема 1: *прокуратор, Пилат, гость, Иешуа, Ершалаим*…(рис. 1). Остальные темы 2…30 связаны с главами, где описываются различные эпизоды, связанные с персонажами из романа-М: ср. тема 4: *Варенуха, Римский, финдиректор, кабинет, администратор, Лиходеев*…. (рис. 2); тема

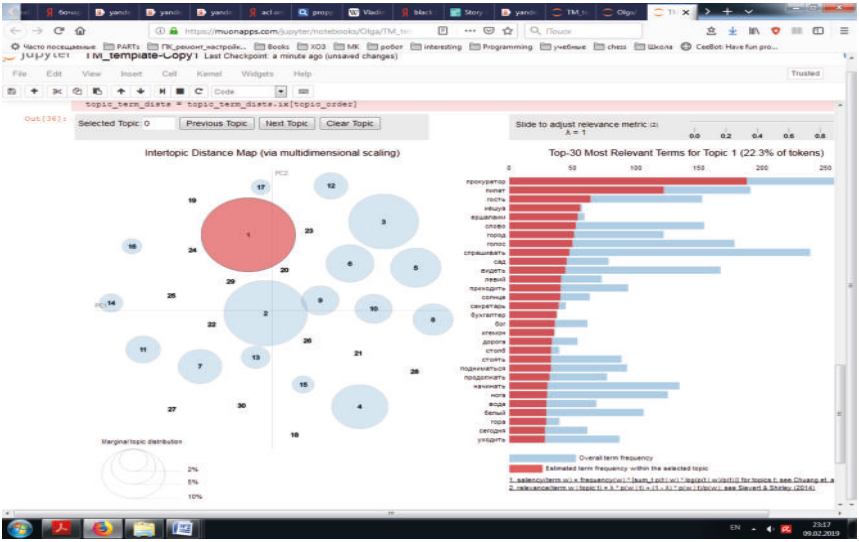


Рис. 1. **Тема 1: *Прокуратор, Пилат, гость…***

Рис. 2. **Тема 4: *Варенуха, Римский, финдиректор…***

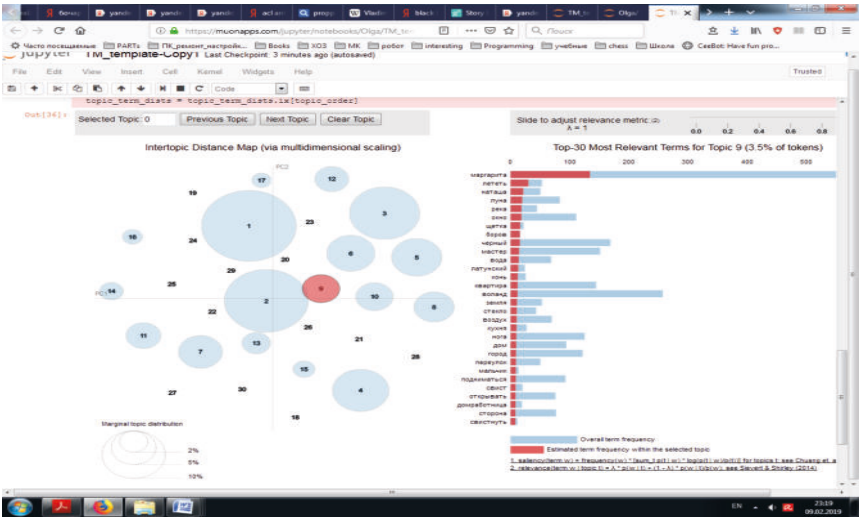


Рис. 3. **Тема 9:** ***Маргарита, лететь, Наташа…***

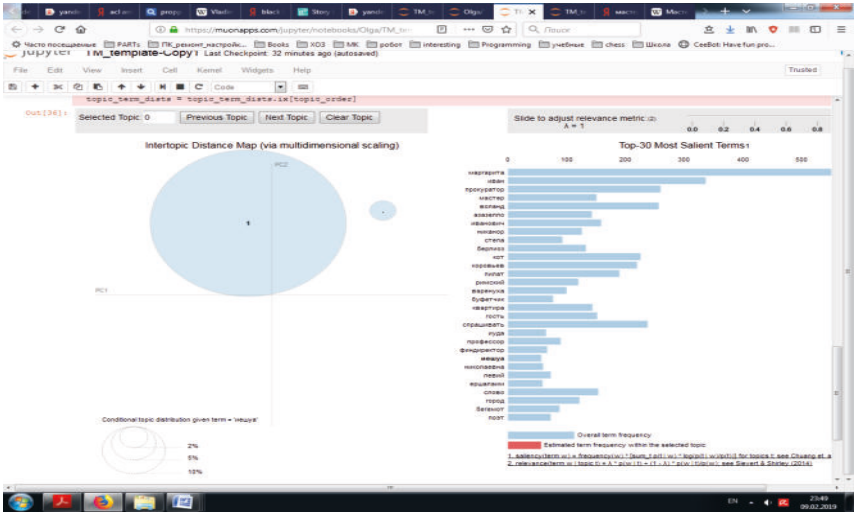


Рис. 4. **Иешуа:** ***распределение по темам***

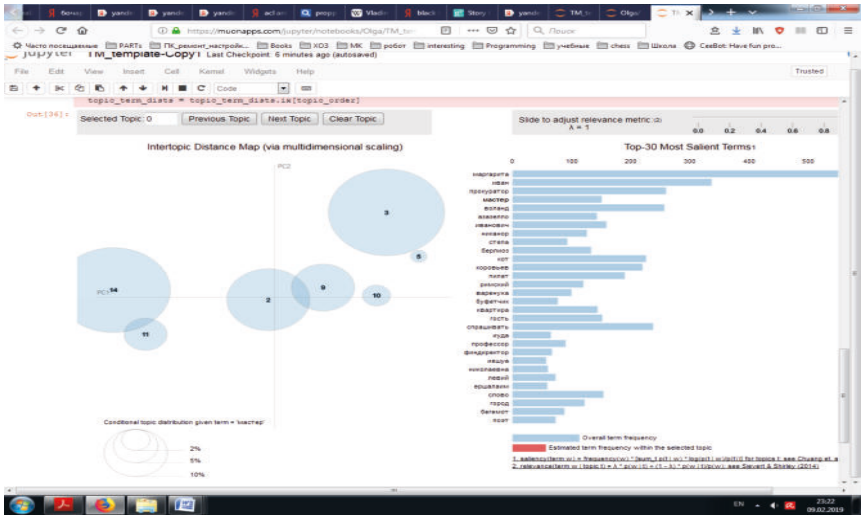


Рис. 5. **Мастер:** ***распределение по темам***

9: *Маргарита, лететь, Наташа, Луна, река, окно, щетка, боров* … (рис. 3) и т. д. В темах есть отдельные «шумы», но их доля не превышает 1 % в основной части выдачи (первые 30 слов в теме), ср. тема 1 (*бухгалтер* и *секретарь*). Пограничной можно считать тему 3, связанную с фрагментами из последних глав романа.

2. Имена персонажей распределены в темах не случайным образом, а в соответствии с тем, в каких частях романа и в каких внутренних эпизодах они упоминаются. Имя *Иешуа* присутствует как ядерный элемент в теме 1 для романа-Е (рис. 4), тогда как имя *Мастер* отражено в темах 2, 3, 5, 9, 10, 11, 14, связанных с главами романа-М (рис. 5).

4. Выводы

Полученные результаты содержат эмпирические данные о тексте романа «Мастер и Маргарита», подтверждающие литературоведческие наблюдения о композиционных особенностях романа. Это свидетельствуют о применимости тематического моделирования в задачах исследования сюжета.

С одной стороны, четкое содержательное противопоставление тем романа-Е и романа-М отражает состоятельность построенной нами тематической модели. С другой стороны, поскольку тематическая

модель основана на лингвостатистических параметрах текста, есть основания связывать различия между фрагментами текста с тем, что а) роман создавался на протяжении длительного времени, это могло оказать влияние на идиостиль автора, б) противопоставляемые части романа стилистически неоднородны — роман-Е близок и историческому повествованию, в то время как в романе-М используются художественные приемы, характерные для сатирических произведений.

Автор выражает благодарность литературоведу М. Н. Никольской и студентам кафедры математической лингвистики СПбГУ А. В. Крюковой и Е. В. Третьяк за обсуждение экспериментов.

Литература

1. *Вулис А. З.* (1991) Роман М. А. Булгакова «Мастер и Маргарита». М.
2. *Гаазе-Рапопорт М. Г., Поспелов Д. А., Семенова Е. Т.* (1980) Порождение структур волшебных сказок. М.
3. *Мартемьянов Ю. С.* (2004) Логика ситуаций. Строение текста. Терминологичность слов. М.
4. *Митрофанова О. А.* (2015). Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика–2015». СПб.
5. *Пропп В. Я.* (2001) Морфология волшебной сказки. М.
6. *Рафаева А. В.* (2014) Компьютер — слово — фольклор. М.
7. *Соколов Б. В.* (2006) Расшифрованный Булгаков. Тайны «Мастера и Маргариты» М.
8. *Blei D. M., Ng A. Y., Jordan M. I.* (2003) Latent Dirichlet Allocation // Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. URL: http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf
9. *Bodrova A. A., Bocharov V. V.* (2014) Relationship Extraction from Literary Fiction // Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2014». Moscow. URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/BodrovaAABocharovVV.pdf
10. *Iyyer M., Guha A., Chaturvedi S., Boyd-Graber J., Daumé III H.* (2016) Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships // NAACL 2016. URL: https://www.aclweb.org/anthology/N16-1180
11. *Mitrofanova O. A., Sedova A. G.* (2017). Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose) // Information Technology and Computational Linguistics (ITCL-2017). Association for Computing Machinery.
12. *Rhody L. M.* (2012) Topic Modeling and Figurative Language // Journal of Digital Humanities. Vol. 2(1). Winter 2012. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

References

1. *Vulis A. Z.* (1991) Roman M. A. Bukgakova «Master i Margarita». [The Novel by M. A. Bulgakov «Master i Margarita»] M.
2. *Gaaze-Rapoport M. G., Pospelov D. A., Semenova E. T.* (1980) Porozhdenije Struktur Volshebnych Skazok [Generation of Structures for Fairy Tales]. M.
3. *Martemyanov Ju. S.* (2004) Logika Situacij. Strojenije Teksta. Terminologichnost' Slov. [The Logic of Situations. Text Structure. Terminology of Words]. M.
4. *Mitrofanova O. A.* (2015) Verojatnostnoje Modelirovanije Tematiki Russkojazychnyh Korpusov Tekstov s Ispol'zovanijem Kompjuternogo Instrumenta GenSim. [Probabilistic Topic Modelling of the Russian Text Corpora by Means of GenSim Toolkit] // Trudy Mezhdunarodnoj Konferencii «Korpusnaja Lingvistika — 2015». [Proceedings of the International Conference «Corpus Linguistics — 2015»]. St.-Petersburg.
5. *Propp V. Ja.* (2001) Morphologija Volshebnoj Skazki. [Morphology of Fairy Tales]. M.
6. *Rafajeva A. V.* (2014) Kompjuter — Slovo — Folklor. [Computer — Word — Folklore]. M.
7. *Sokolov B. V.* (2006) Raschyfrovannyj Bulgakov. Tajny «Mastera i Margarity» M.
8. *Blei D. M., Ng A. Y., Jordan M. I.* (2003) Latent Dirichlet Allocation // Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. URL: http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf
9. *Bodrova A. A., Bocharov V. V.* (2014) Relationship Extraction from Literary Fiction. Computational Linguistics and Intellectual Technologies // Proceedings of International Conference «Dialog–2014». Moscow. URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/BodrovaAABocharovVV.pdf
10. *Iyyer M., Guha A., Chaturvedi S., Boyd-Graber J., Daumé III H.* (2016) Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships // NAACL 2016. URL: https://www.aclweb.org/anthology/N16-1180
11. *Mitrofanova O. A., Sedova A. G.* (2017). Topic Modelling in Parallel and Comparable Fiction Texts (the Case Study of English and Russian Prose) // Information Technology and Computational Linguistics (ITCL-2017). Association for Computing Machinery.
12. *Rhody L. M.* (2012) Topic Modeling and Figurative Language // Journal of Digital Humanities. Vol. 2(1). Winter 2012. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Митрофанова Ольга Александровна
Санкт-Петербургский государственный университет (Россия)
Mitrofanova Olga
Saint Petersburg State University (Russia)
E-mail: oa-mitrofanova@yandex.ru

Г. Я. Мартыненко
G. Ya. Martynenko

СТИЛИЗОВАННЫЕ СИНТАКСИЧЕСКИЕ ТРИАДЫ В РУССКОМ РАССКАЗЕ ПЕРВОЙ ТРЕТИ XX ВЕКА[1]

STYLIZED SYNTACTIC TRIADS IN THE RUSSIAN SHORT STORY OF THE FIRST THIRD OF THE 20TH CENTURY

Аннотация. В статье на материале рассказа Артема Веселого «Реки огненные» рассматриваются структурные особенности стилизованного на основе сказа и орнаментализма индивидуального нарратива. В центре внимания находятся минимальные единицы такого повествования, называемые в статье триадами, а также более крупные единицы – текстовые фрагменты, построенные с использованием триад. Описываются особенности таких фрагментов, корреспондирующие с «огненной рекой» революционных событий, освещаемых в рассказе. Попутно рассматриваются стилистические средства, усиливающие эффективность орнаментальных структур, например «лестничные» конструкции. Предлагаемая методология позволяет объединить традиционный литературоведческий анализ с методами корпусной лингвистики.
Ключевые слова. Русская литература, русский рассказ, XX век, нарратив, сказ, орнаментальность, стилизация, синтаксические триады, «лесенки», квантитативный анализ.

Abstract. The article proposes methodology of structural features analysis of the individual narrative style. The methodology is shown on "Rivers of Fire" by Artem Vesely, which is a fine sample of Russian post-revolutionary ornamental narration. The paper focuses on the so-called "triads", which are the minimal ornamental units of narration, as well as on the larger units – text fragments formed by these triads. The features of such fragments corresponding with the "rivers of fire" of the revolutionary events are described both qualitatively and quantitatively. At the same time, stylistic tools that enhance the effectiveness of ornamental structures, such as the "stairsteps"-designs, are considered. The proposed methodology allows to combine traditional literary analysis with the methods of corpus linguistics.
Keywords. Russian literature, Russian short story, the 20th century, narrative, narration, ornamentality, stylization, syntactic triads, "stairsteps", quantitative analysis.

1. Введение

В современной русской словесности серьезное внимание отводится изучению жанра короткой строки (Смирнов, 1883; Нильссон, 1993; Анализ художественного текста, 2018; Мартыненко и др., 1918). Для

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 17-29-09173 офи_м «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

русской художественной прозы первой трети XX века характерна откровенная фигуративность и сказовость (Русская литература, 2005; Новиков, 1990). Первая черта коренится в символизме Андрея Белого. Писатель сокрушал стилистические каноны, втягивая в орбиту своего влияния огромное число последователей. Еще одна черта литературы этого периода — ее все возрастающая демократичность, обусловленная вовлечением в художественное творчество огромных человеческих масс. Это предопределяет ее сказовость. В рассказах и повестях зазвучала разношерстная народная речь, пусть олитературенная, пусть стилизованная, но очень непривычная для читателя, воспитанного на классической литературе.

Объектом наших наблюдений является повесть (или «большой рассказ») выдающегося писателя 20-х гг. Артема Веселого «Реки огненные» — концентрат литературных нововведений 20-х гг.

В рассказе речь идет о приключениях двух дружков-матросов, подхваченных огненной рекой русской революции и брошенных с морских просторов в пекло Гражданской войны. Свою миссию дружки осознали весьма однонаправленно. Они поняли, что в это смутное время можно «хватать все, что плохо лежит». Вернувшись на корабль после окончания Гражданской войны, они по возможности продолжали свои черные делишки. Это были грубые беспринципные существа, проникшие в гущу идейного комсомольского сообщества. Тем самым автор стремился показать «изнанку» революции, а не только ее парадную, пафосную часть. Дружки-герои имели крестьянское происхождение, и это наложило отпечаток на их поведение и особенности речи. В гуще революционных событий они нахватались обрывков псевдореволюционной, а окунувшись в криминальную атмосферу, — блатной фразеологии. Такая смесь породила в творческой лаборатории Артема Веселого весьма специфический сказ, образующий во взаимодействии с авторским орнаментализмом особый стиль. Некоторые черты этого стиля мы рассмотрим ниже.

Наше внимание сосредоточим на синтаксических триадах, а уже потом рассмотрим некоторые сходные сопутствующие явления.

2. Что такое синтаксическая триада?

Начнем с конкретной триады и условий, ее породивших. Попытаемся дать первичное объяснение причины ее появления в определенном фрагменте текста.

Открывает список триад колоритное предложение-абзац: «Чокнулись, уркнули, крякнули». Что характерно для этой триады? Каждый из ее структурообразующих элементов — глагол в одной и той же грамматической форме, последовательность которых образуют минимальный перечислительный ряд. Все глаголы принадлежат к сфере, соотносящейся с выпивкой. Глагол «уркнули», по-видимому, имеет значение звука, издаваемого в предвкушении употребления напитка. Из-за того, что все глаголы употреблены в одной и той же форме, они созвучны.

Далее через определенный интервал следует серия аналогичных триад. От обычных предложений-абзацев их отличает то, что все они предельно коротки, включают три единицы и построены с использованием ограниченного набора моделей.

Триаде «Чокнулись, уркнули, крякнули» предшествует текст, в котором описана встреча Ваньки-граммофона и Мишки-крокодила на корабле с боцманом Федотычем после долгой разлуки, включающей и ритуал распива коньяка, припрятанного добряком-боцманом. Триада здесь подводит итог этой совокупности действий, т. е. она имеет некоторое финальное, результативное значение.

Следующая триада похожа на первую, вселяет семантическую неопределенность: «Охнули, ххакнули, задермушились»… Что означает глагол «охнули» — понятно, а вот какой смысл несут нам второй и третий глаголы, не разобраться без консилиума продвинутых лексикографов и пьянчужек. Перед нами какие-то облачка смысла, очень милые, с весьма расплывчатым смыслом. Этого для конкретности мало. А может, эта конкретность и не нужна. Можно удовлетвориться общим впечатлением.

По этой модели строится большое число триад. С тем отличием, что они содержат в своем составе зависимые от глагола члены. Приведем несколько примеров без комментариев:

«Подмокли, рассолодели в ругатне, полоскались яро»,
«Утакали, удакали, съэтажили яро»,
«Дружков шатало, мотало, подмывало»,
«В уголь ужглись, укачало, утрепало».

Особое подмножество представляют триады с одним и тем же глаголом:

«Море качелилось,
песня качелилась,

качелились блесткие крики чаек».
«Ноги пляшут, теплушки пляшут, степя пляшут».

В остальных триадах в качестве стержневых слов выступают не глаголы, а другие части речи:

«В позевотину, в одеяло, в храп»,
«Тоска, смертный час, тошнехонько»,
«Паровоз в храпе, паровоз в мыле, пыль пылом».

То, что триады строятся и таким способом — не случайность. Их источником является пристрастие Артема Веселого к «элементарным», сверхкратким фразам типа: «Братки в рев», «Бабы в крик», «Ванька поперек», «Братки в скуку», «Пятки градом» и др. Такие фразы как правило двухчленны. Это оказывает влияние и на размер предложения в цельном тексте. В нем широко представлены предложения-абзацы, которые распределены в тексте следующим образом: см. табл. 1. Доля таких структур в общем числе абзацев составляет около трети от их общей массы. Соответствующий график показан на рис. 1. Он откровенно бимодален, указывая на неоднородность распределения. Левая часть представляет преимущественно стилизованные конструкции, правая — стилистически нейтральные.

Таблица 1. Распределение **в рассказе предложений-абзацев**

Число слов в предложении-абзаце	**Частота**	**Число слов в предложении-абзаце**	**Частота**
1	3	10	9
2	39	11	6
3	39	12	3
4	15	13	0
5	15	14	2
6	19	15	1
7	13	17	1
8	14	18	5
9	11	19 и более	5
Сумма			200

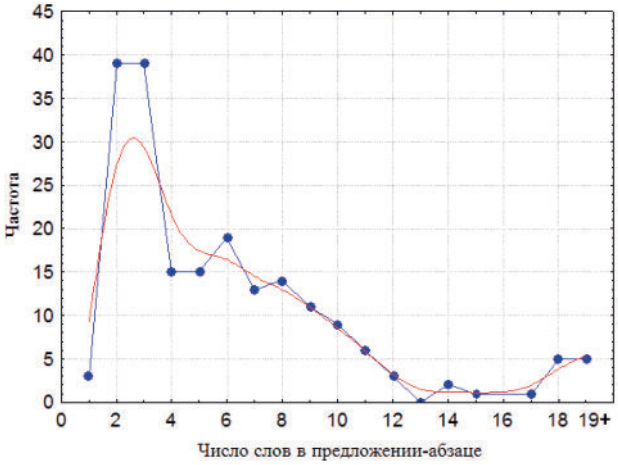


Рис. 1. **Распределение абзацев-предложений по размеру**

Доля сверхкоротких абзацев-предложений (размером не более трех слов) в массе таких структур весьма значительна — 40,5 %. Такие фразы, как правило, играют важную прагматическую роль в композиции текста. Чаще всего они фигурируют в эмоционально напряженных фрагментах.

Предельными вариантами таких бинарных структур являются фразы, подчеркивающие неразрывное эмоциональное единство двух главных героев рассказа. Например: «*Мишка засмеялся, Ванька засмеялся*». «*Мишка в обиде, Ванька в обиде*», «*Ванька зарадовался, Мишка зарадовался*». Полная духовная спаянность, неразрывность героев осуществляется путем введения еще одного объединительного члена:

«В Ваньке сердце стукнуло.
В Мишке сердце стукнуло.
Враз стукнули сердца».

Вот такая триада. Вы когда-нибудь встречались с чем-то подобным?

И совсем всем на удивление через пару страниц вводится еще одна лексически и синтаксически аналогичная триада:

«В Мишке сердце стукнуло,
В Ваньке сердце стукнуло,
Враз стукнули мерзлые, отощалые сердца».

В последней структуре триадность поддерживается трехступенчатой лесенкой. Такие орнаментальные структуры — излюбленный стилистический прием Артема Веселого:

«По палубе хлынул бег,
в парусиновую подвесную койку
укладывался корабль спать».

3. Расширение синтаксических триад и диад

Синтаксические триады и диады могут рассматриваться как перечислительные ряды минимальной длины. Протяженность ряда может быть увеличена на произвольное число членов. Это приводит к повествованию, основанному на однообразном нагнетании коротких глагольных предложений или однородных глаголов-сказуемых в рамках одного предложения.

Эти ряды, однако, не создают впечатление монотонности, не выбиваются из общей канвы нарратива. Они, быть может, более ритмичны, чем остальной текст.

Попытаемся убедиться в этом.

Для этого обратимся к весьма показательному фрагменту. Вот он:

«Густо плескались, пылали тяжелые ветра… Пылали, плескались зноем травы… Поезда бежали, зарывались в горы, с разбегу пробивали туннели. Табунами бродили пожары… Бежали сизые полынные степя… Дороги шумели половодьем.

Вытаптывая города и села, бежали красные, белые, серые и че-о-орная банда. Кованые горы бежали, дыбились, клешились. Бежали, как звери, густошерстые тучи, хвостами мутили игравшие реки.

Партизаны бежали, падали, бежали, плевались трескими, громами, бухами, хохом, ругом… Залпами расстреливали, бросками бросали наливные зерна разбойных дней».

Данный отрывок откровенно стилизован. Особенно режет слух настойчивое нагнетание глаголов в прошедшем времени, преимущественно в начальной позиции коротких предложений.

Для оценки стилистической оригинальности отрывка используем фоновый (обычный, не маркированный) отрывок, а также коллекцию триад. Сначала сравним эти три «текста» с точки зрения распределения частей речи и размера словоупотребления.

Фоновый текст приводится ниже.

«На дружках от военноморской робы одни клеши остались, обхлестанные клеши, шириною в поповские рукава. Да это и не беда! Ваньку с Мишкой хоть в рясы одень, а по размашистым ухваткам да увесистой сочной ругани сразу флотских признаешь. Отличительные ребятки; нахристые, сноровистые, до всякого дела цепкие да дружные. Насчет эксов, шамовки али какой ни на есть спекуляции Мишка с Ванькой первые хваты, с руками оторвут, а свое выдерут. Накатит веселая минутка — чужое для смеха прихватят. Черт с ними не связывайся — распотрошат и шкуру на базар. Даешь-берешь, денежки в клеш и каргала!»

Сравним распределения по двум переменным в трех фрагментах. Но сначала надо сказать несколько слов о распределении частей речи. Как и следовало ожидать, везде доминируют существительное и глагол, образующие основу любого синтаксиса. Однако соотношения этих частей речи в трех фрагментах разные: в стилизованном тексте и триадах доли существительного и глагола значительно выше (см. табл. 2), а доли большинства остальных классов заметно выше в фоновом тексте. Стилизованные фрагменты более однообразны, образуя цепочки, состоящие из комбинации глаголов и существительных. Такое нагнетание создает впечатление лавины, все сметающей на своем пути, подчеркивая и даже имитируя революционный процесс.

Определенный вклад в создание такого настроения производит и размер слова, который примерно одинаков в стилизованных частях, при этом существенно превышая этот показатель в фоновой части (см. табл. 3). Обратное соотношение имеем для коэффициента вариации, который в полтора-два раза меньше, чем в фоновом фрагменте. Это вызвано преимущественно тем, что в фоновом фрагменте доля служебных слов достаточно велика, а в стилизованных ничтожно мала (см. табл. 2).

В целом такие структуры мобилизуются для отражения динамизма эпохи, ее «сверхвысоких энергий». Но энергия эта имеет разную природу. Это в первую очередь «огненная река» революции, преобразующая старый затхлый мир в мир революционного порыва и дерзания. Это и трудовой энтузиазм напряженной и сложной матросской жизни в рамках жесточайшей дисциплины. Но наряду с такой позитивной энергией расцветает разрушительная энергия наживы, стяжания, криминала, разбоя главных героев — неразрывного тандема друзей Ваньки-граммофона и Мишки-крокодила: «Дело идет, контора пи-

Таблица 2. **Распределение частей речи**

Части речи	Фоновый текст	Стилизованный текст	Коллекция триад
Существительное	0,312	0,465	0,432
Глагол	0,229	0,315	0,365
Прилагательное	0,151	0,110	0,014
Наречие	0,011	0,014	0,081
Предлог	0,172	0,027	0,108
Сочинит. союз	0,086	0,027	0,000
Прочие	0,039	0,042	0,000

Таблица 3. Длина **слова в различных фрагментах текста**

Длина словоупотребл.	Фоновый текст	Стилизованный текст	Коллекция триад
1	0,140	0,057	0,108
2	0,129	0,000	0,000
3	0,043	0,014	0,054
4	0,129	0,086	0,108
5	0,161	0,186	0.095
6	0,097	0,243	0,176
7	0,161	0,129	0,176
8	0,022	0,129	0,100
9	0,054	0,057	0,041
10	0,011	0,086	0,068
11	0,054	0,000	0,068
12	0,011	0,014	0,028
13	0,011	0,000	0,000
Среднее	5,05	6,386	6,649
Коэфф. вариации	0,612	0,313	0,268

шет. Ванька-Мишка денежки гребут». У Артема Веселого нет в слоге никакого революционного пафоса. Рафинированная объективность и трезвый взгляд на действительность.

4. Лестничные структуры

Лестничные конструкции приводят не только к образованию триад. Во многих случаях такие структуры распространяются и на более протяженные образования, включающие более трех ступеней. В рассматриваемом рассказе использованы три лесенки, включающие 5, 7 и 9 ступеней. Каждая из этих лесенок строится по трем моделям: глагольно-субстантивной («облака топтали»), предложно-субстантивной («на Оренбург бурей») и атрибутивно- именной («Пляско вино»). В лесенках ступени-предложения коротки. Средний размер такого предложения равен 2,83, а коэффициент вариации — 0,353, что находится в соответствии с другими стилизованными фрагментами.

5. Заключение

Рассмотрены синтаксические триады типа «Чокнулись, уркнули, крякнули» и более крупные образования, имеющие облик связного текста, включающего подобные модели. Установлена их особая роль в повествовании в сравнении с обычным, фоновым текстом. Такие фрагменты текста в выпуклой форме передают накал революционных страстей, их необузданную стихию и мощь народного порыва, сметающего все на своем пути. Но не только. Это и энергия стяжательства и наживы, внедренная в революционные массы для удовлетворения корыстных интересов криминально-буржуазных элементов.

Литература

1. *Анализ художественного текста. Русская литература XX века: 20-е годы.* (2018) Отв. ред. Рогова К. А. СПб: Изд-во СПбГУ.
2. *Мартыненко Г. Я., Шерстинова Т. Ю., Мельник А. Г., Попова Т. И.* Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) / Компьютерная лингвистика и вычислительные онтологии. Вып. 2 (Труды XXI Межд. объед. конф. «Интернет и современное общество, IMS-2018, Санкт-Петербург, 30 мая — 2 июня 2018 г.»). — СПб: Университет ИТМО, 2018. С. 99–104.
3. *Новиков Л. А.* (1990) Стилистика орнаментальной прозы Андрея Белого. М.: Наука.
4. *Нильссон Н. О.* (1993) Русский импрессионизм: стиль «короткой строки». Русская новелла. Проблемы теории и истории / СПб: Изд-во СПбГУ. С. 236–249.
5. Русский рассказ XX века: Антология (2005) / Сост. и автор рис. на обложке Вл. Сорокин. М: Захаров.

6. *Смирнов И. П.* (1993) О смысле краткости / Русская новелла. Проблемы теории и истории. СПб: Изд-во СПбГУ, с. 1–5.

References

1. Analiz khudozhestvennogo teksta. Russkaya literatura XX veka: 20-e gody [Analysis of fiction. Russian literature of the 20th century: the 20th years] (2018) Rogova K. A. (ed.) SPb: SPbGU.
2. *Martynenko G. Y., Melnik A. G., Popova T. I., Sherstinova T. Y.* (2018), Methodological problems of creating the Computer Anthology of Russian short stories as a language resource designed to study language and style of Russian prose in the era of revolutionary changes (in the first third of the 20th century). In: Komp. lingvistika i vychislitel'nye ontologii. Vyp. 2, IMS-2018, SPb: ITMO, pp. 99–104.
3. *Novikov L. A.* (1990) Stilistika ornamental'noj prozy Andreya Belogo [Stylistics of the ornamental prose by Andrei Bely]. M.
4. *Nilsson N. O.* (1993) Russkij impressionizm: stil' «korotkoj stroki». Russkaya novella. Problemy teorii i istorii [Russian Impressionism: the 'Short-line' Style of Russian story. Problems of Theory and History], SPb: SPbSU, pp. 236–249.
5. *Russkij rasskaz XX veka: Antologiya* [Russian Story of the 20th Century: Anthology] (2005) / Sorokin Vl. (Comp.). M: Zakharov.
6. *Smirnov, I. P.* (1993) O smysle kratkosti [On the Meaning of Shortness]. In: Russkaya novella. Problemy teorii i istorii [The Russian Novel. Problems of theory and history]. SPb: SPbGU, pp. 1–5.

Мартыненко Григорий Яковлевич
Санкт-Петербургский государственный университет (Россия)
Gregory Martynenko
St. Petersburg State University (Russia)
E-mail: g.martynenko@gmail.com

В. Ноздева
V. Noseda

ПОЛИВАРИАНТНЫЕ КОРПУСА: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПЯТНАДЦАТИ ИТАЛЬЯНСКИХ ПЕРЕВОДОВ ПОВЕСТИ ГОГОЛЯ «ШИНЕЛЬ»

PARALLEL CORPORA WITH MULTIPLE TRANSLATION VARIANTS: A COMPARED ANALYSIS OF FIFTEEN ITALIAN TRANSLATIONS OF GOGOL'S SHORT STORY "THE OVERCOAT"

Аннотация. Данная статья посвящена сравнительному анализу пятнадцати итальянских переводов повести Николая Гоголя «Шинель», включенных в итальянско-русский параллельный корпус НКРЯ. В частности, рассматриваются две лингвистические особенности, отличающие манеру речи главного героя: частица *того* и синтаксический эллипсис. В обоих случаях наблюдаются две стратегии перевода, показывающие, что выбор переводчиков не всегда зависит от языковых критериев.
Ключевые слова. Параллельные корпуса, поливариантные корпуса, корпусный анализ языка, художественный перевод.

Abstract. This paper focuses on a comparative analysis of fifteen Italian translations of Nikolaj Gogol's short story "The Overcoat", included in the Italian-Russian parallel corpus of RNC. In particular, two linguistic features distinguishing the way of speaking of the main character are examined: the particle *togo* and syntactic ellipsis. In both cases two translation strategies emerged, showing that translators' choices not always depend on linguistic criteria.
Keywords. Parallel corpora, target-variant corpora, corpus analysis, literary translation.

1. Introduction

This paper aims at exemplifying the utility of a corpus with several translation variants, through a compared analysis of 15 versions of Gogol's short story *The Overcoat* (*Šinel'*). This kind of analysis can be useful in a number of areas: i) in Translation Studies, to investigate the properties of translated texts (translation universals); ii) in translation training, to show learners possible renderings and examine professional translators' choices; iii) in linguistics research, to deepen the semantic analysis of a given word or lexical unit. For instance, [Aijmer et al. 2006: 101] consider the corpus based translation method an excellent way to compare discourse markers cross-linguistically as it helps identify how different languages talk about different portions of reality.

1.1. Why Gogol's Šinel'

The Overcoat (*Il Cappotto*, in Italian) has had a great influence from its publication to the present day, and not only in Russia. In Italy, although the interest in Gogol's work emerged relatively late compared to other European countries [Inkova 2014: 47], *Il Cappotto* has been the most translated Russian literary work: over a century — from 1903 to 2012 — 31 versions followed one another (for a complete list see [Inkova 2014: 55–56]). For this reason, it was decided to include this short story in the project of enlargement of the Italian-Russian parallel corpus, started in 2015. Here is the list of translators whose version is now available in the RNC: C. Rebora (1922); E. Carafa D'Andria (1937); T. Landolfi (1941); O. Del Buono (1949); L. Pacini-Savoj (1957); G. Pacini (1963); A Julovic (1964); P. Zveteremich (1967); E. Bazzarelli (1980); F. Mariano (1986); S. Beffa (1986); L. De Nardis (1993); S. Prina (1994); E. Guercetti (1995); F. Legittimo (2001).

2. The focus of the analysis

Rendering the main character's speech is one of the greatest challenges while translating *The Overcoat*. This is how Gogol himself describes Akakij Akakievič's (A. A.) language, which certainly reflects his personality: «Нужно знать, что Акакий Акакиевич изъяснялся большею частью предлогами, наречиями и, наконец, такими частицами, которые решительно не имеют никакого значения»[1]. In Akakij's speech two linguistic features stand out: the obsessive use of particles (or discursive markers)[2] and elliptic constructions. My analysis will take these features into consideration, comparing how translators render them in the target language.

2.1. The particle togo

Akakij's most recurrent 'word' is the particle *togo*, defined by MAS (Malyj akademičeskij slovar') as follows: «Служит для заполнения паузы при заминке в речи, затруднении в выборе слов, иногда взамен какого-л.

[1] «It must be noticed that Akaky Akakyevitch for the most part explained himself by apologies, vague phrases, and particles which have absolutely no significance whatever».

[2] From the beginning of the 20th century, in Russian studies prevailed a formal distinction between parts of speech (časti reči) and particles (časticy). It was Vinogradov who pointed out that particles were to be distinguished from other parts of speech because of their «modal» function [Vinogradov 1975: 56]. Later, thanks to D. Paillard [Kiseleva, Pajar 2003], Russian and Western traditions converged, and Russian particles were all grouped under the category of discourse markers (diskursivnye slova — DM).

слова или словосочетания (…) // При сообщении о каком-л. неблагополучии, неприятности или о чем-л. не совсем благовидном».

In Gogol's text *togo* is in fact employed by the main character in order to fill a void left by the inability to clearly express his thoughts, but, as the narrator points out, its use can be exaggerated: Акакий Акакиевич уже заблаговременно почувствовал надлежащую робость, несколько смутился и, как мог, сколько могла позволить ему свобода языка, изъяснил с прибавлением даже чаще, чем в другое время, частиц «**того**», что была-де шинель совершенно новая, и теперь ограблен бесчеловечным образом (…)[3]

In general, discursive markers help speakers express their thoughts and link texts to reality, especially in spoken language[4]. However, Akakij seems to use these linguistic tools ineffectively, often generating a nonsense that leads to the failure of the perlocutionary aspect of his message.

My purpose was to verify if and how Italian translators produced the same nonsense in the target text.

The examples revealed two tendencies:

A) Six translators used a single term (or phrase) for all cases where the particle appeared: Rebora (1922): *in quanto*; Landolfi (1941): *coso*; Pacini-Savoj (1957): *quella cosa*; Julovic (1964): *vero e coso*; De Nardis (1993): *bè*; Guercetti (1995): *cioè*.

B) The remaining nine translators did not choose a fixed word to render *togo* in Italian, but varied according to the general meaning of the sentence, depending on the context in which it occurred. Sometimes they even omitted the particle (Ø), as in (2a):

(2) — *А я вот* ***того****, Петрович… шинель-то, сукно… вот видишь,* (…) *да вот только в одном месте немного* ***того****… на спине* (…).

(2a) «*Ecco qua, Petrovič…* ***Ø*** *il cappotto, il panno… ecco vedi,* (…),
here here Petrovic… the coat, the cloth… here (you) see

[3] Akaky Akakyevitch, who was overwhelmed with befitting awe beforehand, was somewhat confused and, as far as his tongue would allow him, explained to the best of his powers, with even more frequent "ers" than usual, that he had had a perfectly new overcoat and now he had been robbed of it (…)

[4] [Bazzanella 2006: 449] proposes a useful synthesis of what is universally accepted by linguists regarding DM: 1) they do not affect the propositional content of a sentence; 2) they are linked to the extralinguistic situation; 3) they have modal value and express attitudes and emotions; 4) they can be multifunctional and operate simultaneously on different levels.

ma in un posto, però, è, ***come dire****…. sulla schiena* (…)
but in a spot, but, (it)is, how say…. on-the back
(Bazzarelli, 1980)

As far as the first group is concerned, a diachronic distinction can be made. Most recent versions (for example, Guercetti — see (3b) — and De Nardis) display two discursive markers frequently used in contemporary spoken Italian (*cioè* and *bè*). In this way, they employ a unique translation strategy, while ensuring a more natural style of speech (even if Gogol's dialogues are to be considered caricatures, which do not fully represent the natural and flowing style of authentic language). Other translators' solutions (Rebora — as in (3a) — and Landolfi) sound much more obsolete and express even a higher degree of nonsense:

(3) — *А я вот к тебе, Петрович,* ***тогo****…*

(3a) «*Eccomi ora da te, Petròvic,* ***in quanto****…*». (Rebora, 1922)
Here-I-am now at you, Petrovič, because

(3b) «*Ecco, sono qui, Petrovič,* ***cioè****…*» (Guercetti, 1995)
Here, (I)am here, Petrovič, that-is…

In general, the abovementioned tendencies — A and B — produce two distinct effects in the final text: in the first case, the reader perceives the nonsense that probably Gogol wanted to produce, as well as the existence of a word which is typical of Akakij's linguistic repertoire. In the second case, varying every time the translations of *togo*, the idea of recurrence is missing. This choice could be dictated by stylistic reasons, by the desire to make the Italian text more natural, without the addition of superfluous and redundant elements. Group B is therefore more target-oriented, while group A seems to be truer to the original.

2.2. Syntactic ellipsis

By ellipsis we mean the omission of one or more elements of the discourse normally considered indispensable (even if there is not always unanimity in defining what 'structurally indispensable' means). Despite some unresolved questions regarding the definition of ellipses, in general it can be stated that elliptical sentences present the following features: i) they are semantically complete; ii) they are not necessarily linked with other lexical units in contiguous sentences; iii) they cannot be replaced by complete synonymic versions (by inserting the missing element), since the correct alternative is not known with certainty [Lekant 2004: 232].

This kind of sentences[5] (such as *ja domoj*, literally *I home*) are very interesting from a contrastive point of view, as in Italian (like in English) a similar omission of the verb is not possible. It should be noted, however, that unlike the excessive use of particles, which was intended to ridicule Akakij Akakievič's language, syntactic ellipsis is used by him in an absolutely ordinary way. It was assumed, therefore, that translators would choose an unmarked solution to render elliptic sentences in their own language. The analysis aimed at verifying this hypothesis.

As for the translation of *togo*, two tendencies were noted. Five translators imitated Russian syntactic structure, violating Italian grammar rules:

(4) — *Ну, а если бы пришлось Ø новую, как бы она того…*

(4a) — *Ebbene, e se per caso* **Ø** *uno nuovo, che cosa… coso…*
Well and if by accident a new what thing thing
(Landolfi, 1941)

In (4a) both the main verb and the infinitive were omitted. However, this is perceived as an anomaly in Italian. These choices were probably motivated by the intent to further underline the disconnected character of Akakij's speech.

The remaining translators used a verb in their versions, thus following the syntactic rules of the target language. All of them chose the verb *fare* (make), which refers back to Petrovič's[6] previous sentence: *вам придется новую делать* [you have to make a new one].

Let us take another example, in which two translators omitted the verb in Italian, generating the same effect as in sentence (4a):

(5) (…) *а в это время я ему* **Ø** *гривенничек и того в руку* (…)

(5a) (…) *io invece, ecco, una monetina in mano* **Ø** … (Del Buono 1949)
I whereas, here, a small-coin in hand

In the remaining versions, the position of the elided predicate was filled by a verb belonging to the semantic field of *mettere* (to put), which collocates with the following phrase *in mano* (in his hand). There were also some interesting solutions, like the expression *gli faccio scivolare in mano* (I'll slip into his hand), by Pacini-Savoj and Mariano, which gives the idea of a 'bribe' used by Akakij to corrupt Petrovič and entice him to comply with his request:

[5] For a classification of elliptic sentences see [Zemskaja et al. 1981: 201–206].

[6] His tailor.

(5b) (…) *e io allora, e quella cosa, gli* ***faccio scivolare*** *in mano* and I so and that thing to-him (I)make slip in hand
dieci copeche (Pacini-Savoj, 1957)
ten kopecks

3. Conclusions

A particular translation choice may depend first of all on linguistic criteria: for example, since ellipsis is an ordinary feature of Russian spoken language, it should not be marked in the target text. However, some choices may be influenced by other factors, like in the case of *togo*. The analysis of several translations shows that individual translators can be more source-oriented or, on the contrary, target-oriented. This carries important implications in Translation Studies and in linguistic analysis based on parallel corpora: if it is true that translated texts can be suitable for the study of different linguistic structures, it is equally evident that literary writers sometimes employ linguistic tools not according to the standard use. In Akakij's speech, for example, discourse markers mainly create nonsense, instead of performing their ordinary function.

Литература

1. *Aijmer K., Foolen A., Simon-Vandenbergen A.* (2006), Pragmatic markers in translation: a methodological proposal, Approaches to discourse particles, Elsevier, Oxford, Amsterdam, pp. 101–114.
2. *Bazzanella C.* (2006). Discourse markers in Italian: towards a compositional meaning, Approaches to discourse particles, Elsevier, Oxford, Amsterdam, pp. 449–464.
3. *Inkova O.* (2014), Tradurre il titolo: le traduzioni italiane del "Cappotto" di Gogol'. [Translating a title: the Italian translations of Gogol's "The Overcoat"], Kwartalnik neofilogiczny, LXI, 1, pp. 41–56.
4. *Kiseleva K., Pajar D.* (eds.) (2003), Diskursivnye slova russkogo jazyka: kontekstnoe var'irovanie i semantičeskoe edinstvo [Discoursive markers in Russian: contextual variation and semantic unity], Azbukovnik, Moscow.
5. *Лекант П. А.* (2004), Синтаксис простого предложения в современном русском языке, Высшая Школа, Москва.
6. *Виноградов В. В.* (1975), О категории модальности и модальных словах в русском языке, Исследования по русской грамматики: избранные труды, Наука, Москва, pp. 53–87.
7. *Земская Е. А., Китайгородская М. В., Ширяев Е. Н.* Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис Наука, Москва.

References

1. *Aijmer K., Foolen A., Simon-Vandenbergen A.* (2006), Pragmatic markers in translation: a methodological proposal, Approaches to discourse particles, Elsevier, Oxford, Amsterdam, pp. 101–114.
2. *Bazzanella C.* (2006). Discourse markers in Italian: towards a compositional meaning, Approaches to discourse particles, Elsevier, Oxford, Amsterdam, pp. 449–464.
3. *Inkova O.* (2014), Tradurre il titolo: le traduzioni italiane del "Cappotto" di Gogol'. [Translating a title: the Italian translations of Gogol's "The Overcoat"], Kwartalnik neofilogiczny, LXI, 1, pp. 41–56.
4. *Kiseleva K., Pajar D.* (eds.) (2003), Diskursivnye slova russkogo jazyka: kontekstnoe var'irovanie i semantičeskoe edinstvo [Discoursive markers in Russian: contextual variation and semantic unity], Azbukovnik, Moscow.
5. *Lekant P. A.* (2004), Sintaksis prostovo predloženija v sovremennom russkom jazyke [The syntax of simple sentences in contemporary Russian]. Vysšaja Škola, Moskva.
6. *Vinogradov V. V.* (1975) O kategorii modal'nosti i modal'nych slovach v russkom jazyke [The category of modality and modal words in Russian], Issledovanija po russkoj grammatiki: izbrannye trudy [Studies on Russian grammar: selected works], Nauka, Moskva, pp. 53–87.
7. *Zemskaja E. A., Kitajgorodskaja M. V., Širjaev E. N.* (1981). Russkaja razgovornaja reč'. Obščie voprosy. Slovoobrazovanie. Sintaksis [Russian spoken language. General issues. Word formation. Syntax], Nauka, Moskva.

Валентина Ноземда
Католический университет Святого Сердца (Италия)
Valentina Noseda
Catholic University of the Sacred Heart (Italy)
E-mail: valentina.noseda@unicatt.it

С. Б. Потемкин
S. B. Potemkin

КОРПУС ФОЛЬКЛОРНЫХ ТЕКСТОВ И КЛАСТЕРИЗАЦИЯ УКАЗАТЕЛЕЙ СЮЖЕТОВ

CORPUS OF FOLK TEXTS AND PLOTS IDEXES CLUSTERING

Аннотация. Корпус сказок для русского и английского языка составлен с целью включения в научный оборот фольклорного материала, имеющегося в машиночитаемой форме. Корпус позволяет создавать конкордансы, частотные словари, вызывать полные тексты и пр. На основе корпуса разработаны средства вычисления семантической близости сюжетов, графического представления сюжетов и кластеризации полученных графов. Семантическая близость в данном случае понимается как число совпадающих лексем с учетом последовательности их появления в тексте.
Ключевые слова. сказка, сюжет, граф, семантическая близость, кластеризация.

Annotation. The corpus of fairy tales for Russian and English is compiled with the goal of introduction of folklore material in machine-readable form. The corpus allows one to create concordances, frequency dictionaries, invoke full texts, etc. On the basis of the corpora tools are developed for calculating the semantic proximity of plots, the graphical presentation of plots and clustering of the resulting graphs. In this context semantic proximity is understood as the number of matching tokens, taking into account the sequence of their appearance in the text.
Keywords. fairy tale, plot, graph, semantic proximity, clustering.

Введение

Сказки, как часть фольклора, являются, возможно, самой устойчивой формой сохранения культурной традиции народа. В отличие от подавляющей части достижений ушедших цивилизаций — мифов, утративших силу законов, моральных принципов и технических приспособлений, сказки дошли до нас практически в первоначальном виде. Исследование такого уникального явления как сказка приводит к важным для нашего современника выводам о значимости ее культурного влияния.

В России сказки, собранные выдающимся русским фольклористом Александром Николаевичем Афанасьевым (1826—1871) удивительное и уникальное явление в литературе и шире — в русской культуре Всего для издания было отобрано свыше 600 текстов. Последнее полное издание Сказок вышло в 1984 году. По этому изданию создан электронный ресурс[1], на основе которого составляется Корпус фольклорных текстов.

[1] http://feb-web.ru/feb/skazki/default.asp?/feb/skazki/texts (Дата доступа: 19.02.2019).

В других странах обширные коллекции народных сказок систематизированы, в частности, ученым из США, Ашлиманом [Ashliman 2004] систематизировавшим более 1000 сюжетов разных народов. Тексты сказок в этом собрании представляют собой скорее синопсис сказки в переводе на английский[2].

Подготовка фольклорного корпуса сказок Афанасьева

Работа со Сборником включала несколько этапов, как чисто технических, так и требующих привлечения специалистов-филологов или студентов, уже прошедших определенную филологическую подготовку. Вначале каждая сказка была переписана с вышеупомянутого ресурса в отдельный файл в формате UTF8 для сохранения разметки (знаки ударения и пр.). Затем текст каждой сказки был разбит на отдельные словоформы с отделением знаков препинания. Были выделены последовательности, «предложения», оканчивающиеся на знаки: точка, вопросительный и восклицательный знак, а также двоеточие и точка с запятой. Был проведен морфологический анализ каждой словоформы на основе словаря А. А. Зализняка[3] и частичный синтаксический анализ, включая согласование существительное — прилагательное, предлог — существительное, глагол — существительное и пр. Многие архаичные, диалектные, иноязычные (в основном белорусские и украинские) слова, не найденные в Словаре Зализняка, были выделены в отдельный список, насчитывающий около 4000 словоформ. Этот список с привязкой каждой словоформы к сказке, где она встречается, давался студентам для обработки.

Указатели фольклорных сюжетов

Фольклорные сюжеты обычно классифицируются по системе, введенной финским исследователем Аарне [Aarne 1910]. По этой системе составлены указатели фольклора разных народов, что автоматически вводит в широкий научный контекст вновь собранный или восстановленный по письменным источникам первичный материал. Система Аарне относит каждую «сказку» к определенному сюжету, без вычленения в ней составляющих сюжетных ходов и обобщения персонажей — если в сказке говорится о черте, его ролевая привязка не может

[2] Ashliman: http://www.pitt.edu/~dash/ashliman.html (Дата доступа: 19.02.2019).
[3] http://starling.rinet.ru/download/zaliznia.exe (Дата доступа: 13.02.2019).

быть заменена на, скажем, колдуна, хотя функции этих персонажей в целом будут совпадать. Каталог Аарне был переведен на многие европейские языки, на русский язык его перевел Н. П. Андреев [Андреев 1929]. В системе Томпсона [Thompson 1973], являющейся развитием системы Аарне, сложные сюжеты разложены на несколько элементов и для каждого элемента приводятся различные реализации. К сожалению, составители наиболее представительного русского указателя сюжетов СУС: [Бараг и др. 1979] воспользовались системой Аарне без томпсоновских дополнений.

В мировой науке постоянно составляются новые указатели, относящиеся к разным национальным традициям и жанрам. На русском материале существуют указатели различных фольклорных жанров том числе электронные, наиболее полным из которых является ресурс[4]. Представляет интерес указатель детских «страшилок»[5]. Компьютерная система СКАЗКА [Рафаева 1998], реализованная в СУБД Starling, дает возможность отвечать на различные типы запросов относительно волшебных сказок.

Потребность в кластеризации

Построенные указатели основаны на интуиции исследователей и, таким образом, не чужды субъективности. Необходимы формальные методы отнесения того или иного сюжета к той или иной рубрике указателя сюжетов. С этой целью мы применяем методы кластерного анализа и многомерного шкалирования. Для исследования использован ресурс[6], позволяющий получить более 4000 страниц с описаниями сюжетов и снабженный идентификатором сюжета вида ***andr_87.htm.*** Была предпринята попытка выделить особые «сказочные» термины, сравнивая частотность лемм в словнике сказочных сюжетов и в частотном словаре общей лексики Шарова[7]. Дальнейший анализ проводился только на знаменательных словах, составляющих более половины словоупотреблений. Сюда же можно добавить слова, для которых часть речи не определилась, это скорее всего, редко встречаемые архаизмы и регионализмы. Список наиболее употребительных лексем использовался для определения самых богатых кластеров сюжетов,

[4] http://ruthenia.ru/folklore (Дата доступа: 13.02.2019).
[5] http://www.unn.ac.ru/folklore/sukaz.htm (Дата доступа: 13.02.2019).
[6] http://www.ruthenia.ru/folklore/sus (Дата доступа: 13.02.2019).
[7] http:/www.artint.ru/projects/frqlist.php (Дата доступа: 13.02.2019).

напр., кластер, содержащий лексему «Царь» с частотностью 310, должен включать значительно больше элементов, чем кластер для слова «Избушка» с частотностью 5.

Метод последовательного сопоставления

Поскольку сюжет фольклорной сказки развивается последовательно, можно принять в качестве гипотезы, что в паре совпадающих сюжетов последовательности слов также должны совпадать. Для проверки этой гипотезы был составлен алгоритм сопоставления последовательностей, основанный на методе динамического программирования. Среди всех возможных путей от начала до конца предложений алгоритм динамического программирования выбирает тот, который включает наибольшее число совпадающих лексем. Это число и принимается за меру близости двух сюжетов.

Визуализация взаимного расположения объектов (в нашем случае — сюжетов сказок) на основании меры их сходства/различия позволяет наглядно представить совокупность сюжетов, близких к заданному, а также эвристически выделить кластеры близкорасположенных объектов. Задача визуализации решается методом многомерного шкалирования, применимого в случае, если выборка содержит тысячи объектов, (в нашем случае число пар сюжетов, совпавших хотя бы по одному слову– более 600 тысяч), и, следовательно, пространство объектов существенно многомерно.

Задача многомерного шкалирования (multidimensional scaling, MDS) заключается в отображении пространства большой размерности в пространство меньшей размерности, в частности, размерности 2, что позволяет отобразить выборку в виде множества точек на плоскости (scatter plot). Плоское представление отражает основные структурные особенности многомерной выборки, в частности, семантическое расстояние между сюжетами и кластерную структуру выборки. Поэтому многомерное шкалирование часто используют как инструмент предварительного анализа для понимания данных. Разработанная при участии автора программа [Кедрова, Потемкин, 2007] позволяет построить scatter plot для сильно разреженной матрицы расстояний между сюжетами сказок. На рис. 1 представлена окрестность сюжета А161. Окрестность сюжета «преданная собака» разделилась на 2 кластера, с центрами в точках ANEK886 и А198.

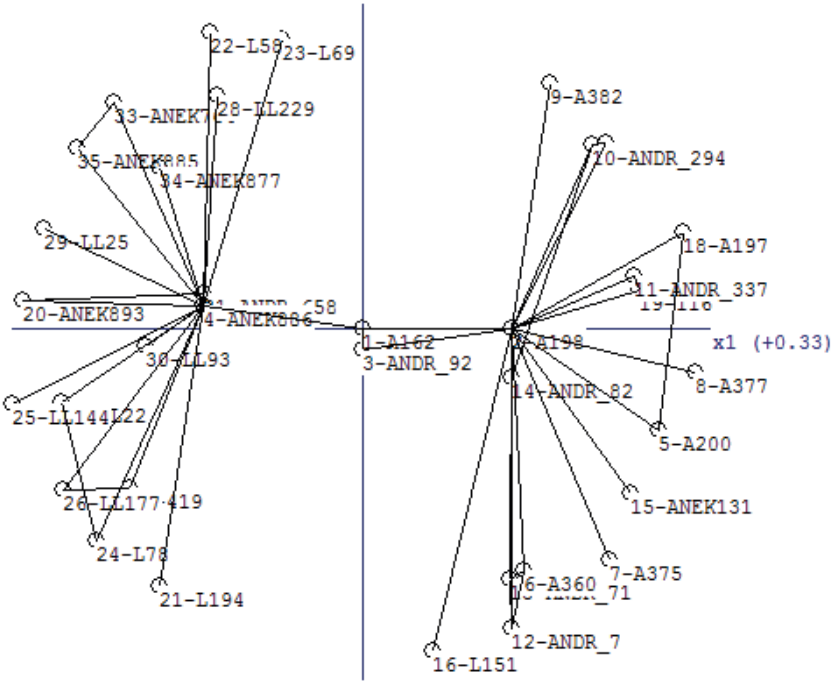


Рис. 1. **Центр: 1 — А161 Преданная собака: незаслуженно наказывается. Кластер А): Баба хуже черта; Кластер Б): Старая хлеб — соль забывается**

Аналогичная методика была применена к сборнику сюжетов Ашлимана. В этом сборнике каждый сюжет описан достаточно пространным текстом, средняя длина которого составляет около 800 слов. Для сравнения в Индексе Андреева (АН) средняя длина описания сюжета сказки составляет около 30 слов. Отсюда для сравнения сюжетов из сборника Ашлимана (АШ) пришлось не ограничиваться последовательным сопоставлением лексем. Мы выделили в текстах сюжетов АШ последовательности глаголов в предположении, что такая последовательность составляет фрейм или структуру сюжета. Структуры сюжетов уже можно сопоставлять последовательно. Для более точного сопоставления сравнивались не сами глаголы V, а их гиперонимы, hyper(V). Полученные результаты (расстояние между сюжетами) подвергались многомерному шкалирования и проецировались на плоскость (Рис. 2)

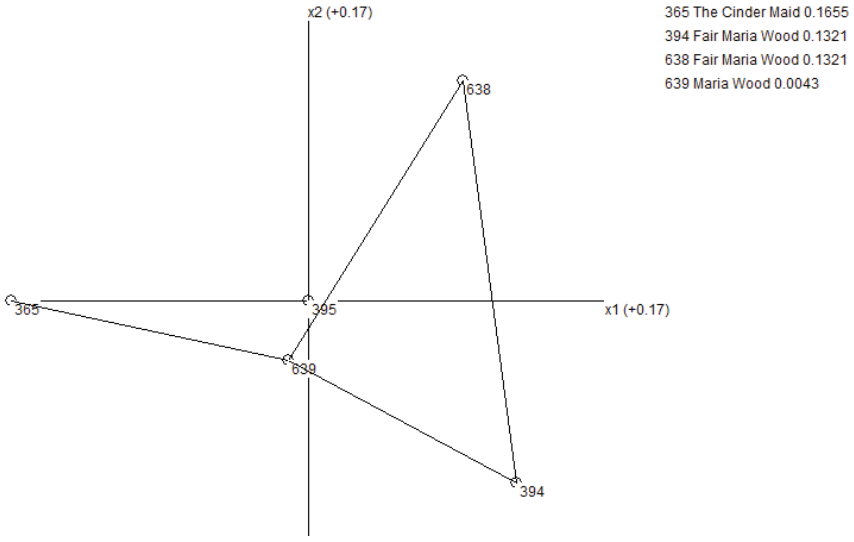


Рис. 2. **Окрестности сказки Cindarella (Золушка) из собрания Ашлимана**

Заключение

Корпус фольклорных текстов позволяет изучать подлинно народный язык конца 19 — начала 20 века в исчерпывающем и объективном смысле впервые (чего нельзя сказать о многих, селективных и поэтому субъективных исследованиях).

В статье также описан новый подход к формированию указателя сказочных сюжетов на основе анализа лексики, определения расстояния между сюжетами и кластеризации сюжетов, которая выполняется путем визуального изучения двумерных образов, полученных методом многомерного шкалирования. Исследование проводилось на материале русских сказок Афанасьева и собрании сказок разных народов Ашлимана в переводах на английский. Результаты исследования могут быть использованы для выявления сходства и кластеризации между однородными объектами при разработке различных онтологий.

Литература

1. *Андреев Н. П.* (1929), Указатель сказочных сюжетов по системе Аарне. Л. 120 с.
2. *Бараг Л. Г. и др.* (1979), Сравнительный указатель сюжетов: Восточнославянская сказка. Л. 437 с.

3. *Кедрова Г. Е. , Потемкин С. Б.* (2007), Использование корпуса параллельных текстов для пополнения специализированного двуязычного словаря // Труды и материалы III Международного Конгресса исследователей русского языка «Русский язык: исторические судьбы и современность», М., сс. 627–628.
4. *Рафаева А. В.* (1998), Полуавтоматический анализ волшебных сказок в компьютерной системе СКАЗКА // Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям, том 2, сс. 701–706.
5. *Aarne A.* (1910), Verzeichnis der Maerchetypen. Helsinki, p. 64.
6. *Ashliman D. L.* (2004), Folk and Fairy Tales: A Handbook. Westport (Connecticut), New York, and London: Greenwood Press, ISBN 0-31332810-2, p. 256.
7. *Thompson S.* (1973), The Types of the Folktale. Helsinki, p. 588.

References

1. *Andreev N. P.* (1929), Ukazatel skazochnykh sujetov po sisteme Aarne [Index of fairy tales according to Aarne's system]. Leningrad. 120 p.
2. *Barag L. G. et al.* (1979), Sravnitelnyi ukazatel sujetov. Vostochnoslavianskaia skazka [Comparative Index of Plots: East Slav Tale.]. Leningrad. 437 p.
3. *Kedrova G. E., Potemkin S. B.* (2007), Ispolzovanie korpusa parallelnykh tekstov dlia popolnenia specializirovannogo dvuiazychnogo slovaria [Use of the corpus of parallel texts for expanding the specialized bilingual dictionary] // Proceedings and Materials of the III International Congress of Researchers of the Russian Language "Russian Language: Historical Destiny and the Present" Moscow, pp. 627–628.
4. *Rafaeva A. V.* (1998), Poluavtomaticheskii analiz volshebnykh skazok v computernoi systeme SKAZKA [Semi-automatic analysis of fairy tales in the computer system TALE] // Proceedings of the International Dialog'98 seminar on computational linguistics and its applications, volume 2, pp. 701–706.
5. *Aarne A.* (1910), Verzeichnis der Maerchetypen. Helsinki, p. 64.
6. *Ashliman D. L.* (2004), Folk and Fairy Tales: A Handbook. Westport (Connecticut), New York, and London: Greenwood Press, ISBN 0-31332810-2, p. 256.
7. *Thompson S.* (1973), The Types of the Folktale. Helsinki, p. 588.

Потемкин Сергей Борисович
Московский Государственный Университет им. М. В. Ломоносова (Россия)
Potemkin Sergey
Moscow State University (Russia)
E-mail: prolexprim@gmail.com

B. Rabu, F. Mélanie and T. Poibeau

"A NOVEL OF CHARACTER": TOWARDS THE AUTOMATIC ANNOTATION OF CHARACTERS IN A LARGE CORPUS OF FRENCH NOVELS

Abstract. In this paper, we apply named entity recognition techniques to a corpus of literary texts, i.e. French novels from the 18th, 19th and 20th century. We obtain results that are usable but could be improved by using advanced annotation techniques. We discuss the use of active learning in this context, as well as the different applications that could be derived from this kind of annotation. In particular, we show that the automatic annotation of large literary corpora makes it possible to check whether traditional classifications exhibit specific structural patterns that could be identified automatically.
Keywords. Named Entity Recognition; Digital Humanities; Literature Analysis; Text Mining; Distant Reading.

1. Introduction

The recent availability of large literary corpora in different languages has open new pathways for the study of literature. This approach is often called "distant reading" (Moretti, 2013) since corpora are then too large to be read directly and can only be accessed through specific tools that create a "distance" between the text and the reader. This approach has given birth to new research avenues and researchers are now able to observe tendencies over a large number of texts, instead of focusing on isolated observations concerning a few novels.

A specific research programme includes for example the investigation of the structure of novels, through the notion of "character": How central are the different characters of a novel? How do they interact with each other in the course of the novel? In other words, are there specific patterns that emerge from different novel traditions, from different period of times or from different subgenres? (Piper *et al*., 2017)

There are now several tools available for different languages that are able to recognize person names in texts and, more generally, named entities like locations, artefacts or organizations. Named entity recognition is a well-established task, but existing tools are far from perfect: they make errors and need to be re-trained to reach acceptable performance on different corpora (Finkel et al., 2005). Their performance over literary texts also need to be properly evaluated, as they are generally trained on news or other kinds of Web sources.

In this paper, we propose an experiment on a corpus of French novels. We annotate person names, as well as other related text sequences (like ti-

tles, functions, or occupations) that can be used to refer to a character. The question of "what to annotate" is a highly complicated one, and we will just give a brief overview of our annotation principles below. We first present the corpus, then our annotation scheme and the tool we used for our experiments. We then present our results on the different novels, we discuss these results and conclude with some observations for future work.

2. The Corpus

For our study, we chose different novels from the 18th, 19th and 20th century. The choice is of course quite subjective as a large number of novels is directly available online in an electronic format. We wanted to get a balanced corpus among the three centuries considered.

Title	Author	Publication date	Size (approx. # of words)
De l'esprit des lois	Montesquieu	1748	65.000
Candide	Voltaire	1759	32.000
L'an 2440	L-S. Mercier	1771	93.000
Les liaisons dangereuses	P. C. de Laclos	1782	140.000
Les Rêveries du promeneur solitaire	J-J. Rousseau	1782	40.000
Notre-Dame de Paris	V. Hugo	1831	156.000
La Maison Nucingen	H. de Balzac	1838	34.000
Madame Bovary	G. Flaubert	1857	116.000
Alice au pays des merveilles	L. Carroll	French: 1869 Original: 1865	30.000
À l'ombre des jeunes filles en fleurs	M. Proust	1919	205.000
Les Faux-Monnayeurs	A. Gide	1925	115.000
La Gloire de mon père	M. Pagnol	1957	47.000

3. Annotation Principles

One of the most difficult part of the task is to define the entities that should be annotated. Some examples are easy to recognize and annotate, but lots of others are difficult.

Person names: Both fictive and real names can be found in novels. Person's names correspond to proper names like first names (*Odette*), last names (*Swann*) or a combination of both (*Odette Swann*). These proper names can be preceded with a title (*Madame de Crécy*, *M. de Norpois*), which can lead to complex noun phrases, especially with nobility titles (*Son Éminence monseigneur le cardinal de Bourbon*, in *Notre-Dame de Paris* from Victor Hugo). The same phenomenon is observed with function or occupation names (*le marquis de Norpois*, *le professeur Cottard*, or *l'abbé Frayssinous*). Texts often contain references to characters through their function, occupation or title, without mentioning any proper name, especially when the character has already been introduced or when there is no ambiguity left with just the title or the function mentioned (*le Principal*, *le Vicomte*). Generic groups of people are not annotated as they cannot be directly considered as characters (*les Anglais*, *les Parisiens*), but specific groups must be annotated (like *les Swann*, in Proust's *A l'ombre des jeunes filles en fleurs*). Other difficult cases are words like *God* or *the Divinity*, whose status is unclear.

Other entities: The software we used for the annotation by default also annotates other kinds of entities (location names, companies, *etc.*). This is of course interesting for the study of literary texts, especially location names since one could imagine a joint study of people (characters) and places. However, this is outside the scope of the present study and, in what follows, we will just focus on person names.

We cannot give all the details used for the annotation here, but the interested reader can refer to existing guidelines, for example the one proposed by Rosset *et al.* (2011) that offers valuable principles for French, especially to practically solve difficult cases. Other guidelines exist for other languages but the most important principle is to be consistent throughout the annotation phase, since a part of the decisions to take is subjective, as there is no formal distinction between named entities and other referential expressions in natural languages.

4. The Annotation Tool

We used a tool called SEM for our experiments (Dupont, 2018). SEM is an open piece of software, freely available online, and based on machine learning techniques. More specifically SEM is based on Wapiti (Lavergne *et al.*, 2010), a CRF toolbox (Conditional random Fields, Lafferty *et al.*, 2001). CRF are simpler than neural networks, and they obtain competitive results for the annotation of sequences. They are thus especially indicated for tasks

like named entity recognition, since our goal is to recognize local and continuous sequences of texts (sequences without gaps). SEM can also very easily be trained using an annotation interface. Practically the end user can just annotate a few examples before training a new model that can be tested on new data, which is what we needed to do since our results will highly depend on the training phase using a representative sample of our corpus.

We have trained a new model from scratch for each century, but this is of course far from optimal since it would normally require annotating huge quantities of data to achieve reasonable performance. There is moreover a serious risk of overfitting since we train a new model for each novel / century. One solution to this problem would be to dynamically update an existing model based on new data. Recent machine learning techniques makes this approach possible, but it has not been explored yet in our context. The other approach consists in using active learning techniques to accelerate and optimize the annotation phase. In our case, unlabelled data is abundant but manually labelling is expensive. Learning algorithms can actively query the user for labels, making it possible to dynamically and automatically identify interesting examples for training, i.e. discriminative and ambiguous examples that the system cannot annotate directly (typically, because contradictory indices can be found in the context). This approach is for example the one already used by Prodigy, the annotation tool developed in relation with Spacy by Montani and Honnibal (2017).

5. Results and Discussion

The results are given in table 1. All the results are expressed using F-measure, the harmonic mean of precision (the percentage of sequences that are accurately recognized among what has been recognized by the system) and recall (the percentage of sequences actually recognized among all those that should have been recognized).

Table 1. **annotation results (all the results are expressed using F-measure)**

Annotation model	18th century corpus	19th century corpus	20th century corpus
18th century	**0,68**	0,63	0,68
19th century	0,61	**0,70**	0,67
20th century	0,62	0,69	**0,73**
All	**0.72**	**0,77**	**0,86**

We can make two main observations: *i*) logically, a model trained on texts from a specific century works better on texts from that century, and vice versa (e.g. novels from the 19th century are more accurately analysed by the model trained on 19th century texts, than on the one trained on 18th or 20th century texts) and *ii*) more surprisingly maybe, the global model aggregating all the different sub-models works better than any other one on all the different corpora by a significant margin (*i.e.* by a statistically significant margin).

We can also observe that our results so far are not very impressive. There are several reasons for this, but the first one is clearly due to our approach. For both practical and theoretical reasons, our training sets are quite small, because we did not have enough time to provide large annotation sets and because we also wanted to avoid overfitting since we just considered a few works and a few authors (see above).

However, our results show that it is possible to develop only one model to annotate the different corpora, although each novel is specific. This is probably true because French has not evolved so much from the 18th century[8]. However, the model may still need some adaptation depending on the novels considered (some are known to have very specific ways to name people for example). This is why the ability to update an existing model and use active learning for training would be especially interesting in our case. It would also help to solve the annotation issue, since active learning makes it possible to reach high performance, while reducing drastically the annotation effort.

Lastly, we may want to explore neural network techniques for annotation, which are known to be slightly more efficient than CRF (Lample *et al.*, 2016), although, as said above, CRF as such are simpler and quite powerful for the annotation of continuous sequences.

6. Conclusion

We have presented an experiment aiming at showing that it is possible to develop accurate models for the annotation of characters in a corpus of French novels. Our current results, although far from perfect, are nevertheless sufficient for practical use. The next steps will consist in annotating many more novels and then develop large scale character analysis, i.e. detecting patterns in character networks, character interactions or character

[8] This is also why we did not include older texts in our corpus: it is known that French has dramatically evolved in the 16th century, and even in the 17th, so it is advisable to be careful when dealing with texts prior to 1700 in French.

structures. For example, some novels put forward one character, or a few number of characters, whereas some others are based on the interaction of a larger group of characters. These characteristics are known to be relevant for literary studies. The approach makes it possible to group together different kinds of novels, and also to check whether traditional classifications exhibit specific patterns, following some proposals made by Moretti (2005).

References

1. Dupont, Y. (2018). Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique. Conf. Traitement Automatique des Langues Naturelles (TALN), 2018.
2. Finkel, J. R.; Grenager, T.; and Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 363–370.
3. Lafferty, J.; McCallum, A.; Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. of the 18th International Conf. on Machine Learning,* Morgan Kaufmann, p. 282–289, 2001.
4. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. (2016). Neural architectures for named entity recognition. *Conf. of the North American Chapter of the Association for Computational Linguistics* (NAACL). San Diego, USA, p. 260–270.
5. Lavergne, T., Cappé, O. and Yvon, F. (2010). Practical Very Large Scale (CRFs). *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL), Uppsala, Sweden, p. 504–513.
6. Montani, I. and Honnibal M. (2017). Prodigy: A new annotation tool for radically efficient machine teaching. https://explosion.ai/blog/prodigy-annotation-tool-active-learning.
7. Moretti, F. (2005). Graphs, Maps, Trees. Abstract Models for a Literary History. London/New York, Verso, 2005, 160 p.
8. Moretti, F. (2013). *Distant Reading*. London/New-York, Verso, 244 p.
9. Piper, A.; Algee-Hewitt, M.; Sinha, K.; Ruths, D.; Vala, H. (2017). Studying Literary Characters and Character Networks. *Digital Humanities 2017*, Montreal, Canada.
10. Rosset, S., Grouin, C. and Zweigenbaum, P. (2011). Entités nommées structurées: guide d'annotation Quaero. Notes et Documents 2011-04, LIMSI, Orsay, France.

Benjamin Rabu
Frédérique Mélanie
Thierry Poibeau
Laboratoire LATTICE (CNRS & ENS / PSL et U. Sorbonne nouvelle / USPC)
Paris, France
E-mail: benjamin.rabu@icloud.com
E-mail: frederique.melanie@ens.fr
E-mail: thierry.poibeau}@ens.fr

Model	F-measure per novel					Mean F-measure
18th century model	De l'esprit des lois	Candide	L'an 2440	Les liaisons dangereuses	Les rêveries du promeneur solitaire	**0.68**
	0.77	0.62	0.69	0.66	0.66	
		Notre-Dame de Paris	La Maison Nucingen	Madame Bovary	Alice au pays des merveilles	**0.63**
		0.77	0.56	0.60	0.60	
			À l'ombre des jeunes filles en fleurs	Les Faux-Monnayeurs	La Gloire de mon père	**0.67**
			0.75	0.73	0.55	
19th century model	De l'esprit des lois	Candide	L'an 2440	Les liaisons dangereuses	Les rêveries du promeneur solitaire	**0.61**
	0.55	0.65	0.69	0.63	0.53	
		Notre-Dame de Paris	La Maison Nucingen	Madame Bovary	Alice au pays des merveilles	**0.70**
		0.76	0.65	0.73	0.68	
			À l'ombre des jeunes filles en fleurs	Les Faux-Monnayeurs	La Gloire de mon père	**0.67**
			0.70	0.74	0.57	
20th century model	De l'esprit des lois	Candide	L'an 2440	Les liaisons dangereuses	Les rêveries du promeneur solitaire	**0.62**
	0.76	0.64	0.71	0.50	0.49	
		Notre-Dame de Paris	La Maison Nucingen	Madame Bovary	Alice au pays des merveilles	**0.69**
		0.66	0.76	0.75	0.59	
			À l'ombre des jeunes filles en fleurs	Les Faux-Monnayeurs	La Gloire de mon père	**0.73**
			0.74	0.79	0.64	

Т. Г. Скребцова
T. G. Skrebtsova

СТРУКТУРА НАРРАТИВА В РУССКОМ РАССКАЗЕ НАЧАЛА XX ВЕКА[1]

NARRATIVE STRUCTURE OF THE RUSSIAN SHORT STORIES IN THE EARLY XX CENTURY

Аннотация. Статья посвящена рассмотрению структуры повествования в русских рассказах начала XX века на материале соответствующего корпуса. Основное внимание уделяется нарушениям принятой композиционной схемы, которые наблюдаются приблизительно в 30 % от общего числа рассказов. Автор анализирует альтернативные варианты композиции, причины отступления от привычной нарративной структуры, а также корреляции между нестандартной структурой повествования и его содержанием.
Ключевые слова: нарративный текст, нарративная структура, кульминация, развязка.

Abstract. The focus of the paper is on the corpus of Russian short stories of the early XX century and their narrative structure, in particular. Non-traditional narrative structure appears to be present in roughly 30 % of the whole set. The paper puts forward reasons for such a high percentage, identifies alternative structures and outlines possible correlations between the non-traditional narrative structure and the short story semantics.
Keywords: narrative text, narrative structure, climax, resolution.

1. Постановка задачи

Наше исследование построено на материале корпуса рассказов русских писателей, созданного усилиями научного коллектива под руководством проф. Г. Я. Мартыненко. Корпус включает в себя рассказы, опубликованные в первой трети XX века в составе отдельных сборников или журнальных выпусков, и на настоящий момент насчитывает несколько тысяч единиц (подробнее о принципах построения корпуса см.: [Мартыненко и пр. 2018]). На основе этого корпуса сформирована выборка из 100 рассказов, написанных в период 1900-1917 гг. Она служит начальным полигоном для разностороннего изучения материала и выдвижения гипотез, которые в дальнейшем будут проверяться на более обширном массиве текстов. Означенный принцип работы при-

[1] Исследование поддержано грантом РФФИ № 17-29-09173 «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

меняется и при анализе рассказов с точки зрения их композиционной структуры, чему и посвящена настоящая работа.

Изучение композиции повествовательного текста имеет давнюю традицию в филологических исследованиях. При некоторых несущественных разночтениях, стандартная структурная схема выглядит как движение от завязки через развитие действия к кульминации и далее к развязке. Считается, что подобная организация текста является отличительной особенностью повествования — в отличие от других функционально-смысловых типов речи, прежде всего описания и рассуждения.

При рассмотрении вышеупомянутой выборки, однако, оказалось, что эта привычная схема нередко нарушается: лишь около 70 из 100 рассказов построены по классическому принципу. Наше внимание будет сосредоточено на оставшихся 30. Нас интересуют альтернативные варианты композиции, причины отступления от привычной нарративной схемы, а также корреляции между нестандартной структурой повествования и его содержанием. Обратимся к материалу.

2. Анализ материала

Есть рассказы, вообще лишенные кульминации и развязки. Это можно наблюдать, к примеру, в произведениях, описывающих внутренний монолог героя — его мысли, рассуждения, воспоминания. Таковы рассказы Ю. Балтрушайтиса «Капли» (1901) и Б. Никонова «Накануне отъезда» (1906). В них соблюдается обязательное для нарратива требование темпоральности, предполагающее временно́е упорядочение событий (в широком смысле), но отсутствует причинность (ср. [Cortazzi 2002: 85]). Их структура очевидным образом не соответствует традиционному взгляду на строение рассказа, ср.: «Минимальная законченная фабула состоит в переходе от одного состояния равновесия к другому. Идеальный рассказ начинается с некоторого устойчивого положения, которое затем нарушается действием какой-то силы. Возникает состояние неравновесия; благодаря действию некоторой противоположной силы равновесие восстанавливается; новое равновесие подобно исходному, но они никогда не тождественны» [Тодоров 1978: 453].

Другой вариант отсутствия кульминации и развязки представлен описаниями рутинного течения жизни — на примере конкретного дня

(Б. Верхоустинский «Лесное озеро» (1912)), типичного дня (Ф. Крюков «У окна» (1909)) или более долгого периода (Г. Гребенщиков «Как гуляет Тихоныч» (1909)), череды событий и перипетий (В. Башкин «Потянуло» (1910), И. Бунин «Хорошая жизнь» (1911), З. Гиппиус «Сумасшедшая» (1903)). В каждом из этих рассказов можно выделить ряд эпизодов, однако в совокупности они не выстраиваются в стандартную схему, предполагающую поступательное развитие действия «по нарастающей» вплоть до некоего пика, после которого напряжение спадает и ситуация разрешается (ср. англоязычные термины *rising action u falling action*).

Есть также рассказы-зарисовки, которые при наличии продолжения могли бы восприниматься в качестве завязки истории, ср. рассказы Н. Лейкина «На хрен да на редьку, на кислую капусту» (1906), А. Серафимовича «Жара и грузчики» (1902). Неполнота композиции отчасти компенсируется выраженной оценкой, которая считается одной из двух основных социальных функций нарратива (другая — референциальная) [Labov, Waletsky 1967: 33].

Оценка сильно (хотя и имплицитно) выражена и в рассказе Л. Толстого «Ягоды» (1905), где праздный образ жизни семьи барина противопоставляется труду крестьянских детей. Автор здесь использует свой любимый прием — композиционный параллелизм [Эрлих 1996: 243]. Специфика построения рассказа (переход от одной группе персонажей к другой) также препятствует построению стандартной схемы повествования.

Интересный случай представляют собой рассказы, насыщенные действием и в то же время лишенные выраженной кульминации и, как следствие, развязки. Это характерно в частности для произведений, в которых нашли отражение предреволюционные настроения в российском обществе. Помещик обходит свое имение, опасаясь «красного петуха» (Л. Авилова «Власть» (1906)), солдаты усмиряют крестьянский бунт (В. Свенцицкий «Солдат задумался» (1906)), пристав умирает от пули студента (Л. Кармен «За что?!» (1904)). Герои нескольких рассказов участвуют в выступлениях и митингах (Б. Зайцев «Завтра!» (1906)) и оказываются заключены в тюрьму — это и студент (М. Горький «Тюрьма» (1904)), и гимназистка (Г. Яблочков «Баррикада» (1913)), и «политические» (Ф. Крюков «У окна» (1909)).

3. О возможных причинах нарушения стандартной нарративной структуры

Рассматривая всю совокупность отмеченных случаев, можно выделить несколько причин отступления от стандартной нарративной структуры. Одна достаточно банальна и связана с недостаточной одаренностью авторов произведений. В эпоху, когда литературоведение не было развитой наукой, писателям приходилось полагаться исключительно на собственную интуицию и талант. Проект создания корпуса русских рассказов предполагает сбор возможно большего числа рассказов без оглядки на их художественный уровень, и при создании выборки из 100 единиц этот фактор также не принимался во внимание.

Есть и другая, более глубокая, причина. Можно утверждать, что существует корреляция между основной темой рассказа и его композицией. Подробному анализу семантики русских рассказов начала XX века будет посвящено отдельное исследование, и тогда можно будет статистически соотнести между собой доминирующую тему и характер композиции. Однако даже навскидку нетрудно заметить, что отсутствие выраженной кульминации и развязки характерно для рассказов, повествующих о тоскливых серых буднях и неизбывной нищете, в которую погружены персонажи. Напротив, мы не найдем эту особенность в произведениях, основным содержанием которых является любовь, ревность, измена, убийство или самоубийство.

Тот факт, что «ущербная» структурная схема наблюдается в рассказах, повествующих о революционных настроениях той поры, вероятно, можно объяснить неясностью, амбивалентностью тогдашней политической ситуации. Протест зреет, но еще не принимает массового характера и не приводит к результату; неразрешенность конфликта находит отражение в незаконченности композиции.

Настроением безысходности проникнуты и другие произведения на общественно значимые темы — о русско-японской войне (В. Вересаев «В мышеловке» (1906)) и переселении крестьян из черноземных областей на Урал (П. Заякин-Уральский «Переселенцы» (1912)). Несмотря на преобладание динамичных эпизодов, они лишены общей кульминации и развязки. Возможно, это связано с желанием авторов передать свое отношение к затяжной и непобедоносной войне, проекту освоения Сибири и заселения новых земель. В некотором смысле здесь можно усмотреть параллель с исследованием Т. А. ван Дейка, об-

наружившим отсутствие развязки в половине устных рассказов белых голландцев о мигрантах [Дейк 1989: 268–304].

Возвращаясь к художественным произведениям, отметим, что характер развязки исторически изменчив и зависит от эпохи. В исследовании, посвященном сравнительному анализу завершающих фраз в американских рассказах, было обнаружено, что в первой половине XIX века рассказы имели глобальную, объективную и четко выраженную сюжетную развязку (обычно смерть персонажа или решение ключевой проблемы), в то время как в XX веке она становится более имплицитной, субъективной, связанной с локальными темами повествования [Lohafer 1994]. Разумеется, периодизация русской и американской литературы различается, но в целом этот фактор не следует исключать из рассмотрения. Можно предположить, что выявленный нами процент нарушения структурной схемы повествования в русских рассказах начала XX века выше, чем у рассказов, скажем, начала или середины XIX века. Это может оказаться плодотворным направлением дальнейшего исследования.

Литература

1. *Дейк Т. А. ван* (1989). Язык. Познание. Коммуникация. М.
2. *Мартыненко Г. Я., Шерстинова Т. Ю., Попова Т. И., Мельник А. Г., Замирайлова Е. В.* (2018). О принципах создания создания корпуса русского рассказа первой трети XX века // Труды XV Международной конференции по компьютерной и когнитивной лингвистике «TEL 2018». Казань. С. 180–197.
3. *Тодоров Ц.* (1978). Грамматика повествовательного текста. Новое в зарубежной лингвистике. Вып. 8. М. С. 450–463.
4. *Эрлих В.* (1996). Русский формализм. СПб.
5. *Cortazzi M.* (2002). Narrative Analysis. London.
6. *Labov W., Waletzky J.* (1967). Narrative analysis: oral versions of personal experience. J. Helm (ed.). Essays on the Verbal and Visual Arts: Proc. of the 1966 Annual Spring Meeting of the American Ethnological Society. Seattle; London. P. 12–44.
7. *Lohafer S.* (1994). A cognitive approach to storyness. Ch. May (ed.). The New Short Story Theories. Athens, Ohio. P. 301–311.

References

1. *Dijk T. A.* (1989). Jazyk. Poznanie. Kommunikacija [Language. Cognition. Communication]. Moscow.
2. *Martynenko G. Ja., Sherstinova T. Ju., Popova T. I., Mel'nik A. G., Zamirajlova E. V.* (2018). O principah sozdanija sozdanija korpusa russkogo rasskaza pervoj treti XX veka [On principles of creating a corpus of the Russian short stories of 1900–1930].

Trudy XV Mezhdunarodnoj konferencii po komp'juternoj i kognitivnoj lingvistike «TEL 2018» [Proceedings of XV International conference on computational and cognitive linguistics «TEL 2018»]. Kazan'. P. 180–197.
3. *Todorov Ts.* (1978). Grammatika povestvovatel'nogo teksta [Narrative Text Grammar]. Novoje v zarubezhnoj lingvistike. Vyp. 8. [New Trends in Foreign Linguistics, 8]. Moscow. P. 450–463.
4. *Erlikh V.* (1996). Russkij formalism [Russian Formalism]. St. Petersburg.
5. *Cortazzi M.* (2002). Narrative Analysis. London.
6. *Labov W., Waletzky J.* (1967). Narrative analysis: oral versions of personal experience. J. Helm (ed.). Essays on the Verbal and Visual Arts: Proc. of the 1966 Annual Spring Meeting of the American Ethnological Society. Seattle; London. P. 12–44.
7. *Lohafer S.* (1994). A cognitive approach to storyness. Ch. May (ed.). The New Short Story Theories. Athens, Ohio. P. 301–311.

Скребцова Татьяна Георгиевна
Санкт-Петербургский государственный университет (Россия)
Skrebtsova Tatiana
St. Petersburg State University (Russia)
E-mail: t.skrebtsova@spbu.ru

О. Е. Фролова
O. E. Frolova

КОРПУС КАК ИНСТРУМЕНТ ИССЛЕДОВАНИЯ ХУДОЖЕСТВЕННОГО ТЕКСТА

CORPUS AS A TOOL FOR RESEARCH OF FICTIONAL TEXT

Аннотация. В статье рассматриваются возможности использования корпуса для исследования текстовых категорий художественного текста, в частности, категории пространства. Спациальность анализируется на материале повести Андрей Платонова «Котлован». Конструирование пространства художественного текста подчиняется нескольким закономерностям: а) отношению локусов к сильным позициям текста, б) роли локуса в сюжете, в) перцептивности локусов, г) воспринимаемому субъекту. Обращение к корпусу позволяет выявить, как организуют пространства частотность, тематическая и семантическая близость лексических единиц. Лексика с пространственной семантикой формирует семантическое поле пространства. Употребление глаголов восприятия выявляет перцептивные локусы. Анализ контекстного окружения «спациальных» единиц показывает, какое пространство отнесено к планам прошлого, настоящего и будущего.
Ключевые слова. корпус, художественный текст, текстовые категории, частотность.

Abstract. The article discusses the possibility of using the corpus for the study of textual categories of fictional text, in particular, the category of space. Spatiality is analyzed on the material of Andrei Platonov's "The Pit". The construction of the space of an fictional text is subject to several laws: a) the relation of loci to strong positions of the text, b) the role of the locus in the plot, c) the perceptiveness of the loci, d) to the perceiving subject. Appeal to the corpus allows you to identify how space is organized by frequency, thematic and semantic proximity of lexical units. Vocabulary with spatial semantics forms a semantic field of space. The use of perceptual verbs reveals perceptual loci. An analysis of the contextual environment of "spacial" units shows what space is assigned to the plans of the past, present, and future.
Keywords. corpus, fictional text, text categories, frequency.

Категории пространства повествовательного художественного текста посвящен ряд главным образом литературоведческих работ [Бахтин 1975; Лотман 1988]. Нелинейность организации пространства подчиняется сильным позициям текста: заголовку, начальной и финальным фразам текста [Арнольд 1978], а также включением в перцептивно оформленные фрагменты текста.

Статус локуса в тексте определяется: 1) характером номинации: апеллятив или оним, 2) принадлежностью природе, деревне или городу, 3) ролью в сюжете — локализацией события и персонажа, 4) наличием или отсутствием модусного, субъективного, компонента, благодаря которому локус становится объектом восприятия, мечты,

фантазии; 5) характером локуса — реальным или ирреальным; 6) временной характеристикой локуса — отнесением к плану прошедшего, настоящего или будущего времени.

Мы рассматриваем только один, достаточно сложный по стилю текст без сопоставления с другими, исходя из того, выделенные выше позии имеют универсальный характер и приложимы к описанию пространственного устройства прозаического художественного текста как класса текстов.

Мы намереваемся выяснить с помощью Национального корпуса русского языка, как «построено» пространство повести А. Платонова «Котлован», проанализированное в работе [Левин 1998]. Ю. И. Левин рассматривает обстоятельства места при глаголах, сочетаемость глаголов с пространственными предлогами и отмечает важную роль в тексте модусных предикатов [Левин 1998: 414; 416, 417].

Корпус открытвает новые возможности исследования, поскольку позволяет определить частотность лексической единицы и грамматической конструкции в тексте [Захаров, Богданова 2011]. Мы видим свою задачу в том, чтобы соотнести частотность единицы или конструкции с их семантикой и текстовой позицией.

В качестве единиц поиска были выбраны несколько типов единиц: 1) слова соотносимые с сильными позициями: заглавие, первая фраза, финальный абзац текста; 2) большие локусы текста; 3) общественные локусы, связанные с совместным трудом или являющиеся местом собраний; 4) малые локусы, являющиеся жильем; 5) метаслова, описывающие пространственную организацию чего-либо; 6) глаголы зрительного восприятия; 7) глаголы интеллектуальной деятельности; 8) глагол *быть* в будущем времени; 9) способы характеризации локуса в атрибутивных словосочетаниях. Поиск велся по подкорпусу НКРЯ, равному тексту «Котлована» и составившему 2495 предложений, 34867 слов.

Частотность, заданная словом, вынесенным в заглавие, приобретает большую важность и по отношению к названиям других локусов. Существительное *котлован* встречается 44 раза. Локусы, заданные в инициальной (пример 1) и финальной фразах (пример 2): *завод* (общее количество 13), *камень* (общее количество 13), *барак* (общее количество 30). Заглавие и начало текста определяют характер пространства, которое соотносимо со строительством и производством. Существительное *барак*, называющее жилище, является локусом, «закрывающим» пространство текста. Локусом, помещенным в сильную

позицию, мы также считаем *кафельный завод*, где Чиклин находит умирающую Юлию и ее дочь Настю.

(1) «В день тридцатилетия личной жизни Вощеву дали расчет с небольшого механического **завода**, где он добывал средства для своего существования». [НКРЯ].

(2) «Отдохнув, Чиклин взял Настю на руки и бережно понес ее класть в **камень** и закапывать. Время было ночное, весь колхоз спал в **бараке**, и только молотобоец, почуяв движение, проснулся, и Чиклин дал ему прикоснуться к Насте на прощанье». [НКРЯ].

Большие локусы текста, между которыми развертывается сюжет: *город* (26 случаев) и *деревня* (30 случаев). Количество упоминаний данных единиц в тексте соотносимо.

Частотность существительных, называющих общественные локусы, в частности, заданные в инициальной фразе, такова: *завод* (13 случаев), *кафельный завод* (7), *организационный двор* (5 случаев) / *оргдвор* (25 случаев). К этой же группе относится существительное *колхоз*, у которого из 99 случаев употребления 11 интерпретируются как пространство. Именно в данной группе обнаруживается еще один локус в сильной позиции текста, поскольку он выражен топонимом: *колхоз имени Генеральной Линии.*

Локусы жилья выражены несколькими существительными *дом* (36), *изба* (18), *барак* (30). НКРЯ также обнаруживает 5 случаев употребления сложной единицы *изба-читальня,* утратившей функцию служить жильем для человека.

Соотношение метаслов *пространство* (12 случаев) и *место* (66 случаев) показывает, что философский дискурс в тексте уступает бытовому. Пространство достаточно скупо охарактеризовано в атрибутивных словосочетаниях: *общее*, *трудное*. *Место* в тех же конструкциях показывает достаточно большое разнообразие: *далекое*, *лишнее*, *маточное*, *мягкое*, *новое*, *открытое*, *порожнее*, *просторное*, *пустопорожнее*, *родное*, *светлое*, *скучное*, *сырое*, *тесное*, *тихое*, *теплое*, *узкое*, *чужое*.

Поскольку пространство является объектом восприятия и интеллектуальной деятельности, мы проверили по корпусу частотность модусных предикатов, выраженных глаголами *видеть* (63 случая), *увидеть* (26 случаев), *смотреть* (20 случаев), *посмотреть* (11 случаев), *наблюдать* (21 случай), а также *знать* (74 случая), *понять* (7 случаев), *понимать* (9 случаев). Сложность анализа модуса в тексте с по-

мощью корпуса заключается в том, что корпус позволяет обнаружить лишь эксплицированное присутствие модусного предиката в рамке *X увидел, как / что P.* Корпус бессилен в выявлении имплицированных модусных рамок. Поиск употреблений с частицей *не* показывает, что в перцептивной группе у глагола *видеть* соотношение 63 общих употреблений и 15 с отрицанием, у *увидеть* 26 и 1, корпус не обнаруживает употреблений с отрицательной частицей у глаголов *смотреть*, *посмотреть* с интервалом 1 слово. С глаголами интеллектуальной деятельности ситуация такая: у глагола *понять* соотношение общего количества и отрицательных употреблений 7 и 1, у *понимать* 9 и 5. Глагол *знать* демонстрирует такое распределение: из 74 употребления и 37 отрицательных. Соотношение употреблений глаголов *видеть*, *увидеть, смотреть, посмотреть* позволяет выявить значительное превалирование глаголов восприятия над конструкциями без отрицания с глаголами интеллектуальной деятельности. Объектом восприятия становится и освоенное, и неосвоенное пространство (примеры 3, 4).

(3) «Вощев долго **наблюдал** строительство неизвестной ему башни; он **видел**, что рабочие шевелились равномерно, без резкой силы, но что-то уже прибыло в постройке для ее завершения». [НКРЯ].

(4) «Он жил так в недавнее время, … и сколько годов он ни **смотрел** из деревни вдаль и в будущее, он **видел** на конце равнины лишь слияние неба с землею, а над собою имел достаточный свет солнца и звезд». [НКРЯ].

Поиск по формам будущего времени глагола *быть* (98 случаев), а также по частотности употребления прилагательного *будущий* (44 случая) позволяет выявить интересную особенность хронотопа в тексте (примеры 5, 6) **—** пренебрежение прошлым и настоящим во имя будущего.

(5) «Здесь **будет дом**, в нем **будут** храниться люди от невзгоды и бросать крошки из окон живущим снаружи птицам». [НКРЯ].

(6) «Пусть **будущее будет** чужим и **пустым**, а **прошлое** покоится в **могилах —** в тесноте некогда обнимавшихся костей, в прахе сотлевших любимых и забытых тел». [НКРЯ].

План прошедшего был проанализирован на показателях частотности наречий *раньше* (8 случаев) и *прежде* (6 случаев), которые представлены в тексте достаточно скупо.

Подведем итог нашим наблюдениям. Частотность единиц, конструирующих пространство текста «Котлована», отражена в таблице 1.

Таблица 1. **Частотность единиц, конструирующих пространство текста**

Группа единиц	единица	Частотность
Сильные позиции текста		
Заглавие	*котлован*	44
Начало текста	*завод*	1
Событие	*кафельный завод*	1
Финал текста	*камень*	1 (общее количество 15)
	барак	1
Топоним	*Колхоз имени Генеральной Линии*	1
Большие локусы текста		
	город	26
	деревня	30
Общественные локусы		
	завод	13
	кафельный завод	7
	Организационный двор / оргдвор	30
	колхоз	11
Локусы жилья		
	дом	39
	барак	30
	изба	18
Метаслова		
	место	66
	пространство	12
Модусные средства текста		
Модусные глаголы интеллектуальной деятельности		
	знать	74

Окончание табл. 1

Группа единиц	единица	Частотность
	понять	7
	понимать	9
Глаголы зрительного восприятия		
	видеть	63
	увидеть	26
	смотреть	20
	посмотреть	11
	наблюдать	21
Хронотоп.		
План будущего		
	быть	98
	будущий	44
План прошедшего		
	прежде	6
	раньше	8

Итак, анализ частотности показывает, что пространство текста «Котлована» оказывается сложно организованным: по преимуществу безымянным, разделенным между городом и деревней, устремленным в будущее. Локус *котлован*, являющийся заглавием, сильной позицией текста, обнаруживает большую частотность употребления и также непосредственно связан с будущим, поскольку данное слово называет основание еще не построенного здания. Модус зрительного восприятия представлен шире, чем модус знания, а значительное количество употреблений глагола *знать* с отрицанием для субъектов модуса делает будущее неопределенным.

Литература

1. *Арнольд И. А.* (1978) Значение сильной позиции для интерпретации художественного текста // Иностранные языки в школе. № 4. С. 23–31.
2. *Бахтин М. М.* (1975) Формы времени и хронотопа в романе // Бахтин М. М. Вопросы литературы и эстетики. Исследования разных лет. М.

3. *Захаров В. П., Богданова С. Ю.* (2011) Корпусная лингвистика. Иркутск.
4. *Левин Ю. И.* (1998) От синтаксиса к смыслу и далее («Котлован» А. Платонова) // Левин Ю. И. Избр. труды: Поэтика. Семиотика. М. С. 292–419.
5. *Лотман Ю. М.* (1988) Художественное пространство в прозе Гоголя // Лотман Ю. М. В школе поэтического слова: Пушкин. Лермонтов. Гоголь. М. С. 251–292.
6. Национальный корпус русского языка. Электронный ресурс. URL: http://ruscorpora.ru/search-main.html

References

1. Arnol'd I. A. (1978) Znachenie sil'noj pozicii dlya interpretacii hudozhestvennogo teksta [The value of a strong position for the interpretation of fictional text]. In: Inostrannye yazyki v shkole [Foreign languages at school]. № 4. P. 23–31.
2. Bahtin M. M. (1975) Formy vremeni i hronotopa v romane [Forms of time and chronotope in the novel]. In; Voprosy literatury i ehstetiki. Issledovaniya raznyh let [Questions of literature and aesthetics. Studies of different years]. Moscow.
3. Zaharov V. P., Bogdanova S. Yu. (2011) Korpusnaya lingvistika [Corpus linguistics]. Irkutsk.
4. Levin Yu. I. (1998) Ot sintaksisa k smyslu i dalee («Kotlovan» A. Platonova) [From syntax to sense and further (A. Platonov's "Pit")]. In: Izbr. trudy: Poehtika. Semiotika [Selected. works: Poetics. Semiotics]. Moscow. P. 292–419.
5. Lotman Yu. M. (1988) Hudozhestvennoe prostranstvo v proze Gogolya [Space in Gogol's prose]. In: V shkole poehticheskogo slova: Pushkin. Lermontov. Gogol' [In the school of poetic words: Pushkin. Lermontov. Gogol]. Moscow. P. 251–292.
6. Natsional'nyy korpus russkogo yazyka. [Russian National Corpus]. URL: http://ruscorpora.ru/search-main.html

---

Фролова Ольга Евгеньевна
Московский государственный университет им. М. В. Ломоносова (Россия)
Olga Frolova
Moscow State University (Russia)
E-mail: olga_frolova@list.ru

Т. Ю. Шерстинова
T. Yu. Sherstinova

БИОГРАФИЧЕСКАЯ БАЗА ДАННЫХ РУССКИХ ПИСАТЕЛЕЙ (К СОЗДАНИЮ КОРПУСА РУССКОГО РАССКАЗА XX ВЕКА)[1]

BIOGRAPHIC DATABASE OF RUSSIAN WRITERS: TOWARDS CREATION OF THE RUSSIAN SHORT STORIES CORPUS OF THE 20TH CENTURY

Аннотация. Рассматриваемая в статье биографическая база данных русских писателей создается как модуль Корпуса русского рассказа первой трети XX века. При построении формальной модели литературно-художественной системы исследуемой эпохи, а также при изучении отдельных авторских стилей и языка конкретных писателей представляется целесообразным учитывать определенные аспекты биографии и социологические характеристики авторов. Разрабатываемая биографическая база данных позволит проводить исследования языка и стиля художественных произведений в зависимости от целого ряда параметров – социального происхождения писателей, их образования, вида деятельности, возраста автора на момент написания конкретного произведения и др.
Ключевые слова. Корпусная лингвистика, русская литература, русский язык, русский рассказ, биографическая база данных.

Abstract. The article discusses the structure of the biographical database of Russian writers, created for the Russian short stories corpus of the first third of the 20th century. For building a formal model of a literary system, as well as for studying individual author's styles and artistic trends, it seems appropriate to take into account certain sociological characteristics and biography features of the writers. The anticipated biographical database will allow to conduct language and style studies of literary works depending on a number of parameters – writer's social origin, education, type of activity/profession, author's age at the time of writing a particular work, etc.
Keywords. Corpus linguistics, Russian literature, Russian language, Russian short story, biographical database, sociolinguistics.

1. Введение

В последние годы в рамках общей тенденции к дигитализации гуманитарного знания у ученых-словесников наблюдается рост интереса к исследованию литературных произведений квантитативными методами — появляются новые интересные ресурсы (напр., корпус про-

[1] Работа выполнена при поддержке РФФИ, грант № 17-29-09173 офи_м «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

изведений Чарльза Диккенса и английской литературы XIX века CLiC Dickens [CLiC] или корпус русских драматических текстов RusDraCor [Скоринкин 2018]) и проводятся многоаспектные корпусные исследования художественных произведений [Craig & Kinney 2009; Fischer-Starcke 2010; Balossi 2014 и др.].

Количество корпусов, посвященных русской литературе, а, соответственно и проводимых на их материале исследований, до сих пор существенно уступает множеству и разнообразию ресурсов, реализуемых для англоязычной литературы. Кроме того, следует отметить, что большинство цифровых литературных проектов — как для русского, так и других языков — обычно имеют целью изучение творчества только одного выдающего автора. Так, для английской литературы это в первую очередь — Шекспир, Диккенс, Остин, Вулф; для русской — Пушкин, Лермонтов, Достоевский, Чехов, Толстой.

Рассматриваемая в данной работе биографическая база данных русских писателей создается как отдельный модуль Корпуса русского рассказа первой трети XX в. [Мартыненко и др. 2018а]. Отличительной особенностью этого цифрового ресурса является стремление его разработчиков включить в корпус тексты по возможности максимального количества авторов-прозаиков, публиковавшихся в рассматриваемый период. Тем самым станет возможным анализировать художественную прозу изучаемой эпохи как единую «литературно-художественную систему» (этот термин был предложен выдающимся отечественным литературоведом и писателем Ю. Н. Тыняновым, который говорил о необходимости такого подхода еще 90 лет назад [Тынянов 1929]), а также моделировать изменения языка и стиля русской литературы в синхронии и диахронии на представительном языковом материале [Мартыненко и др. 2018б].

Именно разнообразие представленных авторов-писателей при сохранении единства языка и жанра можно считать главным преимуществом создаваемого корпуса, который позволит изучать динамику языка русской литературы этого насыщенного событиями и воистину драматического периода в истории нашей страны методами математической и компьютерной лингвистики.

При построении формальной модели литературно-художественной системы, а также при изучении отдельных авторских стилей и художественных направлений представляется целесообразным учитывать определенные социологические характеристики авторов, а также некоторые аспекты их биографий. Разрабатываемая биографическая

база данных позволит проводить исследования языка и стиля художественных произведений в зависимости от целого ряда параметров — социального происхождения авторов, их образования, вида деятельности, возраста писателя на момент написания конкретного произведения и др.

2. Об отборе авторов для включения в Корпус

Как уже было отмечено, Корпус русского рассказа имеет заданные временные рамки — в него включаются произведения, написанные или впервые опубликованные с 1900 по 1930 г. включительно. Формирование полного списка прозаиков изучаемой эпохи основывается на библиографиях и литературных энциклопедиях, словарях писателей, каталогах крупнейших библиотек, периодических изданиях того времени, электронных онлайн библиотеках и др. интернет-ресурсах [Мартыненко и др. 2018б].

Ставится задача включения в корпус рассказов не только «столичных», но и региональных авторов, писавших на русском языке и проживавших на территории Российской империи (до 1917), а позже — на территории РСФСР и СССР. Полагается, что для включения произведения в корпус достаточно оснований, даже если это единственный рассказ, написанный литератором. Поэтому в корпусе представлены рассказы не только прозаиков, но и поэтов, драматургов, публицистов. Исключением является проза для детей и юношества — писатели, работающие в «детском» жанре в корпус не включаются, равно как и произведения писателей, написанные в эмиграции.

3. Источники и структура биографической базы

Для каждого автора, включаемого в корпус, производится поиск биографической информации. Основными источниками здесь являются:

Русский биографический словарь — электронная версия Энциклопедического Словаря Брокгауза и Ефрона (1890–1907 гг.) и Нового Энциклопедического Словаря (1910–1916 гг.) [РБС].

Литературные энциклопедии [ЛЭ; КЛЭ].

Библиографические указатели [Муратова 1963], словари писателей, биографические словари, литературные сайты, библиотеки, словари псевдонимов и др.

В настоящее время биографическая база данных русских прозаиков состоит из двух модулей, представленных в виде реляционных таблиц: 1) полный список писателей и 2) биографические данные.

1) **Список писателей** состоит из следующих основных полей описания:

№	Поле описания
1	Номер или код писателя в базе данных
2	ФИО писателя
3	Комментарий
4	Количество электронных текстов рассказов автора в собранной коллекции
5	Класс/категория, отражающая степень готовности текстов
6	Наличие иных электронных ресурсов для данного автора (URL-ссылки)

2) **Биографические данные** содержат следующие поля:

№	Поле описания
1	Номер или код писателя в базе данных
2	ФИО писателя
3	Основной литературный псевдоним
4	Другие имена, под которыми печатался автор
5	Год рождения
6	Год смерти
7	Дата рождения
8	Дата смерти
9	Место рождения
10	Место смерти
11	Эмиграция (если да, то с какого года)
12	Социальное происхождение
13	Образование
14	Профессии
15	Где издавался

16	Наиболее известные произведения/сборники
17	Основные жанры
18	Художественное направление
19	Комментарии о творчестве
20	Основные места проживания
21	Особенности биографии
22	Первая публикация
23	Последняя публикация
24	Интересные факты о жизни/творчестве
25	Интернет-ссылки (биографические данные)
26	ЭСБЕ/НЭС [Русский биографический словарь]
27	ЛЭ-1929 [Литературная энциклопедия 1929]
28	КЛЭ-1962 [Краткая литературная энциклопедия 1962]
29	ИРЛ-1963 [Муратова 1963]
30	Другие ссылки (библиогр. указатели, словари писателей)
31	Прочие комментарии (что не вошло в прочие рубрики)

Тестовое заполнение биографической базы данных было осуществлено студентами образовательной программы «Филология» НИУ ВШЭ, Санкт-Петербург. Ставилась задача сбора биографической информации с целью накопить материал и оценить перспективы его оптимальной нормализации, поэтому жесткой стандартизации требований для заполнения полей не вводилось.

В результате, из первоначального списка в 300 имен прозаиков удалось найти биографические данные для 265 персоналий. При этом степень заполнения полей описания для разных авторов достаточно сильно варьируется. Так, для известных и хорошо изученных прозаиков биографическая информация представлена с избытком. Информация о писателях «второго эшелона» как правило более скудна. Еще хуже дело обстоит с малоизвестными и забытыми писателями. Так, для 35 прозаиков из 300 к настоящему моменту не удалось найти никакой биографической информации (для некоторых авторов, публиковавшихся под псевдонимами, нет информации даже об их настоящих именах). Можно предположить, что при расширении выборки авто-

ров в корпусе доля авторов «с белыми пятнами в биографии» будет еще выше, так как в первоначальный список были включены преимущественно маститые и популярные русские писатели.

4. Распределение русских писателей по продолжительности жизни

В качестве примера «автономного» использования биографической базы данных рассмотрим не требующую нормализации переменную — продолжительность жизни прозаиков рассматриваемой эпохи. В табл. 1 приведены основные статистики по этому параметру, полученные для выборки в 227 персоналий. Гистограмма полученного распределения, которое оказалось близко к нормальному, представлена на рис. 1.

Таблица 1. **Продолжительность жизни русских прозаиков**

N	Средн.зн.	Медиана	Мин.	Макс.	Ст. откл.	Коэф.вар.
227	59,56	60	23	95	14,82	24,88

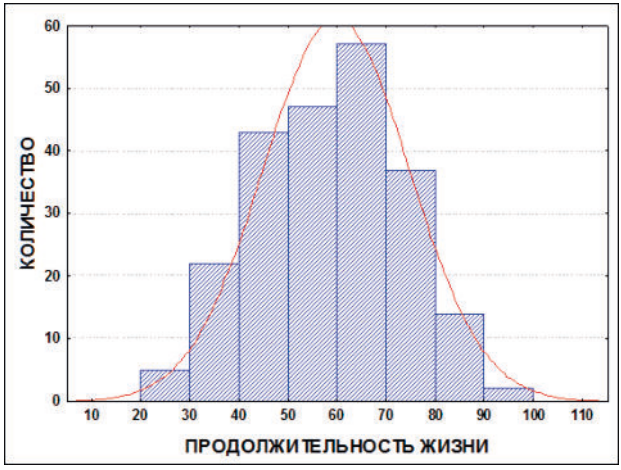


Рис. 1. **Гистограмма продолжительности жизни русских прозаиков первой трети XX в.**

Ограниченный объем данной статьи не позволяет привести исходную таблицу целиком. Однако, по-видимому, стоит отметить писателей, занимающих «предельные» позиции в этом списке. Это Лев Ната-

нович Лунц, прозаик, публицист и драматург, участник объединения «Серапионовы братья», который прожил всего 23 года, и Сергей Иванович Гусев-Оренбургский, один из популярных писателей дореволюционной России, который оказался долгожителем среди писателей, прожив 95 лет (возможно, благодаря своей эмиграции из Советской России в 1921 г.).

В заключение отметим, что следующей важной задачей создания биографической базы данных русских писателей ставится нормализация информации по представленным полям описания, что позволит проводить автоматический поиск и фильтрацию данных по соответствующим параметрам. Совмещение биографической информации с корпусной разметкой художественных произведений даст возможность посмотреть на литературные тексты, особенности языка и стиля под новым углом зрения — с точки зрения отдельных социологических признаков авторов и особенностей их жизненного пути. Предполагается, что такой междисциплинарный подход может быть весьма перспективным.

Литература

1. *КЛЭ — Краткая литературная энциклопедия в* 9 т. М.: Советская энциклопедия, 1962–1978, URL: http://feb-web.ru/feb/kle/kle-abc/default.asp (дата обращения: 04.05.2019).
2. *ЛЭ — Литературная энциклопедия*: В 11 т. — [М.], 1929—1939, URL: http://feb-web.ru/feb/litenc/encyclop/ (дата обращения: 04.05.2019).
3. *Мартыненко Г.Я., Шерстинова Т.Ю., Мельник А.Г., Попова Т.И.* (2018а), Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) / Компьютерная лингвистика и вычислительные онтологии. Вып. 2 (Труды XXI Межд. объед. конф. «Интернет и современное общество», IMS-2018), СПб: Университет ИТМО, 2018. С. 99–104.
4. *Мартыненко Г.Я., Шерстинова Т.Ю., Попова Т.И., Мельник А.Г., Замирайлова Е.В.* (2018б), О принципах создания корпуса русского рассказа первой трети XX века // Труды XV Межд. конф. по компьютерной и когнитивной лингвистике «TEL 2018». Казань, с. 180–197.
5. *Муратова К.Д.* (ред.) История русской литературы конца XIX — начала XX века, Библиографический указатель. М.: АнСССР, 1963.
6. *РБС — Русский биографический словарь*, URL: http://rulex.ru/be.htm (дата обращения: 04.05.2019).
7. *Русские писатели 1800–1917*, Биографический словарь. В 7 т. Под ред. Николаева П. А. 1992–2000. М.: «БРЭ».

8. *Тынянов Ю. Н.* (1929), Архаисты и новаторы. Л.: Прибой.
9. *Balossi G. A.* (2014) Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's 'The Waves'. Amsterdam; Philadelphia: John Benjamins Publishing Company.
10. *CliC Dickens*. URL: http://clic.bham.ac.uk/ (дата обращения: 05.05.2019).
11. *Craig, H. and Kinney, A. F.* (2009) Shakespeare, Computers, and the Mystery of Authorship. Cambridge: Cambridge University Press.
12. *Fischer-Starcke B.* (2010) Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries. London; N-Y: Continuum.
13. *Skorinkin D., Fischer F., Palchikov G.* (2018) Building a Corpus for the Quantitative Research of Russian Drama: Composition, Structure, Case Studies. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii), Vol. 2018 (17), pp. 662–682.

References

1. *Balossi G. A.* (2014) Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's 'The Waves'. Amsterdam; Philadelphia: John Benjamins Publishing Company.
2. *CliC Dickens*. URL: http://clic.bham.ac.uk/ (last accessed: 05.05.2019).
3. *Craig, H. and Kinney, A. F.* (2009) Shakespeare, Computers, and the Mystery of Authorship. Cambridge: Cambridge University Press.
4. *Fischer-Starcke B.* (2010) Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries. London; N-Y: Continuum.
5. *Kratkaya literaturnaya enciklopediya* [A brief literary encyclopedia] in 9 vol. M.: Sovetskaya enciklopediya [Soviet Encyclopedia], 1962–1978, URL: http://feb-web.ru/feb/kle/kle-abc/default.asp (last accessed: 04.05.2019).
6. *Literaturnaya enciklopediya* [The literary encyclopedia]: 11 vol. [M.], 1929–1939, URL: http://feb-web.ru/feb/litenc/encyclop/ (last accessed: 04.05.2019).
7. *Martynenko G. Y., Melnik A. G., Popova T. I., Sherstinova T. Y.* (2018a), Methodological problems of creating the Computer Anthology of Russian short stories as a language resource designed to study language and style of Russian prose in the era of revolutionary changes (in the first third of the 20th century). In: Komp'yuternaya lingvistika i vychislitel'nye ontologii. Vyp. 2 (Trudy XXI Mezhd. ob"ed. konf. «Internet i sovremennoe obshchestvo», IMS-2018), SPb: ITMO, pp. 99–104.
8. *Martynenko G. Y., Melnik A. G., Popova T. I., Sherstinova T. Y., Zamiraylova E. V.* (2018), On the Principles of Creation of the Russian Short Stories Corpus of the First Third of the 20th Century. In: Proc. of the Int. Conference «TEL 2018», Kazan, pp. 180–197.
9. *Muratova K. D.* (*ed.*) (1963) Istoriya russkoj literatury konca XIX–nachala XX veka, Bibliograficheskij ukazatel' [The history of Russian literature of the late XIX — early XX century, Bibliography], Moscow.
10. *Nikolaev P. A.* (*ed.*) (1992–2000) Russkie pisateli 1800–1917, Biograficheskij slovar'. 7 vol. M.: «BRE».
11. *Russkij biograficheskij slovar',* URL: http://rulex.ru/be.htm (last accessed: 04.05.2019).
12. *Skorinkin D., Fischer F., Palchikov G.* (2018) Building a Corpus for the Quantitative

Research of Russian Drama: Composition, Structure, Case Studies. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii), Vol. 2018 (17), pp. 662–682.
13. *Tynyanov Yu. N.* (1929) Arkhaisty i novatory [Archaists and innovators]. Priboj, Leningrad.

Шерстинова Татьяна Юрьевна
Национальный исследовательский университет
«Высшая школа экономики» (Санкт-Петербург, Россия)
Санкт-Петербургский государственный университет (Россия)
Tatiana Sherstinova
St. Petersburg State University (Russia)
National Research University Higher School of Economics
(St. Petersburg, Russia)
E-mail: sherstinova@gmail.com

Научное издание

ТРУДЫ МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2019»
24–28 июня 2019 г., Санкт-Петербург

Компьютерная верстка *Ю. Ю. Тауриной*

Подписано в печать 19.06.2019. Формат 60×84 $^{1}/_{16}$.
Усл. печ. л. 26,04. Тираж 98 экз. Заказ №

Типография Издательства СПбГУ.
199034, Санкт-Петербург, Менделеевская линия, д. 5.