ST PETERSBURG STATE UNIVERSITY

PROCEEDINGS

OF THE INTERNATIONAL CONFERENCE

«CORPUS LINGUISTICS–2021»

July 1–3, 2021, St. Petersburg

СКИФИЯ
принт

St. Petersburg
2021

Организационный комитет конференции
«Корпусная лингвистика–2021»

В. П. Захаров (председатель), И. В. Азарова,
Е. Л. Алексеева, Л. Н. Беляева, А. О. Гребенников,
О. Н. Камшилова, О. В. Митренина, О. А. Митрофанова,
И. С. Николаев (зам. председателя), В. И. Фирсанова,
М. В. Хохлова, А. В. Чижик

Программный комитет конференции
«Корпусная лингвистика–2021»

В. П. Захаров (председатель), Н. Абдурахмонова (Узбекистан), И. В. Азарова,
Е. Л. Алексеева, В. А. Баранов, Л. Н. Беляева, В. Бенко (Словакия),
О. В. Блинова, С. Ю. Богданова, Н. В. Борисов, В. В. Бочаров,
Р. фон Вальденфельс (Германия), А. М. Галиева, Р. Гарабик (Словакия),
А. Горак (Чехия), А. О. Гребенников, Д. О. Добровольский,
Т. Елинек (Чехия), Л. Л. Иомдин, Е. Каллас (Эстония), О. Н. Камшилова,
К. И. Коваленко, М. С. Коган, А. В. Колмогорова, Е. Н. Колпачкова,
М. В. Копотев (Финляндия), М. Кршен (Чехия), Д. А. Кочаров,
А. М. Лаврентьев (Франция), У. Лоу (Великобритания), О. Н. Ляшевская,
О. В. Митренина, О. А. Митрофанова, М. М. Михайлов (Финляндия),
А. Д. Москвина, О. А. Невзорова, И. С. Николаев (зам. председателя),
В. Нозеда (Италия), Х. Нэси (Великобритания), К. Пала (Чехия),
В. Петкевич (Чехия) (зам. председателя), А. Ч. Пиперски,
Л. В. Рычкова (Беларусь), С. О. Савчук, В. П. Селегей, Д. В. Сичинава,
О. Скривнер (США), В. Д. Соловьев, А. Стефанович (Германия),
Е. В. Суворина, Ю. Тао (Китай), Н. Тюренн (Франция), М. В. Хохлова,
А. Я. Шайкевич, С. А. Шаров (Великобритания), Т. Ю. Шерстинова,
С. Эйден (Франция), Е. В. Ягунова, М. Якубичек (Чехия)

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2021»

1–3 июля 2021 г., Санкт-Петербург

СКИФИЯ
принт
Санкт-Петербург
2021

ББК 81.1
Т78

Ответственный редактор издания
В. П. Захаров

Т78 **Труды международной конференции «Корпусная лингвистика–2021».** — СПб.: Издательство Скифия-принт, 2021. — 396 с.

ISSN 2412-9623
ISBN 978-5-98620-557-1

Сборник содержит материалы докладов, представленных на научной конференции «Корпусная лингвистика-2021» 1–3 июля 2021 г. в Санкт-Петербурге.

Создание и использование 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

ISSN 2412-9623
ISBN 978-5-98620-557-1

ОГЛАВЛЕНИЕ

ПЛЕНАРНЫЕ ДОКЛАДЫ

KEYNOTE TALKS

D. Zeman

ENHANCED UNIVERSAL DEPENDENCIES: THE CURRENT STATE AND OUTLOOK[1]

Abstract. Universal Dependencies (UD) is a multilingual collection of corpora featuring morphological and syntactic annotation in a unified style. We discuss an optional layer of deep-syntactic annotation in UD, called Enhanced Universal Dependencies. We survey the existing enhanced representation as of release 2.8 and consider two possible future expansions: semi-automatic addition of existing enhancement types to new languages, and addition of new enhancement types.
Keywords. Dependency syntax, deep syntax, multilingual corpora, gapping, coordination, coreference.

1. Introduction

Universal Dependencies[2] (UD) [Nivre et al. 2020; de Marneffe et al. 2021] is an international community project that strives to define a unified morpho-syntactic annotation scheme applicable to all natural languages, and to collect corpora (treebanks) annotated following that scheme. It started with the first version of annotation guidelines in 2014, and with the first release of 10 treebanks in January 2015; after six years, release 2.8 of UD boasts about 202 treebanks for 114 languages from 24 different families. Some treebanks are just tiny samples of less than thousand tokens while others contain over a million tokens; the total size of the collection amounts to 27 million.

UD has become an indispensable resource for research on multi-lingual natural language processing, especially morphological tagging and syntactic parsing. It has been also used in many linguistic studies, in particular in linguistic typology. Two large CoNLL shared tasks were organized in 2017 and 2018 to evaluate parsing systems on UD data [Zeman et al. 2018].

[1] This work was supported by the Grant No. GX20-16819X (LUSyD) of the Czech Science Foundation (GAČR).

[2] https://universaldependencies.org/

The morphological annotation in UD includes the lemma, universal part-of-speech tag (UPOS), morphological feature-value pairs, and possibly another tag from a treebank-specific tag set (XPOS). The UPOS tag must be picked from a fixed set of 17 categories; any finer distinctions, if desirable, are encoded in the morphological features (in addition to universally defined features, treebanks may also use language-specific values and features). All UD treebanks must have at least the UPOS tags manually checked. Lemmas, features and XPOS are optional and in a few UD corpora they have been assigned automatically.

The *basic* syntactic representation is a rooted dependency tree where every word/node (except the root) has one parent node. Each relation in the tree is labeled with its type; while the main types come from a fixed set of 37 universal relations, it is possible to define language-specific subtypes. Like UPOS tags, the parent nodes and the relation types are manually checked in all UD treebanks.

Examples of UD basic trees with UPOS tags are given in Figures 1 (English) and 2 (Russian), respectively. The sentences are parallel and so is their syntactic annotation: Relations between content words are identical in both structures, although English also has a number of additional relations be-

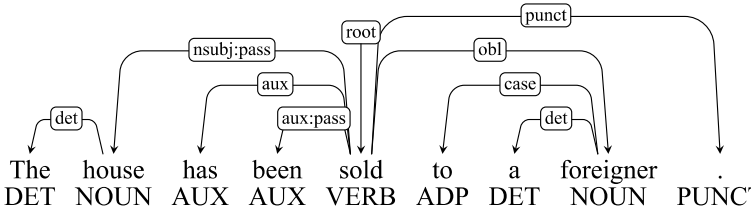


Fig. 1. **Basic UD tree of the English sentence *The house has been sold to a foreigner***

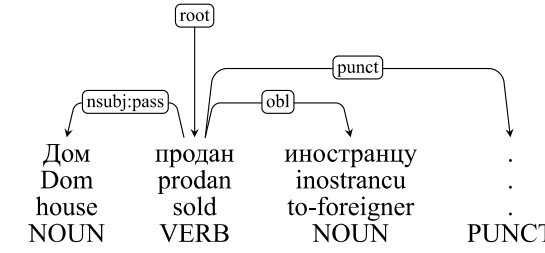


Fig. 2. **Basic UD tree of the Russian sentence *Дом продан иностранцу* (The house has been sold to a foreigner)**

tween content and function words, which are not present in Russian. Such parallelism is possible thanks to the fact that content words are attached higher in UD, and function words (articles, prepositions, auxiliaries etc.) are normally attached as leaf nodes — an approach that is relatively uncommon in dependency frameworks outside UD.

2. Enhanced Representation

Besides the basic syntactic representation, UD may optionally contain an *enhanced* dependency structure, which is still a directed and rooted graph, but not necessarily a tree. The enhanced UD layer was proposed by [Schuster and Manning 2016] but its first official specification appeared in the version 2 of the UD guidelines [Nivre et al. 2020: § 3.4]. The purpose of the enhanced representation is to facilitate downstream language understanding tasks by making certain relations explicitly annotated. In most cases, the enhanced graph is only a moderate modification of the basic tree. There are six types of enhancements defined in the guidelines (a UD treebank may annotate only some types and ignore the others):

1) gapping (empty nodes for elided predicates);
2) parent of coordination (propagated relations to non-first conjuncts);
3) shared dependent of coordination (propagated relations from non-first conjuncts);
4) external subject of a controlled or raised verb;
5) relative clause (modified noun attached instead of the relative pronoun, thus forming a directed cycle);
6) case information in the relation label.

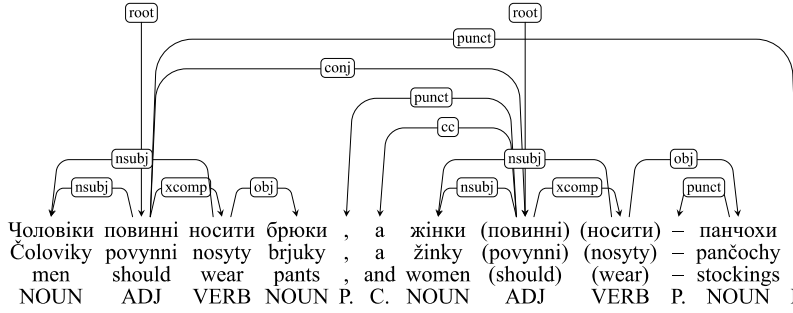


Fig. 3. **Enhanced UD graph of the Ukrainian sentence *Чоловіки повинні носити брюки, а жінки — панчохи* (Men should wear pants and women should wear stockings)**

Figure 3 illustrates three enhancement types. There are two 'empty' nodes that represent elided predicates in a gapping construction, *повинні* "should" and *носити* "wear". Propagation of coordination parent results in the second root relation, pointing to the empty node representing the second instance of *повинні*. And finally, both *чоловіки* "men" and *жінки* "women" have two incoming relations each, making them subject not only of *повинні* but also of *носити*. Figure 4 illustrates enhanced representation of relative clauses; note the directed cycle between *szamponu* "shampoo" and *myje* "washes". For more details on the six enhancement types, see [Droganova and Zeman 2019; Nivre et al. 2020].

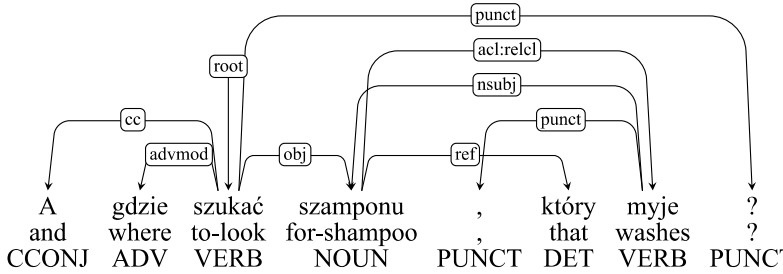


Fig. 4. **Enhanced UD graph of the Polish sentence *A gdzie szukać szamponu, który myje?* (And where to look for shampoo that works?)**

Some of the enhancements can be computed deterministically from the basic tree, as they really just foreground information that is already there. This is the case with coordination parents and case-enhanced relation labels. Other types will benefit from manual disambiguation but they still can be approximated algorithmically using relatively simple heuristics. At the other end of the scale, distinguishing between shared dependents of coordination and private dependents of the first conjunct clearly requires extra knowledge and may be sometimes hard even for a human annotator. Several tools have been proposed that use heuristics to compute some of the enhancements from the basic tree; see [Nivre et al. 2018] for a comparison. Among the UD treebanks that currently have some enhanced dependency structures, little annotation (if any at all)[3] has been added manually. In some cases, the treebanks were converted from non-UD annotation schemes where the extra information was available. In others, heuristic enhancers

[3] Exact information about the origin of enhanced annotation in individual treebanks is not available.

were employed. Table 1 shows statistics of enhancement types in UD 2.8, which is the most recent release at the time of writing. Overall, 30 treebanks of 18 languages have at least one enhancement type, and 15 treebanks of 8 languages have all six types.

Table 1. **Overview of enhancements in UD 2.8. Number of nodes includes empty nodes. Enhanced relations per 1,000 nodes: G = gapping relations (i. e., to or from empty nodes); P = relations propagated from parents of coordination; S = relations propagated to shared dependents of coordination; X = relations to controlled external subjects; R = relations added in relative clauses (usually 2 for each clause); C = relation labels enhanced with case information**

Treebank	Nodes	G	P	S	X	R	C
Arabic PADT	282,460	2	48	10		16	286
Belarusian H.	305,406	4	44	0	6	16	148
Bulgarian BT.	156,149		7	4	3	16	118
Chukchi HSE	6,207	15					1
Czech CAC	495,497	15	65	22	5	24	184
Czech FicTree	167,371	11	55	19	5	21	127
Czech PDT	1,509,052	11	45	11	7	23	173
Czech PUD	18,623	4	39	1	7	26	172
Dutch Alpino	208,747	2	14	5	6	18	130
Dutch LassyS.	98,242	5	25	8	2	12	138
English EWT	254,857	1	21	9	14	11	128
English GUM	134,553	2	26	8	13	11	143
English GU. R.	16,286		19	11	18	14	123
English PUD	21,183	1	21	7	12	16	150
Estonian EDT	438,175	4	0	0		15	42
Estonian EWT	68,968	6				14	53
Finnish PUD	15,817	1					
Finnish TDT	202,453	6	33	32	5	0	
Italian ISDT	298,380	1	24	5	5	20	169
Latvian LVTB	252,961	13	36	32	9		158

End of the Table 1

Treebank	Nodes	G	P	S	X	R	C
Lithuanian A.	70,051	0	77	27	6	11	233
Polish LFG	130,967		12	8	7	0	85
Polish PDB	350,036		41	20			
Polish PUD	18,389		37	20			
Russian Syn.	1,107,741	4					3
Slovak SNK	106,184	5	39	7	4	15	151
Swedish PUD	19,085	2	22	8	11	28	158
Swedish Tal.	96,859	2	34	9	11	22	149
Tamil TTB	9,581		25			7	271
Ukrainian IU	122,324	10	51	8	9	17	

Similarly to the basic dependencies, there are parsing models that can generate the enhanced graphs for previously unseen text. Some successful parsers take advantage of the fact that many enhancements can be guessed based on the basic tree and combine a tree parsing model with enhancing heuristics. Enhanced UD parsers have been evaluated in two shared tasks run in connection with the IWPT 2020 and 2021 conference [Bouma et al. 2020].

3. What Is Next?

As Table 1 clearly shows, the enhanced representation, being optional, is only available for a fraction of the UD treebanks, and it does not grow as quickly as the data with the basic representation. [Droganova and Zeman 2018] note that this is unlikely to change, as more complex annotation requires more annotation effort, and it is thus difficult to get sufficient manpower to annotate data in a new language. They propose to at least apply a heuristic enhancer to all UD treebanks after each release and make this data available to the users. In addition, they propose heuristics to normalize syntactic alternations such as passive vs. active diathesis; they call the resulting data Deep UD, to distinguish it from the Enhanced UD defined in the official UD guidelines. Various other "enhanced-plus" variants have been proposed by [Schuster and Manning 2016] and others, but they are

not (yet?) approved as a part of UD. Even the existing guidelines sometimes leave room for multiple interpretations, leading to enhancement 'sub-types' that only appear in some treebanks.

The treatment of relative clauses can be extended to attributively used participles, as in the French example in Figure 5, where *fusée* "rocket" is modified by the participle *pouvant* "able", and at the same time it is also annotated as the external subject of the participle (as well as its complement infinitive *menacer* "threaten").

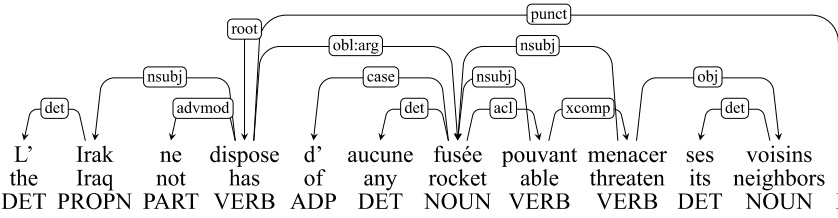


Fig. 5. **Enhanced UD graph of the French sentence *L'Irak ne dispose d'aucune fusée pouvant menacer ses voisins* (Iraq has no rockets that could threaten its neighbors)**

In pro-drop languages, empty nodes could help restore the coreference of controlled subjects if the main subject is missing, as the empty node (*on*) is used in Figure 6.

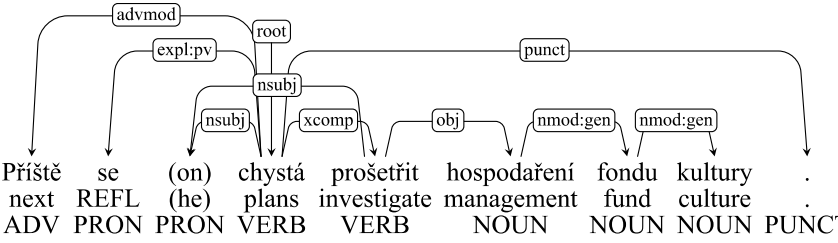


Fig. 6. **Enhanced UD graph of the Czech sentence *Příště se chystá prošetřit hospodaření fondu kultury* (Next time they are going to investigate the management of the culture fund)**

Finally, empty nodes in enhanced UD have been used to show the attachment of constituents that are incorporated in the verb and lack a node in the basic representation, like the Chukchi adverb *ныткы* "again" in Figure 7.

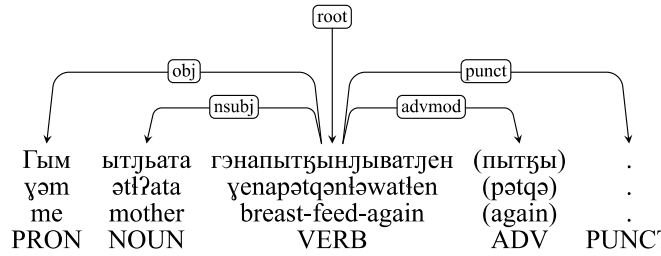


Fig. 7. **Enhanced UD graph of the Chukchi sentence *Гым ытԓьата гэнапытӷынԓыватԓен* (My mother was breast-feeding me again)**

4. Conclusion

We have presented the current state of the *enhanced representation* in Universal Dependencies. Being an optional and more complex annotation layer, it is only available for a fraction of the UD treebanks. Fortunately, a significant part of it can be computed or estimated with simple heuristics from the basic representation.

The guidelines for enhanced graphs are not considered as frozen as the basic guidelines in the UD community, and some details are still being elaborated as more languages are added. There is room for future additions of new variants of existing enhancement types or even completely new types with the same general motivation: to make otherwise implicit syntactico-semantic relations explicit and thus more easily accessible for language understanding applications.

References

1. *Bouma G., Seddah D., Zeman D.* (2020), Overview of the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies. In: Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies, pp. 151–161, ACL, Stroudsburg, PA, USA. URL: https://www.aclweb.org/anthology/2020.iwpt-1.16/ (date of access: 01.07.2021).
2. *de Marneffe M.-C., Manning C., Nivre J., Zeman D.* (2021), Universal Dependencies. In: Computational Linguistics. doi.org/10.1162/coli_a_00402
3. *Droganova K., Zeman D.* (2019), Towards Deep Universal Dependencies. In: Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019), Paris, France, pp. 144–152, URL: https://www.aclweb.org/anthology/W19-7717/ (date of access: 01.07.2021).
4. *Nivre J., de Marneffe M.-C., Ginter F., Hajič J., Manning C., Pyysalo S., Schuster S., Tyers F., Zeman D.* (2020), Universal Dependencies v2: An Evergrowing Multilingual

Treebank Collection. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), ELRA, Marseille, France, pp. 4034–4043. URL: https://www.aclweb.org/anthology/2020.lrec-1.497/ (date of access: 01.07.2021).
5. *Nivre J., Marongiu P., Ginter F., Kanerva J., Montemagni S., Schuster S., Simi M.* (2018), Enhancing Universal Dependency Treebanks: A Case Study. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), Bruxelles, Belgium, pp. 102–107. URL: https://www.aclweb.org/anthology/W18-6012 (date of access: 01.07.2021).
6. *Schuster S., Manning C.* (2016), Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Portorož, Slovenia, pp. 2371–2378. URL: https://www.aclweb.org/anthology/L16-1376/ (date of access: 01.07.2021).
7. *Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S.* (2018), CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Bruxelles, Belgium, pp. 1–21. URL: https://www.aclweb.org/anthology/K18-2001/ (date of access: 01.07.2021).

Daniel Zeman
Charles University, Faculty of Mathematics and Physics, ÚFAL (Czech Republic)
E-mail: zeman@ufal.mff.cuni.cz

С. О. Савчук

S. O. Savchuk

НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА В ЗЕРКАЛЕ СТАТИСТИКИ[1]

THE RUSSIAN NATIONAL CORPUS THROUGH THE PRISM OF STATISTICS

Аннотация. Основными критериями оценки текстовых корпусов, принятыми в корпусной лингвистике, являются объем, репрезентативность, сбалансированность. Особенностью НКРЯ как большого представительного корпуса является то, что за все время его существования рост объема осуществлялся пропорционально, с соблюдением баланса в функциональном разнообразии текстов. В связи с планами дальнейшего развития корпуса был проведен масштабный анализ состава текстов НКРЯ по основным параметрам, которые традиционно используются для оценки репрезентативности корпусов. Среди них дата создания текстов, сфера функционирования, жанр и тематика текстов. В статье представлены результаты анализа статистической информации об основных метатекстовых параметрах корпуса.
Ключевые слова. Национальный корпус русского языка, состав текстов, репрезентативность, сбалансированность.

Abstract. The main criteria for evaluating text corpora accepted in corpus linguistics are corpus size, representativeness, and balance. A peculiarity of the RNC design as a large representative corpus is that its size increasing was proportional, while maintaining a balance in the functional diversity of texts. A large-scale analysis of text composition of the RNC was carried out using the parameters traditionally applied for evaluating corpora representativeness. Among them are date of text creation, functional sphere, genre, and text topic. The article presents the results of the analysis of statistical information about the main metatextual parameters of the RNC.
Keywords. The Russian National Corpus, corpus composition, representativeness, balance.

1. Введение

Национальный корпус русского языка проектировался в начале 2000-х годов как корпус современного русского языка по образцу больших национальных корпусов объемом 100 млн словоупотреблений. Это особый тип большого представительного корпуса, отражающий функционирование национального языка в современный период. За 20 лет существования ядро корпуса увеличилось в объеме и обросло системой специальных корпусов, но при этом, согласно статистике запросов, основной корпус письменных текстов остается самым вос-

[1] Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках Соглашения № 075-15-2020-793.

требованным корпусом в составе НКРЯ. Но за это время в корпусной лингвистике была налажена технология создания интернет-корпусов очень большого объема, от одного до десятков миллиардов, так что на современном этапе НКРЯ сосуществует с другими корпусами русского языка, значительно превосходящими его по объему. Среди них ГИКРЯ [Беликов и др. 2013], семейство корпусов Araneum Russicum [Benko 2014], корпусы, доступные в системе Sketch Engine, например, RuTenTen [Jakubíček et al. 2013], корпусы русских книг в составе Google Books [Захаров и др. 2014]. Первоначальное увлечение гигантскими объемами веб-корпусов сменилось более взвешенной позицией. Специальные исследования, проведенные на материале корпусов разных типов, объективно, на наш взгляд, оценивают преимущества и недостатки каждого типа [Захаров 2015; Пиперски 2020; Davies 2019; Nesset 2019; Sharoff 2020 и др.].

В связи с возникшей возможностью и необходимостью сравнения разных корпусов в зоне внимания опять оказались такие критерии оценки корпусов, как репрезентативность и сбалансированность. Имеются разные подходы к определению репрезентативности. В частности, есть мнение, что под репрезентативностью понимается способность корпуса отражать свойства всей популяции текстов, но поскольку мы не знаем этих свойств в полном объеме, ни один корпус не является зеркалом того, с чем мы сталкиваемся в реальной жизни, поэтому репрезентативность — это внутреннее свойство корпуса. Согласно другой позиции, репрезентативность — это способность корпуса представлять максимальное разнообразие текстов на каком-либо языке. Предлагается ориентированная на восприятие интерпретация репрезентативности, хотя отмечается, это понятие невозможно рассчитать и описать строго математически. В целом в определении репрезентативности исследователи сходятся в том, что это нечто недостижимое, к чему нужно стремиться [Viana et al. 2015: 3–4, 66–68, 102, 118, 160]. Сбалансированность — другая сторона репрезентативности. Это свойство корпуса представлять разнообразные тексты в определенных заданных пропорциях помогает, в частности, избежать субъективизма при отборе текстов при составлении корпуса.

НКРЯ предназначен как для исследований современного языка, так и для изучения микродиахронических изменений. Обладает ли он необходимыми свойствами для этого? Согласно М. Дэвису, корпус для изучения языковых изменений должен соответствовать нескольким критериям: 1) иметь большой объем (не менее 100 млн словоу-

потреблений), что позволяет отслеживать редкие языковые явления; 2) включать современные тексты (желательно пополняться ежегодно); 3) содержать тексты, относящиеся к разным жанрам, в определенном соотношении; 4) сохранять жанровый баланс от периода к периоду; 5) иметь архитектуру, которая позволяла бы прослеживать частотные характеристики слов в разные периоды и сравнивать их между собой [Davies 2012].

Как представляется, НКРЯ отвечает всем этим требованиям, что будет показано в дальнейшем. Его текущий объем составляет более 320 млн словоупотреблений, то есть почти в 10 раз превышает первоначальный объем при открытии ресурса на сайте ruscorpora.ru. При этом увеличение объема осуществлялось пропорционально, с соблюдением баланса в функциональном разнообразии текстов.

2. Анализ текущего распределения текстов основного корпуса

В связи с проводимыми в настоящее время системными и инфраструктурными изменениями в программном обеспечении корпуса и разработкой перспектив его дальнейшего развития было предпринято исследование текстового состава корпусов по основным метатекстовым параметрам, служащим для оценки его репрезентативности. Программное обеспечение для анализа разработано А. Хаджийской. Были получены данные о распределении текстов по дате создания, сферам функционирования, жанрам художественной литературы, типам и основной тематике по состоянию на декабрь 2020 г. Ниже перечислены результаты анализа статистической информации об основных метатекстовых параметрах корпуса письменных текстов.

2.1. Распределение по дате создания

Поскольку основной корпус используется для диахронических исследований, он должен давать возможность прослеживать изменения в значениях слов, сочетаемости и пр. на протяжении определенных периодов. Поэтому в идеальном случае каждый период должен быть представлен приблизительно равным количеством словоупотреблений. На уровне подкорпусов, из которых исторически складывался основной корпус, — XIX в., 1-й пол. XX в., 2-й пол. XX в. и XXI в. — картина близка к идеальной (см. рис. 1).

Однако если увеличить масштаб и рассмотреть распределение текстов по 10-летиям, то его нельзя назвать сбалансированным (рис. 2).

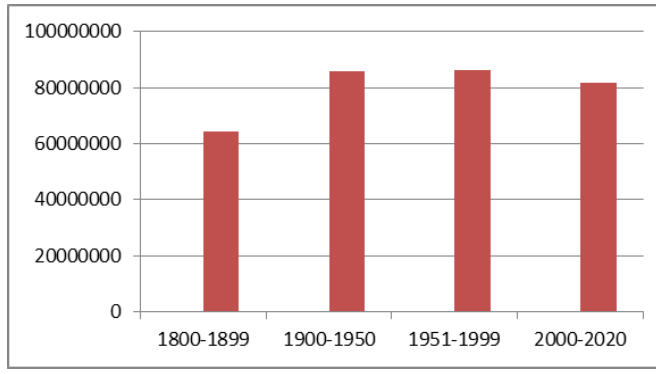


Рис. 1. **Распределение текстов основного корпуса письменных текстов по периодам**

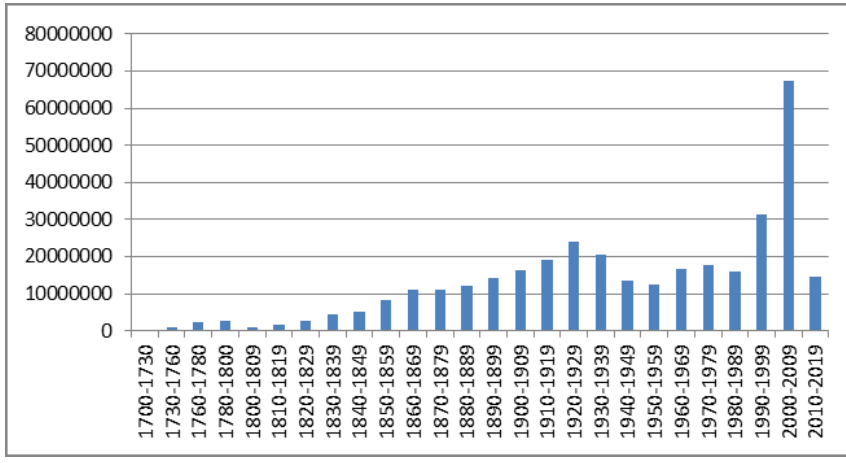


Рис. 2. **Распределение текстов основного корпуса письменных текстов по дате создания**

Подкорпус XVIII века пока очень мал и пополняется медленно из-за сложности подготовки текстов. Возможно, его следует рассматривать отдельно, а для сравнительного анализа использовать больший хронологический интервал, например, 1700–1730, 1730–1760, 1760–1780, 1780–1800.

Этому есть свое объяснение. Пик на графике, который приходится на начало 2000-х годов, соответствует той текстовой коллекции, с которой начинался НКРЯ как корпус современных текстов. Дальнейшее

развитие корпуса — составление корпусов предшествующих периодов, которое еще не завершено. Особого внимания требуют тексты 1800–1830-х годов — по сложности подготовки они мало отличаются от текстов XVIII в. Другая задача, которую предстоит решить, — создание полноценной коллекции текстов 2010–2020 гг., сбалансированной по разным функциональным сферам.

2.2. Распределение по сферам функционирования

Центральным элементом организации текстового состава НКРЯ является распределение текстов по функциональным сферам. Это самая общая типологическая характеристика текста, которая определяет отнесение текста к одной из социально значимых областей общественно-речевой практики и используется в соответствии с традицией отечественной функциональной стилистики и типологии текста [Савчук 2005: 70–71]. Для описания текстов НКРЯ выделено 9 функциональных сфер: художественная, публицистика (и массовая информация), церковно-богословская, учебно-научная, официально-деловая, производственно-техническая, реклама, бытовая, электронная коммуникация. Отметим, что содержательно это деление соответствует тому, что называется жанровой классификацией в других национальных корпусах (ср. «художественная литература», «журналы», «газеты», «академические тексты», «разговорные тексты» в британском корпусе или «художественная литература», «журналистика», «научные» и «административные тексты» в чешском корпусе). Результаты распределения текстов по речевым сферам в статике представлены в табл. 1.

Для того чтобы можно было сравнивать частотные характеристики единиц в разные периоды, необходимо, чтобы корпус сохранял примерно одинаковое соотношение текстов, относящихся к разным функциональным сферам в разные периоды. В этом случае различия в частотных характеристиках языковых единиц будут иметь языковую природу, а не объясняться несбалансированностью состава корпуса.

Табл. 2 объединяет две характеристики корпуса и представляет распределение текстов по сферам функционирования в динамике. Соотношение текстов разных функциональных сфер показано в разные периоды — в XIX в., в 1-й пол. XX в., во 2-й пол. XX в., в нач. XXI в.

Как видно из таблицы, в целом баланс распределения текстов по сферам функционирования сохраняется в каждом из выделенных периодов, за исключением отдельных отклонений, которые предстоит исправить.

Таблица 1. **Распределение текстов по сферам функционирования**

Сфера функционирования	Кол-во словоупотр.	Доля в %
Художественная	134 963 105	42,75 %
Публицистика	123 213 170	39,03 %
Учебно-научная	34 956 396	11,07 %
Официально-деловая	4 653 591	1,47 %
Производственно-техническая	1 469 354	0,47 %
Церковно-богословская	4 282 472	1,36 %
Реклама	713 666	0,23 %
Бытовая	10 915 179	3,46 %
Электронная коммуникация	3 534 997	1,12 %
Всего словоформ	315 691 990	100,95 %

Таблица 2. **Распределение текстов разных сфер функционирования по периодам**

Сферы функционир.	1800–1899	1900–1950	1951–1999	2000–2020
Художественная	49,94 %	44,31 %	49,28 %	31,03 %
Публицистика	32,96 %	36,48 %	37,13 %	49,41 %
Учебно-научная	10,49 %	11,63 %	8,30 %	12,35 %
Офиц.-деловая	1,05 %	1,50 %	0,86 %	1,32 %
Произв.-техн.	0,02 %	0,42 %	0,88 %	0,42 %
Церк.-богосл.	0,79 %	1,44 %	2,00 %	0,31 %
Реклама	0,01 %	0,06 %	0,09 %	0,69 %
Бытовая	6,35 %	5,52 %	1,96 %	0,42 %
Электр. комм.	0,00 %	0,00 %	0,00 %	4,32 %
Всего словоформ	64 314 169	85 872 427	86 120 684	81 804 968

Для более детального анализа получены данные распределения текстов по десятилетиям каждого периода — как для всего корпуса, так и для каждой текстовой коллекции, представляющей разные функциональные сферы. Анализ позволяет найти участки, которые

в первую очередь нуждаются как в количественном пополнении, так и в выравнивании соотношения текстов разных функциональных сфер.

2.3. Соотношение художественных текстов по жанрам в разные периоды

Было исследовано более мелкое членение текстов корпуса на жанры и типы. Для художественных текстов важна отнесенность текста к литературному жанру, она во многом определяет тематическое содержание и языковые особенности текста (ср., например, разницу в колорите научно-фантастической повести и документальной прозы).

В разные периоды соотношение по жанрам должно быть ожидаемо разным, поскольку жанры возникают, развиваются и исчезают или трансформируются. Например, детектив как жанр в русской литературе появился в XX в., его предшественники в XIX в. — авантюрный роман (например, «Петербургские трущобы» В. Крестовского) и судебный очерк. Как показал анализ жанрового состава основного корпуса, нежанровая проза (которая в электронных библиотеках маркируется как «современная проза»), в XVIII–XIX вв. преобладает над всей жанровой, вместе взятой (62,6 %), и составляет около 50 % в XX–XXI вв. Объясняется это тем, что классификация жанров в корпусе отражает жанровый состав современной литературы, а в XIX в. эта система была иной, какие-то жанры только складывались в пределах нежанровой прозы. Жанровое соотношение художественных текстов корпуса в разные периоды представлено в табл. 3.

2.4. Соотношение нехудожественных текстов по типам

Для корпуса значимо соотношение нехудожественных текстов по типам текстов. В НКРЯ используется подробная система кодирования текстовых типов, привязанная к различным сферам функционирования. Первоначальный список типов был открытым: ограничения на состав типов не делалось, поскольку производилось сплошное описание текстов, а не отбор текстов по заданному списку типов. Предполагалось, что помимо распространенных типов в корпус будут попадать и менее частотные, так что список будет расширяться и со временем стабилизируется. Так и произошло, и в настоящее время мы имеем некоторый основной набор типов, представленный на панели отбора текстов, описывающий подавляющее большинство текстов. И есть некоторый набор типов, представленный малым количеством и даже единичными текстами.

Таблица 3. **Соотношение художественных текстов по жанрам в разные периоды**

Жанры худож. лит-ры	1800–1899	1800–1899, %	1900–1949	1900–1949, %
Детектив, боевик	197 022	0,61 %	463 317	1,22 %
Детская	402 639	1,25 %	2 133 894	5,61 %
Документ. проза	1 438 694	4,48 %	3 719 641	9,78 %
Историч. проза	4 645 672	14,46 %	5 967 662	15,68 %
Любовная история	185 250	0,58 %	229 501	0,60 %
Нежанр. проза	20 112 241	62,61 %	20 444 581	53,73 %
Приключения	1 244 519	3,87 %	2 476 796	6,51 %
Фантастика	474 126	1,48 %	1 821 586	4,79 %
Юмор и сатира	2 252 473	7,01 %	924 997	2,43 %
Драматургия	1 351 398	4,21 %	999 619	2,63 %
Прочее	511 438	1,59 %	38 091	0,10 %
Всего	32 121 476	102,16 %	38 050 694	103,07 %

Жанры худож. лит-ры	1950–1999	1950–1999, %	2000–2020	2000–2020, %
Детектив, боевик	1 816 749	4,28 %	5 338 492	21,03 %
Детская	3 018 028	7,11 %	510 539	2,01 %
Документ. проза	5 624 811	13,25 %	1 238 415	4,88 %
Историч. проза	2 083 389	4,91 %	426 530	1,68 %
Любовная история	139 553	0,33 %	668 577	2,63 %
Нежанр. проза	22 315 795	52,59 %	14 304 336	56,36 %
Приключения	1 398 094	3,29 %	130 203	0,51 %
Фантастика	4 184 241	9,86 %	2 200 695	8,67 %
Юмор и сатира	1 353 040	3,19 %	855 915	3,37 %
Драматургия	861 158	2,03 %	232 524	0,92 %
Прочее	38 189	0,09 %	304 560	1,20 %
Всего	42 436 986	100,93 %	25 380 417	103,27 %

Проведенный анализ позволил оценить действующий стандарт разметки текстовых типов и определить, какие типы текстов в пределах каждой функциональной сферы плохо представлены (менее

100 текстов). В частности, в сфере публицистики недостаточно представлены такие публицистические жанры (типы), как *автобиография* (33), *аннотация* (54), *блог* (18), *доклад* (85), *заявление* (55), *календарь* (14), *лекция* (19), *листовка* (35), *манифест* (13), *описание* (18), *памфлет* (7), *послание* (17), *пресс-конференция* (2), *путеводитель* (12), *рецепт* (66). В учебно-научной сфере меньше 100 текстов приходится на типы: *доклад* (59), *задача* (30), *инструкция* (44), *конспект* (20), *лекция* (49), *обзор* (35), *отзыв* (5), *отчет* (80), *справочник* (3), *тезисы* (3), *учебник* (20), *учебное пособие* (61), *методические материалы* (54) и др. Составлен список таких типов текстов для каждой функциональной сферы. Поиск, обработка и включение в состав корпуса текстов указанных типов рассматривается как одна из первоочередных задач пополнения основного корпуса.

2.5. Соотношение нехудожественных текстов по тематике

Включение в состав НКРЯ текстов всех сфер функционирования обеспечило тематическое разнообразие текстов, поскольку каждая сфера бытования языка связана с определенным набором тематик. Наиболее широк этот набор в сфере публицистики, всеобъемлющ в художественной сфере (настолько, что кодирование по этому параметру становится бессмысленным) и требует мониторинга и коррекции в учебно-научной сфере, чтобы при отборе текстов обеспечить представленность в корпусе всех отраслей науки в равной мере.

Анализ корпуса по данному параметру позволил определить, какие тематики в пределах каждой функциональной сферы плохо представлены (менее 100 текстов). В частности, в официально-деловой сфере подавляющее число текстов относится к тематическим областям *политика и общественная жизнь* (913), *администрация и управление* (516), *армия и вооруженные конфликты* (480), *бизнес, коммерция, экономика, финансы* (199), и напротив, слабо представлены такие тематики, как *сельское хозяйство* (8), *спорт* (4), *строительство и архитектура* (11), *техника* (13), *транспорт* (12) и др. Поиск, обработка и включение в состав корпуса официальных документов, относящихся к тематическим областям, недостаточно представленным в коллекции деловой речи, будут учтены при составлении плана пополнения основного корпуса. Аналогичные списки и планы пополнения составлены для других функциональных сфер.

Остановимся подробнее на соотношении тематических областей, связанных с разными науками. Отдельно оценивалось общее состо-

яние по всему массиву текстов и соотношение в текстах с 2010 года. Наилучшим образом представлены гуманитарные науки (*история, политология, религиоведение, право, образование, психология, культурология*), а из прикладных и естественных наук — *медицина, военное дело, техника, биология, геология и география*. Сравнительно небольшой объем текстов (до 2 млн словоупотреблений) приходится на такие области, как *химия, физика, строительство и архитектура, сельское хозяйство, транспорт*. Совсем слабо, до 1 млн словоупотреблений, представлены *астрономия, информатика, энергетика, статистика и лесное хозяйство*.

Что касается текстов, созданных после 2010 года, то первая группа научных областей, наиболее полно представленных в текстах последнего десятилетия, по составу в основном совпадает с общей по корпусу. Однако количество тематических областей, недостаточно представленных в новейших текстах, несколько больше — к *информатике*, *статистике*, *лесному хозяйству* присоединились *геология* и *география*, *математика*, *химия*, *сельское хозяйство, искусствоведение*. На основе оценки состояния были установлены целевые показатели для репрезентации разных наук в пределах учебно-научного стиля. Ситуация с современными научными текстами будет исправлена в ближайшее время, с включением в состав корпуса большой коллекции текстов научных журналов и популярных научных сайтов (Arzamas, postnauka.ru, polit.ru, pro-science.ru, indicator.ru и др.).

3. Заключение

Проведенный анализ основного корпуса письменных текстов по основным метатекстовым параметрам показал, что по состоянию на начало 2021 года состав корпуса отличается достаточным разнообразием. Что касается распределения по хронологической оси, то в количественном отношении удовлетворительно представленным можно считать период со 2-й пол. XIX в. до наших дней: в этом интервале объем подкорпуса за каждое 10-летие превышает 10 млн словоупотреблений, причем в отдельные 10-летия многократно. Задача ближайшего будущего состоит в том, чтобы сделать это распределение более равномерным. Особого внимания требуют подкорпус текстов раннего периода — XVIII в. — 1-й пол. XIX в., столь важный для диахронических исследований, а также коллекция новейших текстов, отражающих текущий момент. Если вторая задача решается достаточно про-

сто, то первая, при соблюдении высоких стандартов качества текстов, требует значительных усилий и времени.

Что касается объема корпуса, то не следует забывать о растущем газетном корпусе (период 2000–2019 гг., объем, включая региональные газеты, 350 млн словоупотреблений) и корпусе устных текстов (13,3 млн словоупотреблений), которые представляют современный язык. Историческую часть дополняет старорусский корпус (8 млн словоупотреблений) и поэтический корпус, в котором значительную долю составляет поэзия XIX в. (около 5 млн словоупотреблений). Организация кросс-корпусного поиска объединит все эти коллекции в один массив (объемом около 700 млн словоупотреблений), тем самым частично решив проблему увеличения размера корпуса.

Для того чтобы при пополнении корпуса новыми текстами добиться равномерного распределения по основным текстовым параметрам в пределах каждого периода, предстоит сочетать две стратегии. Традиционная стратегия «естественного» прироста, которая при отборе текстов опирается на библиографические критерии (автор, издание, год издания) без учета жанровой или тематической принадлежности, будет дополнена стратегией «искусственной поддержки» редких жанров, при которой тексты отбираются по текстологическим критериям в заданном количестве, что поможет повысить представленность в корпусе отдельных текстовых типов и жанров. Распределение текстов по функциональным сферам будет осуществляться в соответствии с целевыми показателями баланса корпуса, что обеспечит пропорциональное соотношение текстов разных функциональных сфер по периодам.

Подобная программа развития корпуса обеспечит поступательный рост его объема, который в обозримом будущем может достигнуть 1 млрд словоупотреблений. Что касается «больших данных», то они остаются прерогативой интернет-корпусов русского языка, развитие которых идет как по пути наращивания объемов, так и усложнения структурной организации данных.

Литература

1. *Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А.* (2013), Корпус как язык: от масштабируемости к дифференциальной полноте. Компьютерная лингвистика и интеллектуальные технологии: Материалы международной конференции «Диалог-2013». Вып. 12(19), т. 1, с. 84–95.

2. *Захаров В. П.* (2015), Сочетаемость через призму корпусов. Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог». Вып. 14(21), М.: РГГУ, с. 667–682.
3. *Захаров В. П., Масевич А. Ц.* (2014), Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer. Структурная и прикладная лингвистика. Вып. 10. СПб: Изд-во С.-Петерб. ун-та, 2014, с. 303–327.
4. *Пиперски А. Ч.* (2020), Русский язык и корпусное разнообразие. Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог». Вып. 19(26), с. 84–95.
5. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka, A. Horak, I. Kopeček, K. Pala (eds.). Text. Speech and Dialogue. 17th International Conference. TSD 2014. Brno, Czech Republic. September 8–12, 2014. Springer International Publishing, pp. 257–264.
6. *Davies M.* (2012), Looking at Recent Changes in English with the Corpus of Contemporary American English (COCA). 21st Century Text (Peer-reviewed, online journal).
7. *Davies M.* (2019), Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In: C. Suhr, T. Nevalainen, I. Taavitsainen (eds.). From data to evidence in English language research (Digital Linguistics). Leiden: Brill, pp. 66–87.
8. *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V.* (2013), The TenTen Corpus Family. In Proceedings of the International Conference on Corpus Linguistics, pp. 125–127.
9. *Nesset T.* (2019), Big data in Russian linguistics? Another look at paucal constructions. Zeitschrift für Slawistik. Vol. 64(2), pp. 157–174.
10. *Sharoff S.* (2020), Topography of internet corpora. Компьютерная лингвистика и интеллектуальные технологии: По материалам международной конференции «Диалог». Вып. 19(26), доп. том, с. 1134–1137.
11. *Viana V., Zyngier S., Barnbrook G.* (eds.) (2011), Perspectives on Corpus Linguistics: Connections & Controversies. Philadelphia: John Benjamins.

References

1. *Belikov V. I., Kopylov N. Yu., Piperski A. Ch., Selegej V. P., Sharov S. A.* (2013), Korpus kak yazyk: ot masshtabiruemosti k differencialnoj polnote [Corpus as language: from scalability to register variation]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Materialy mezhdunarodnoj konferentsii "Dialog-2013" [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog-2013"]. Vol. 12(19), pp. 84–95.
2. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka, A. Horak, I. Kopeček and K. Pala (eds.). Text. Speech and Dialogue. 17th International Conference. TSD 2014. Brno. Czech Republic. September 8–12, 2014. Springer International Publishing, pp. 257–264.
3. *Davies M.* (2012), Looking at Recent Changes in English with the Corpus of Contemporary American English (COCA). 21st Century Text (Peer-reviewed, online journal).

4. *Davies M.* (2019), Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In: From data to evidence in English language research (Digital Linguistics). In: C. Suhr, T. Nevalainen, I. Taavitsainen (eds). Leiden: Brill, pp. 66–87.
5. *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V.* (2013), The TenTen Corpus Family. Proceedings of the International Conference on Corpus Linguistics, pp. 125–127.
6. *Nesset T.* (2019), Big data in Russian linguistics? Another look at paucal constructions. Zeitschrift für Slawistik. Vol. 64(2), pp. 157–174.
7. *Piperski A. Ch.* (2020), Russkij yazyk i korpusnoe raznoobrazie [Russian language and corpus diversity]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii "Dialog" [Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog"]. Vol. 19(26), pp. 84–95.
8. *Sharoff S.* (2020), Topography of internet corpora. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii "Dialog" [Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog"]. Vol. 19(26), pp. 1134–1137.
9. *Viana V., Zyngier S., Barnbrook G.* (eds.). (2011), Perspectives on Corpus Linguistics: Connections & Controversies. Philadelphia: John Benjamins.
10. *Zakharov V. P.* (2015), Sochetaemost cherez prizmu korpusov [Set Phrases: a View through Corpora]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii "Dialog" [Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog"]. Moscow: RGGU, vol. 14(21), pp. 667–682.
11. *Zakharov V. P., Masevich A. C.* (2014), Diahronicheskie issledovaniya na osnove korpusa russkih tekstov Google Books Ngram Viewer [Diachronic investigations on the base of Russian corpus of Google Books Ngram Viewer]. In: Strukturnaya i prikladnaya lingvistika [Structural and applied linguistics]. Iss. 10. St. Petersburg: St. Petersburg University Publishing house, pp. 303–327.

Савчук Светлана Олеговна
Институт русского языка им. В. В. Виноградова РАН (Россия)
Savchuk Svetlana
Vinogradov Russian Language Institute or
the Russian Academy of Sciences (Russia)
E-mail: savsvetlana@mail.ru

С. П. Тимошенко, Л. Л. Иомдин,
С. А. Гладилин, Е. С. Иншакова

S. P. Timoshenko, L. L. Iomdin,
S. A. Gladilin, E. S. Inshakova

СИНТАГРУС В СОСТАВЕ НКРЯ: НОВЫЕ ВОЗМОЖНОСТИ[1]

SYNTAGRUS WITHIN THE RUSSIAN NATIONAL CORPUS: NEW POSSIBILITIES

Аннотация. Данная статья знакомит читателей с актуальным состоянием корпуса СинТагРус, уделяя особое внимание структурной организации разметок разных типов. На основании этого материала формулируются проблемы, без решения которых невозможно осуществление полноценного поиска по корпусу, и предлагаются некоторые решения.
Ключевые слова. Синтаксис зависимостей, анафорическая разметка, лексико-функциональная разметка, микросинтаксическая разметка, темпоральная разметка, эллипсис, поиск.

Abstract. The paper describes the current state of the SynTagRus corpus, with a particular focus on the structure of its annotation. There are 8 different types of annotation interplaying in the corpus: morphological, syntactic, lexical-functional, anaphoric, microsyntactic, temporal, the annotation of elliptical sentences and the annotation of word meanings. Based on their description, we address the challenges with which the search engine developers are faced. In particular, the search interface should allow multiword queries, where a word can be linked to any other in as many ways as there are annotation types in corpus. Unlike morphological characteristics and other word-based annotations, which are unique, links between words satisfying the query conditions may not be unique, so, the search algorithm must be able to iteratively look for any of them. So, when formulating a query, the user should indicate how the search mechanism should treat the link specified: stop after the first occurrence is found or continue to find all of them.
Keywords. Dependency syntax, anaphora marking, lexical functions, microsyntax markup, temporal markup, ellipsis, search.

1. Вводные замечания

Входящий в состав Национального корпуса русского языка «Синтаксически размеченный корпус русского языка» (или СинТагРус, от **Syn**tactically **Tag**ged **Rus**sian Corpus), разрабатываемый в лаборатории компьютерной лингвистики ИППИ РАН, пополняется новыми типами разметки — см., в частности, описание в [Иншакова и др. 2019]. На данный момент в корпусе имеется 8 видов разметки: 1) морфологическая; 2) синтаксическая; 3) разметка эллипсиса; 4) лексико-функциональная; 5) микросинтаксическая; 6) анафорическая; 7) темпоральная; 8) разметка значений многозначных слов.

[1] Данная работа выполнена при поддержке гранта 19-07-00842 РФФИ.

Авторы корпуса хотят сделать общим достоянием все результаты своей работы. Для этого требуется выполнить два условия: во-первых, обеспечить корпус достаточным количеством справочных материалов, чтобы разобраться в его содержимом мог любой желающий; во-вторых, обеспечить физическую доступность данных. Что касается физической доступности, то она пока осуществляется Национальным корпусом русского языка (НКРЯ) не в полном объеме: далеко не все, чем корпус в принципе располагает, можно найти с помощью действующего поискового интерфейса. Из перечисленных разметок на сайте НКРЯ доступны для ознакомления только морфологические, синтаксические и лексико-функциональные данные.

Разнообразие разметки порождает ряд специфических проблем поиска: как сделать так, чтобы пользователь мог переходить от одного типа разметки к другому и сочетать их между собой? На первый взгляд, задача кажется чисто технической, однако ее решение во многом зависит от внутренних лингвистических установок того, кто осуществляет поиск.

Данная статья призвана, во-первых, описать актуальное состояние корпуса, уделяя особое внимание формату разметки. Во-вторых, сформулировать требования к организации поиска, которые бы учитывали, с одной стороны, структурные особенности разметки, а с другой — потребности и внутренние установки пользователей. Посвященная поиску часть статьи требует особого внимания корпусного лингвистического сообщества: широкое обсуждение этих вопросов может определяющим образом повлиять на эффективность проведения новых корпусных исследований.

2. Структура разметки СинТагРуса

В самом абстрактном виде любая разметка текста — это набор помет при элементах этого текста. Лингвист базовым элементом текста считает слово, так что первая стадия создания любой разметки — это разбиение на слова, точнее, на словоформы, которые считаются конкретными реализациями слов в тексте. Как правило, словоформа определяется формально как последовательность букв (шире — знаков), расположенная между пробелами. На месте пробела также может выступать пунктуационный знак. Подход, используемый для нахождения границ в СинТагРус'е, составляет одну из его особенностей. Границы словоформ не во всех случаях совпадают с традиционным делением.

Своим существованием СинТагРус обязан лингвистическим процессорам серии ЭТАП, которые задумывались прежде всего как средство межязыкового перевода. Для тех случаев, когда выражение из нескольких слов функционирует как цельная единица и не путается (во всяком случае, систематически) со словосочетаниями, в морфологическом и комбинаторном словарях системы ЭТАП имеется соответствующая статья. Например, существует отдельная статья для выражения *что бы то ни было*. Это означает, что данное выражение — с точки зрения словарей — приравнено к слову. Во время работы процессор разбивает текст на словоформы в соответствии со статьями в своих словарях. Соответственно, вхождение в текст выражения *что бы то ни было* будет представлять одну словоформу, а не пять. С одной стороны, деление текста на словоформы не по пробелам, а «по словарю» является более содержательным — грамматически, синтаксически и семантически мотивированным. Можно сказать, что текст действительно поделен на слова — единицы, заданные сочетанием грамматической формы и синтаксической функции. Используя в дальнейшем термин «слово» применительно к СинТагРус'у, мы будем понимать под ним именно такое сочетание: отрезок текста (возможно, с пробелом или некоторыми другими неалфавитными символами), соответствующая ему начальная форма (лемма) и набор грамматических характеристик.

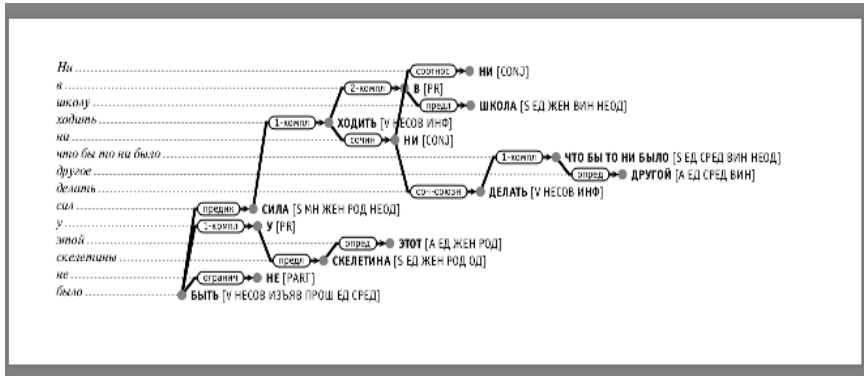


Рис. 1. **Синтаксическая структура фразы, содержащей выражение *что бы то ни было***

Опора на словари имеет и другое преимущество. Если рассматривать корпус как базу данных, то наличие словарей делает ее структуру

нормализованной: словари работают как справочная таблица, которая используется для контроля корректности корпусных данных.

С другой стороны, нерегулярность, выражающаяся в том, что в большинстве случаев пробелы отделяют одну словоформу от другой, но иногда оказываются «внутри» словоформы, усложняет работу с корпусом как для пользователей, так и для разработчиков. Поэтому поисковый интерфейс НКРЯ в текущем варианте не позволяет искать единицы типа *что бы то ни было*.

На момент написания этой статьи в состав СинТагРус'а входит 731 текст. Технически это 731 файл в формате, представляющем собой надстройку над xml. Тексты в файлах разбиты на 87 676 фрагментов, каждый из которых заключен в тег <S> — это предложения. Содержимое тега S, в свою очередь, разбито на фрагменты, заключенные в тег W, — это слова. Всего в СинТагРусе 1 239 942 таких слова.

3. Базовая разметка слова

Мы проиллюстрируем базовый формат разметки слова на примере вхождения *что бы то ни было* в предложение, приведенное на рис. 1. В исходном файле оно выглядит так:

```
<W DOM=»8» FEAT=»S ЕД СРЕД ВИН НЕОД» HYPOT=»1-компл.11» ID=»6»
KSNAME=»ЧТО$БЫ$ТО$НИ$БЫЛО» LEMMA=»ЧТО БЫ ТО НИ БЫЛО» LINK=»1-
компл»>что бы то ни было</W>
```

Текстовый фрагмент никак не выделен, синим цветом выделен собственно тег, красным — его атрибуты, а фиолетовым — их значения. Атрибут DOM содержит информацию о синтаксическом хозяине слова. В качестве значения используется порядковый номер слова в предложении. Атрибут FEAT содержит набор грамматических характеристик. Атрибут HYPOT может показаться дублером атрибута LINK, хранящего имя синтаксического отношения между словом и его хозяином. На самом деле HYPOT предназначен для хранения имени правила ЭТАПа, обеспечившего построение этого синтаксического отношения между словом и его хозяином. ID — идентификатор, порядковый номер слова в предложении. KSNAME — это ссылка на статью комбинаторного словаря (далее — КС) системы ЭТАП, а LEMMA — начальная форма слова.

Разметка эллипсиса и разметка значений многозначных слов хранятся в виде атрибутов слова. Анафорическая и микросинтаксическая разметки хранятся в виде атрибутов предложения. Лексико-функцио-

нальная и темпоральная разметки хранятся в виде специальных тегов, которые так же, как и тег слов W, «вложены» в тег предложения S.

4. Разметка эллипсиса

В СинТагРус'е принят следующий подход: эллипсис понимается как пропуск на поверхностном уровне одного или нескольких синтаксических элементов. Соответственно, пропущенные слова восстанавливаются в структуре и имеют все полагающиеся им атрибуты, а также помету FANTOM (значение дополнительного атрибута при теге W), позволяющую понять, что данному слову соответствует фрагмент текста нулевой длины (однако у него есть линейная позиция).

На рис. 2 представлена структура предложения *К людям врожденную любовь слон испытывает, особенно к красивым женщинам, но еще большую — к маленьким детям.* В нем восстановлено два пропуска, каждый из которых включает два слова — *испытывать* и *любовь*. В левом столбце видны пустые строки. Таково наглядное представление эллипсиса в действующей версии СинТагРус'а на сайте НКРЯ, однако использовать соответствующую помету для поиска там нельзя.

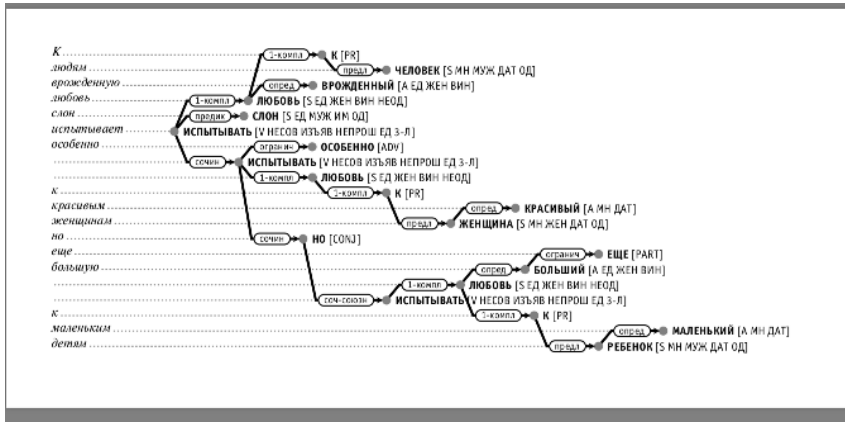


Рис. 2. **Синтаксическая структура предложения с эллипсисом**

К эллипсису относятся как случаи, когда можно дословно восстановить пропущенное, так и случаи, когда очевиден только смысл пропуска, но не его конкретное лексическое наполнение. Нулевая связка не считается эллипсисом и в структуре не восстанавливается. На дан-

ный момент в СинТагРус'е количество восстановленных элидированных слов составляет 2693.

5. Разметка значений многозначных слов

Как уже говорилось, атрибут слова KSNAME хранит имя статьи комбинаторного словаря (КС) систем ЭТАП. Различных лемм в СинТагРус'е на данный момент насчитывается 50 350, а статей КС цитируется 41 812. Тот факт, что лемм больше, чем статей КС, отчасти объясняется тем, что одна и та же лемма может соответствовать двум и более статьям КС. Это происходит в первую очередь потому, что в корпусе встречается заметное число слов, отсутствующих в комбинаторном словаре (это могут быть собственные имена, названия и просто редкие слова или редкие значения многозначных слов). Для всех таких единиц в качестве леммы используется начальная форма слова, а в качестве имени статьи КС выступает имя некоторой дежурной статьи КС (например, ФИКТ-ФАМИЛИЯ, ФИКТ-ЛИЧИМЯ, ФИКТ-МЕСТО и т. п.): понятно, что такие единицы встречаются многократно.

Принципы выделения значений в словарях системы ЭТАП обусловлены, во-первых, традицией академических словарей, во-вторых, исторической задачей обеспечения автоматического перевода, а в-третьих, ориентацией на автоматическую разметку. Отдельные значения выделяются тогда, когда это продиктовано нуждами перевода или семантического анализа и когда их можно достаточно надежно детектировать автоматически на основе контекста, в частности, по реализованным валентностям или лексическим функциям. Поэтому значения в КС по сравнению с обычными толковыми словарями укрупнены: слово ТОЧКА представлено в КС двумя статьями — ТОЧКА 1 объединяет все центральные значения («малый пространственный или временной объект», «знак препинания»), а ТОЧКА 2 обозначает действие по глаголу *точить*. А многозначное с точки зрения конвенционального словаря ЯДРО и вовсе представлено единственной статьей.

Из 41 812 статей КС, цитируемых в СинТагРус'е, 4414 имеют индекс. Если сгруппировать их по леммам, то получится 2787 наборов омонимов или многозначных слов. Эти наборы уступают в подробности и точности сведениям из толковых словарей, но, несомненно, могут быть использованы для прицельного поиска. Кроме того, интересно исследовать, насколько такие укрупненные и ориентированные

на контекст значения подходят для решения задач автоматического разрешения неоднозначности.

6. Микросинтаксическая разметка

Под микросинтаксическими единицами (далее — МЕ) понимается широкий класс конструкций, которые находятся на стыке грамматики и лексики и характеризуются высокой степенью идиоматичности (см., например, [Иомдин 2006]). Словарь МЕ на сегодня насчитывает 2132 единицы. Общее количество их вхождений в СинТагРус составляет 18 939. Эти вхождения распределены по 14 836 предложениям: в некоторых предложениях содержится больше одной МЕ. Таким образом, МЕ содержатся примерно в 17 % предложений. Как уже отмечалось выше, информация о МЕ в xml-файл записывается в виде атрибута предложения. Начало и конец единицы задаются с помощью идентификаторов ее первого и последнего слова. Кроме того, указывается имя единицы: по нему ее можно найти в соответствующем словаре.

Следует иметь в виду, что некоторые МЕ совпадают со словами-оборотами типа *что бы то ни было*. В частности, само это слово тоже представляет собой МЕ. Однако эта информация записывается не в теге W, а в теге предложения S в таком виде:

```
<S DATE=»19 03 2021 14:15:34» ID=»181» MICROSYNT=»(что бы то ни
было,{6:что бы то ни было…6:что бы то ни было})&#xA;»>
```

В данном случае МЕ однословна (в понимании СинТагРус'а), но подавляющее большинство МЕ не таковы. Неоднословные с точки зрения СинТагРус'а МЕ имеют внутреннюю синтаксическую структуру, и при поиске по синтаксическим условиям лингвисту может потребоваться исключить из выборки примеры, если они находятся внутри МЕ, или, наоборот, исследовать именно МЕ с точки зрения их внутренней синтаксической организации.

7. Анафорическая разметка

Анафорическая разметка устанавливает связи между местоимениями (3-го лица, *себя, свой, друг друга, тот, который, кто, чей*) и синтаксическими вершинами их антецедентов. Черновой вариант разметки выполняется процессором серии ЭТАП, снабженным специальными правилами. Затем разметка редактируется лингвистом. На сегодняшний день анафорическая разметка охватывает не весь корпус

и недоступна для поиска на сайте НКРЯ. Подробнее об этой разметке см. [Иншакова 2019].

8. Лексико-функциональная разметка

Лексико-функциональная разметка отмечает словосочетания, которые могут быть описаны в терминах лексических функций модели «Смысл-Текст» (см., в частности, [Mel'chuk 1996; Апресян 2011]). Она охватывает весь корпус. Лексико-функциональную разметку можно мыслить как разметку связей (хотя и не синтаксических в изложенном выше понимании): лексико-функциональное отношение связывает слово-аргумент с его значением. Таких связей на сегодняшний день в СинТагРус'е установлено 34 106. Количество предложений с лексико-функциональной разметкой составляет 24 516 (приблизительно 28 % всех предложений), в одном предложении может содержаться больше одной лексико-функциональной связи. Есть примеры, когда слово является аргументом сразу двух различных лексических функций, и примеры, когда значение одной лексической функции является аргументом другой. Рассмотрим предложение

Но я никак не мог ***найти ответ на вопрос****: почему я вынужден здесь* ***придерживаться строгого распорядка*** *дня, когда, пребывая у бабушки, я мог бы вставать, ложиться и развлекаться по собственному усмотрению?*

В этом предложении реализовано 4 лексико-функциональных связи: слово *распорядок* является аргументом лексических функций REAL1-M и MAGN, значениями которых являются слова *придерживаться и строгий* соответственно. А слово *ответ* является аргументом лексической функции OPER1 со значением *найти* и одновременно значением лексической функции S0-REAL3-M для аргумента *вопрос*. Эта информация записана в xml-файле следующим образом:

```
<LF LFARG=»16» LFFUNC=»_REAL1-M» LFVAL=»14»/>
<LF LFARG="16" LFFUNC="_MAGN" LFVAL="15"/>
<LF LFARG="9" LFFUNC="_S0_REAL3-M" LFPREP="8" LFVAL="7"/>
<LF LFARG="7" LFFUNC="_OPER1" LFVAL="6"/>
```

За каждую лексико-функциональную связь отвечает отдельный элемент LF, имеющий атрибуты LFARG (аргумент), LFFUNC (имя функции), LFVAL (значение), LFPREP (вспомогательный предлог). Значениями атрибутов являются идентификаторы слов в пределах предложения, как и во всех остальных типах разметки.

9. Темпоральная разметка

Темпоральная разметка основывается на синтаксической. Каждое темпоральное выражение — это выделенный в синтаксической структуре подграф. Ему приписан набор специальных признаков, описывающих его семантику. Например, в предложении *17 октября пал Таганрог* темпоральным выражением является *17 октября*. Оно относится к событию, обозначенному глаголом *пасть*, и имеет признаки «локализация» и «календарное». Первый признак означает, что это выражение локализует событие во времени, а второй — что оно передает календарную информацию, нужную для логических выводов о времени. Темпоральными выражениями считаются любые выражения, описывающие протекание событий во времени, в том числе те, которые упорядочивают события друг относительно друга — *после войны*. Сведения о темпоральных выражениях, как и сведения о лексико-функциональных связях, хранятся в специальных элементах, обозначаемых как TE.

```
<TE HEAD="1" WORDS="1, 2" DESCRIPTION="локализация|календарн
ое" RELATED_TO="3"/>
```

Атрибут HEAD указывает синтаксическую вершину временного выражения, WORDS содержит перечисление всех входящих в него слов, DESCRIPTION — набор признаков. Значениями атрибутов, как и в других разметках, являются номера-идентификаторы в пределах текущего предложения.

Это самый новый тип разметки СинТагРуса, первые файлы были размечены в ручном режиме только в прошлом году. Очевидно, что даже в таком простом виде эта разметка многое сообщает о событиях. Планируется обогащение схемы разметки для дальнейших исследований семантики событий в тексте.

10. Организация поиска

Описанные выше разметки можно разделить на два больших класса: на те, которые характеризуют слова, и на те, которые характеризуют связи слов. К первому классу будут относиться морфологическая разметка, разметка значений слов, микросинтаксическая разметка и разметка эллипсиса. Микросинтаксическая разметка попадает в этот класс потому, что существенным лингвистическим фактом является включенность или невключенность слова в состав микросин-

таксической единицы. Разметка эллипсиса относится к этому классу потому, что эллипсис понимается как пропуск слов, соответственно, про каждое слово известно, является оно восстановленным пропуском или нет. Остальные типы разметки — синтаксическая, анафорическая, лексико-функциональная, темпоральная — задают связи между словами. Соответственно, для полноценной поисковой работы требуется интерфейс, позволяющий строить запросы из нескольких слов, произвольным образом связанных между собой. Кроме того, данные небольшого опроса, проведенного нами среди пользователей синтаксически размеченных корпусов, показывают, что информация о линейном расположении слов является важным поисковым критерием. Она востребована тремя четвертями пользователей. Поиск по линейным условиям — это прототипический поиск контекста: в любом корпусе без исключения есть возможность искать не только слово само по себе, но и сочетание слов, и в этом случае задаются не только характеристики второго слова, но и расстояние от первого, в пределах которого оно должно встретиться. Можно сказать, что линейное расстояние в этом случае связывает два слова, так что задание линейного контекста и задание ограничений по всем разным типам разметки, предполагающим связь между словами, при поиске функционально эквивалентны.

Разумеется, запросы составляются с целью найти некий объект, но по достижении запросом определенной сложности перестает быть очевидно, какой именно объект находится в фокусе поиска. Когда ищется определенная словоформа или слово, или отдельное синтаксическое отношение, или лексическая функция определенного типа, определить объект не составляет труда. Однако как только в запросе оказывается несколько частей или условий, появляется неоднозначность. Допустим, ищется слово, участвующее в роли хозяина в двух разных синтаксических отношениях, одном повторимом и одном неповторимом. Это может быть глагол, управляющий одним из комплетивных отношений (они относятся к неповторимым, то есть таким, по которым можно иметь не больше одного зависимого) и обстоятельственным отношением (это повторимое отношение). Условию удовлетворяет следующее предложение:

У наших современников есть возможность встретиться с инопланетянами 24 декабря 2013 года, когда они могут прилететь, то есть после последнего их посещения пройдет как раз 3600 лет.

Глагол *встретиться* управляет предлогом с по первому комплетивному отношению, а также является хозяином двух обстоятельственных связей — с числительным *24* и с союзом *когда*. Фокусом запроса лингвиста в данном случае может быть либо слово *встретиться*, которое в указанном предложении встречается 1 раз, либо вся лексико-синтаксическая конфигурация в целом, которая в предложении встречается дважды. Нужен эксплицитный способ, позволяющий обозначить фокус лингвистического интереса. Для этого следует добавить к каждому поисковому условию, задающему связь между словами, возможность использовать ограничитель «есть хотя бы 1». Это различение актуально не только для синтаксически аннотированных корпусов.

11. Заключение

Чтобы сделать доступным все богатство разметки СинТагРуса, необходим интерфейс, который поддерживает поиск сочетаний слов, связанных разными типами связей. Имеющийся в НКРЯ интерфейс, ориентированный на синтаксис, необходимо существенно перестроить. Должен присутствовать также поиск по линейному контексту. Способ задания линейных условий должен быть аналогичен тому, который применяется для задания прочих связей. При этом должна быть учтена специфика выделения словоформ в СинТагРусе. Кроме того, относительно каждого условия, касающегося какой-либо связи, должна быть возможность указать, как его следует понимать — «хотя бы 1 связь» или «строго 1 связь». Микросинтаксическая разметка и разметка эллипсиса должны быть представлены в виде параметров: является ли слово частью микросинтаксической единицы и является ли оно восстановленным пропуском.

Литература

1. *Апресян Ю. Д.* (2011), К новой версии теории лексических функций (ЛФ). Международная конференция, посвященная 50-летию Петербургской типологической школы: Материалы и тезисы докладов, СПб, с. 21–26.
2. *Иншакова Е. С., Иомдин Л. Л., Митюшин Л. Г., Сизов В. Г., Фролова Т. И., Цинман Л. Л.* (2019), СинТагРус сегодня. Труды Института русского языка им. В. В. Виноградова. М., т. 21, с. 14–41. doi: 10.31912/pvrli-2019.21.1
3. *Иншакова Е. С.* (2019), Система разрешения анафоры для русского языка на базе лингвистического процессора ЭТАП-4. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции

«Диалог» (г. Москва, 29 мая — 1 июня 2019 г.). М.: РГГУ, вып. 18(25), с. 239–251. ISSN 2221-7932.
4. *Иомдин Л. Л.* (2006), Многозначные синтаксические фраземы: между лексикой и синтаксисом. Компьютерная лингвистика и интеллектуальные технологии («Диалог-2006»). Труды международной конференции. Бекасово, 31 мая — 4 июня 2006 г. М.: Изд-во РГГУ, с. 202–206.
5. *Mel'chuk I.* (1996), Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In: L. Wanner (ed.), Lexical Functions in Lexicography and Natural Language Processing, Amsterdam/Philadelphia, pp. 37–102.

References

1. *Apresjan Ju. D.* (2011), K novoj versii teorii leksičeskix funkcij (LF) [The theory of lexical functions renewed]. In: Mezhdunarodnaja konferencija, posvjaschennaja 50-letiju Peterburgskoj tipologicheskoj shkoly: Materialy i tezisy dokladov [The international conference dedicated to the 50th anniversary of the typological school in St. Petersburg: Materials and abstracts], St. Petersburg, pp. 21–26.
2. *Inshakova E. S., Iomdin L. L., Mitjushin L. G., Sizov V. G., Frolova T. I., Cinman L. L.* (2019), SinTagRus segodnja [The SynTagRus Today]. In: Trudy Instituta russkogo jazyka im. V. V. Vinogradova [Proceedings of Vinogradov Russian Language Institute]. Vol. 21, Moscow, pp. 14–41. DOI: 10.31912/pvrli-2019.21.1
3. *Inshakova E. S.* (2019). Sistema razreshenija anafory dlja russkogo jazyka na baze lingvisticheskogo protsessora ETAP-4 [An anaphora resolution system for Russian based on ETAP-4 linguistic processor]. In: Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi mezhdunarodnoi konferentsii "Dialog" [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog"], pp. 239–251.
4. *Iomdin L. L.* (2006), Mnogoznachnye sintaksicheskie frazemy: mezhdu leksikoj i sintaksisom [Ambiguous syntactic phrasemes: between lexicon and syntax]. In: Komp′juternaja lingvistika i intellektual′nye texnologii ("Dialog-2006". Trudy meždunarodnoj konferencii. Bekasovo, 31 maja — 4 ijunja 2006) [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog-2006", Bekasovo, May 31 -June 4, 2006]. Moscow: RGGU, pp. 202–206.
5. *Mel'chuk I.* (1996), Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In: L. Wanner (ed.), Lexical Functions in Lexicography and Natural Language Processing, Amsterdam/Philadelphia, pp. 37–102.

Тимошенко Светлана Петровна
Институт проблем передачи информации им. А. А. Харкевича РАН (Россия)
Timoshenko Svetlana
Kharkevich Institute for Information Transmission Problems of the Russian Academy of Sciences (Russia)
E-mail: timoshenko@iitp.ru

Иомдин Леонид Леонидович
Институт проблем передачи информации им. А. А. Харкевича РАН (Россия)
Iomdin Leonid
Kharkevich Institute for Information Transmission Problems of
the Russian Academy of Sciences (Russia)
E-mail: iomdin@gmail.com

Гладилин Сергей Александрович
Институт проблем передачи информации им. А. А. Харкевича РАН (Россия)
Gladilin Sergey
Kharkevich Institute for Information Transmission Problems of
the Russian Academy of Sciences (Russia)
E-mail: gladilin@iitp.ru

Иншакова Евгения Сергеевна
Институт проблем передачи информации им. А. А. Харкевича РАН (Россия)
Inshakova Evgenia
Kharkevich Institute for Information Transmission Problems of
the Russian Academy of Sciences (Russia)
E-mail: e.s.inshakova@gmail.com

В. И. Беликов, А. Д. Верещагина, В. П. Селегей

V. I. Belikov, A. D. Vereshchagina, V. P. Selegey

НА ГРАНИЦАХ ЛЕКСИКОНА: ДИФФЕРЕНЦИАЛЬНЫЕ КОРПУСНЫЕ ИССЛЕДОВАНИЯ ПАРЕМИЙНОГО ФОНДА РЯ

BOUNDARIES OF THE RUSSIAN LEXICON: DIFFERENTIAL CORPUS RESEARCH OF THE PAREMIOLOGICAL RESOURCES

Аннотация. Статья посвящена пограничным составным элементам Лексикона, сочетающим интертекстуальность (как крылатые цитаты) со встроенностью в систему лексических средств языка. Приводятся предложения как по системе различительных признаков, так и категориям, в системе которых существуют паремии. Рассмотрены примеры переводческих проблем, вызванных «неузнанными» паремиями, и возрастные распределения некоторых из них.
Ключевые слова. Паремия, Лексикон, крылатая цитата, языковой корпус, корпусный анализ, интертекст.

Abstract. The paper is devoted to the marginal compound elements of the Lexicon, which are both intertextual (like quotations) and built-in into the Lexical System. Proposals are given both for a system of relevant distinctive features and categories of paroemias. Examples of translation problems caused by "unrecognized" paroemias and age distributions of some of them are considered.
Keywords. Paroemias, Lexicon, language corpora, corpus analysis, intertext.

1. Введение

Работа с объектами, о которых пойдет речь, началась в 2006 г. в ходе лексикографического интернет-проекта «Охота за цитатами». Его инициатором была группа участников переводческих форумов Lingvo, филологов (увы, не лингвистов и лексикографов!), которые сталкивались в работе с составными единицами Лексикона, совершенно игнорируемыми словарями, даже для богатого словарными ресурсами английского языка.

Помимо авторов, идея проекта принадлежала известному переводчику Павлу Палажченко, среди наиболее активных донаторов были также Константин Душенко, составитель словарей цитат, и целый ряд «рефлексирующих» переводчиков, активно сотрудничавших с лексикографами Lingvo.

Основная цель проекта состояла в том, чтобы разобраться с границами, типологией и принципами описания объектов, которые были для простоты условно отождествлены с самым близким термином «крылатые цитаты» (КЦ). В результате примерно двух лет активного существования проекта были собраны и профессионально откоммен-

тированы (прежде всего в отношении проблемы перевода) около двух тысяч русских и английских КЦ. Выявился также ряд принципиальных проблем. Прежде всего, это полное игнорирование обсуждений лингвистами, вследствие чего акцент в обсуждениях постоянно смещался в сторону сомнительной этимологии и авторизации КЦ. Также стало ясно, что продуктивной работе мешает размытость границ между крылатыми и обычными цитатами, между КЦ и пословицами/поговорками.

По этим и другим причинам проект был приостановлен в 2010 г. Авторы статьи приняли решение возобновить его, но уже на новой корпусной технологической платформе. Для этого есть несколько важных оснований, помимо наличия плохо осмысленного, но ценного материала. За это время появился корпус ГИКРЯ с возрастной разметкой объемом в несколько миллиардов слов. Наконец, в 2021 г. стартовал проект ЯГеЛь[1], полностью унаследовавший материалы форумов Лингво, в том числе «Охоты за цитатами», теперь уже в вики-среде.

Термин «крылатая цитата» на практике сбивает с толку, поскольку цитаты в общем случае не являются единицами Лексикона. Мы предлагаем для такого рода языковых объектов термин *лексическая паремия* (ЛП). Полагаем, что такое именование согласуется с традицией, восходящей к Г. Л. Пермякову. Пора вернуться к проблеме идентификации и описания паремий, уже на надежной корпусной основе. Работа с интернет-корпусами требует автоматической разметки/поиска ЛП, что обусловливает необходимость разработки эксплицитных дифференциальных признаков, различные «пучки» которых и должны определять отнесение фразы к той или иной типологической категории.

2. О границах Лексикона и типологии его составных элементов

Систематизация и типизация собранного материала требует уточнения самой системы терминов, действующих в отношении составных элементов Лексикона. Известно много фразеологических «номенклатур», но все они обладают тем неудобным свойством, что последовательное надежное их применение для корпусной разметки (с опорой на имеющиеся определения и пояснения) удается только их разработчикам.

[1] https://int.webcorpora.ru/reg2/index.php/Языки_Городов_и_Людей_(ЯГеЛь).

В современной постановке задача идентификации и сбора лексических паремий требует больших «золотых» корпусов с ручной согласованной разметкой для последующего машинного обучения. Кроме того, нужна разметка релевантными признаками самих интернет-корпусов для поиска паремий. То есть необходимы эксплицитные дифференциальные признаки, различные «пучки» которых и должны определять отнесение фразы к той или иной типологической категории.

Мы исходим из того, что Лексикон является хранилищем готовых (некомпозициональных) смыслов, слов и словосочетаний, из которых по правилам грамматики можно строить более сложные смыслы. При этом он, вероятно, имеет сетевую иерархическую структуру (хотя и неизвестно, как он устроен в мозгу человека). Лексиконы индивидуальны и дифференциальны (имеют персональные и социально мотивированные составляющие), но при оперировании их единицами говорящие, вполне следуя принципу кооперативности, используют их «незакавыченно».

К сожалению, нам неизвестны попытки определить необходимые и достаточные признаки составных единиц Лексикона, отличающие их от просто прецедентных текстов.

Мы предлагаем следующую систему необходимых требований к единицам Лексикона:

1) Позиционируемость: «попадание» в систему родо-видовых, синонимических и прочих отношений, структурирующих Лексикон.
2) Толкуемость: допускают описание в пределах изобразительных возможностей некоторого явно определенного языка толкований значений.
3) Синтаксический статус: имеют модель подключения зависимых и грамматическое значение, обеспечивающее «бесшовное» встраивание в более крупные синтаксические единицы (без признаков цитирования).
4) Некомпозициональность: не могут быть представлены и поняты (во всем объеме значения, подразумеваемого автором) как композиция составляющих ее более простых единиц Лексикона. Тут много градаций, от *вешать лапшу* до *умный в гору не пойдет*.
5) Разделяемость (универсальность): входят в индивидуальные Лексиконы многих носителей языка с достаточно близкими, хотя и не идентичными значениями.

3. Лексические паремии

Несколько примеров объектов, которые находятся в сфере наших интересов:

- *У каждого мертвого будет припарка. У каждой козы — баян, у каждой свиньи по апельсину, у барана — новые ворота* (Вен. Ерофеев о коммунизме);
- *Ничего-то не понимают египтяне в колбасных обрезках*;
- *По лезвию славы.*

Можно сразу определить круг специфических проблем, связанных с такими фразами:

- как распознать (и даже как не попасть впросак);
- как перевести, сохранив оттенки значения;
- как (и где) описывать.

Паремийный статус такая фраза-цитата получает не сразу. Замечательный пример — *usual suspects* (подсказан П. Р. Палажченко) показывает путь от цитаты до единицы Лексикона:

- Исходный контекст: капитан Рено в фильме «Касабланка»: *Round up the usual suspects*;
- Естественный перевод на РЯ: *Знакомые все лица*;
- Дальнейшая фразеологизация: *In addition to all of the usual high-octane suspects, both foreign and domestic, the Russian Cafe also stocks an impressive spread of artisan vodkas in flavors like black currant, lemon, pepper and cranberry*.

Пример из русского языка — вторичная семантизация мнемонической фразы «*наука умеет много гитик*» (предположительно для карточного фокуса). Здесь семантика приобретена «по дороге» употребления (как и в случае знаменитого *абырвалг*):

- *Статистика умеет много гитик* (название статьи С. Е. Бащинского в международном журнале медицинской практики про псевдонаучное использование статистики);
- *Всегда всегда в глубине политик // наука умеет много гитик* («Окнов и Козлов» у Хармса);
- *Обработка изображений — новые идеи или множество гитик Photoshop.*

Особые сложности вызывает перевод ЛП: переводчик должен различать оригинальное употребление и последующее «тиражирование».

Первое употребление переводится буквально, последующие в общем случае — подбором эквивалента в языке перевода. Замечательные примеры ошибок перевода приводят М. Берди и В. К. Ланчиков [Берди, Ланчиков 2006: 20]: *the man was no gift* (*человек этот не подарочек*); *that's from another opera* (*это из другой оперы*); *a dog's death for a dog* (*собаке собачья смерть*) и т. п.; ср. также [Шмелев 2020]. Существенно, что переводческое сообщество достаточно уверенно относило большую часть обсуждаемого к ведению лексикографии, хотя и не находило таких единиц в словарях. Для некоторых типов паремий доступ в словари «принципиально» закрыт, например, для считалок[2], в результате название рассказа и сборника рассказов Т. Толстой «На золотом крыльце сидели…» было переведено на английский как "On the Golden Porch"[3].

Эксплуатироваться может как смысл паремии, так и ее внутренняя форма. Ср. эссе Л. Рубинштейна «Сбоку бантик» о георгиевских ленточках[4].

4. Лексические паремии: лингвистический анализ

Основным отличием ЛП как элементов Лексикона является интертекстуальность в широком смысле: их употребление осознается компетентными носителями языка в связи с некоторым культурно-историческим контекстом.

Изначально термин *паремия* относился только к цитатам из Библии дидактического характера. Позже он стал использоваться для пословиц, загадок, фольклорных изречений, дразнилок и т. п., которые и стали объектами уже вполне почтенной науки паремиологии. Лингвисты часто отказывают таким единицам в независимом лексикографическом статусе, справедливо полагая, что они весьма разнородны, и относя к разным разделам типологии фразеологии лишь часть из них. Многое из отринутого лингвистикой и для фольклористов оказывается на грани интересов (или даже за гранью).

---

[2] Как и для всего «детского»: слова типа *бибикать* в толковых словарях отсутствуют. Слово *считалка* впервые появилось в т. 14 БАСа (1963) и сопровождается пометой *разг.*, хотя это т е р м и н не только в детских играх, но и в фольклористике.

[3] Ср. удачный перевод названия романа Ле Карре "Tinker, Tailor, Soldier, Spy" как «Шпион, выйди вон!» (пер. В. Прахта, 2004).

[4] Рубинштейн Л. С. Словарный запас. М.: Новое издательство, 2008, с. 77–79.

К сожалению, типология фразеологии остается достаточно противоречивой и неполной. Очевидно всем только то, что, в отличие от лексикологии, изучающей отдельные слова, фразеология изучает устойчивые словосочетания. Принято делить их на коллокации (сочетания с несвободным выбором элементов, как *одержать победу*) и фразеологизмы, куда попадают идиомы, поговорки и пословицы и размытый класс речевых клише и фразеосхем [Баранов, Добровольский 2008].

Наш опыт работы в редсовете конференции «Диалог» показывает, что авторы работ по фразеологии и их рецензенты часто не соглашаются друг с другом в отнесении объектов к тому или иному типу. Поэтому важно не просто предложить понятную типологическую систему, но снабдить ее эксплицитными дифференциальными признаками, позволяющими проводить согласованную разметку в значительных объемах.

4.1. Признаки лексических паремий

ЛП как составные элементы Лексикона имеют все вышеуказанные признаки, но с существенными модификациями:

- позиционируемость и толкуемость: для описания ЛП требуется существенное расширение тезаурусной модели Лексикона и языка толкований значений (см. материалы форума «Охота за цитатами»);
- синтаксическая бесшовность: ЛП обычно используются без признаков цитирования, но наряду с этим могут встречаться и в контекстах цитирования (что важно для их автоматического поиска и идентификации);
- скрытая некомпозициональность или семантическая «бесшовность»: очень часто ЛП либо «семантизируются» из контекста, либо имеют релевантное, но неполное прямое композициональное значение;
- употребление ЛП имеет заметные возрастные смещения (до полного отсутствия в некоторых когортах): такие смещения характерны и для других элементов Лексикона, но для ЛП в силу их интертекстуальности они порождают наиболее чувствительные примеры generation gaps;
- деривационная вариативность (наряду с обычной вариативностью).

Особые черты некомпозициональности делают восприятие ЛП максимально сложным для неподготовленного читателя, незнакомого с первоисточником ЛП.

В отличие от цитат и других паремий (пословиц и т. п.), ЛП хоть и безусловно идиоматичны (стилистически отмечены), но ни в коей мере не претендуют на оригинальность, дидактичность. Они отличаются и упомянутой выше бесшовностью, и относительной простотой семантики (легкость перифразирования, наличие непаремийного синонима).

При этом, в отличие от цитат, ЛП не могут в общем случае переводиться напрямую (ср., например, *Все смешалось в доме Ивановых* или *Мы все женились понемногу*).

И, к сожалению, ЛП решительно отсутствуют в словарях (точнее, частично представлены в самых разных).

Вопрос толкования ЛП мы оставляем за пределами данной статьи. Прототипы толкований можно найти на сайте проекта ЯГеЛь в разделе «Охота за цитатами». В качестве ценного сравнительного материала, который необходимо учитывать, мы рассматриваем Тезаурус идиоматики [Словарь-тезаурус 2008].

5. Технологии поиска и верификации паремий

Пограничные объекты можно пытаться исследовать методами проекта «Охота за цитатами» или других проектов вики-словарей, создаваемых сегодня дешевыми и быстрыми методами краудсорсинга (типа Memepedia.ru или Wikireality.ru, для английского — Urban Dictionary). Результаты такого описания, к сожалению, сложно назвать надежными и адекватными с лексикографической точки зрения.

Одна из причин состоит в том, что, как уже говорилось выше, значительная часть лексических паремий является наиболее быстро меняющейся частью Лексикона. Культурные межпоколенческие различия сказываются здесь заметнее всего. Характерны для нее, разумеется, и другие социолингвистические параметры вариативности. Очевидно, что объективный анализ возрастных, жанровых и гендерных смещений возможен только на больших размеченных корпусных данных.

Насколько большими должны быть «паремиологические» корпуса? Наш выборочный анализ показывает, что частота идиом и лексических паремий в текстах отличается в среднем на порядок. Например,

при грубой оценке книги «Просто Рим. Образы Италии XXI» Аркадия Ипполитова (автора, склонного к идиоматичному письму) мы насчитали примерно одну идиому на 1000 слов и одну паремию на 10 тыс. слов (разумеется, в интертекстуальной поэзии Бродского будет иное соотношение).

Поэтому, например, на корпусе в 10 млрд слов абсолютные частоты паремий позволяют исследовать только ядро паремийного фонда в несколько сотен единиц, а то и менее.

С другой стороны, адекватная стилометрия безусловно должна учитывать концентрацию и типологию паремий в текстах [Баранов, Добровольский 2021]. Для этого совершенно недостаточно специализированных корпусов с ручной разметкой. Но такие корпуса могут с успехом использоваться для машинного обучения. Это позволит перейти к фразеологической разметке больших корпусов, представляющих разные сегменты текстов.

В целом, автоматическая идентификация лексических паремий должна основываться прежде всего на синтаксической бесшовности при наличии прямых цитирований. Эта работа еще предстоит на корпусе ГИКРЯ 2.0. Большим подспорьем будут и имеющиеся open-source словники, для которых важна только полнота исходного списка кандидатов на роль ЛП.

В данной статье мы рассмотрим некоторые результаты корпусного анализа возрастного распределения паремий в корпусе ГИКРЯ, объем и разметка которого позволяют проводить такие исследования.

6. Примеры возрастных смещений по корпусу соц. сети ВКонтакте на базе ГИКРЯ

Для наилучшей иллюстративности результаты возрастной статистики были разбиты на возрастные когорты, основанные на годе рождения автора материала. Так, «Вхожд» — это абсолютное число вхождений, а «Норм» — это нормированные значения, вычисленные нормированием результата выдачи на общую статистику корпуса. По каждой когорте чи́сла вхождений суммировались, после чего делились на суммированные значения общего числа вхождений за авторством представителей этой же самой когорты в статистике всего корпуса ВК. Более объективной оценкой для нас являются именно нормированные числа вхождений. Рассмотрим динамику на примере графиков, полученных из вышеприведенных данных.

Использование фразы из поэмы А. С. Пушкина «Медный всадник» демонстрирует нисходящий тренд. Возможно, это связано с более индифферентным отношением молодежи к школьному курсу литературы.

Лозунг «Россия для русских» всегда был основой для публичной доктрины русского национализма. Во второй половине 2000-х годов стала популярна субкультура эмо, в связи с чем интернет быстро переделал лозунг в созвучный ему «Россия для грустных», а идеологию — в «эмоционал-социализм». Тогда паремия получила свою популярность и используется по сей день.

Фраза из популярного сериала «Следствие вели…» ассоциируется с Леонидом Каневским, для старшей когорты — великим актером из любимого сериала, а для молодежи — «Тем самым усатым мужичком из мема». Таким образом, интертекст знаком и тем, и другим, что делает возрастное распределение относительно равномерным.

Рис. 1. **Нисходящий тренд**

Еще несколько примеров с нормированным числом вхождений приведены в таблице ниже. Разумеется, важны не сами цифры, а их соотношение: в двух первых строках в старшей когорте оно составляет 1 : 2, а в младшей — 5 : 1.

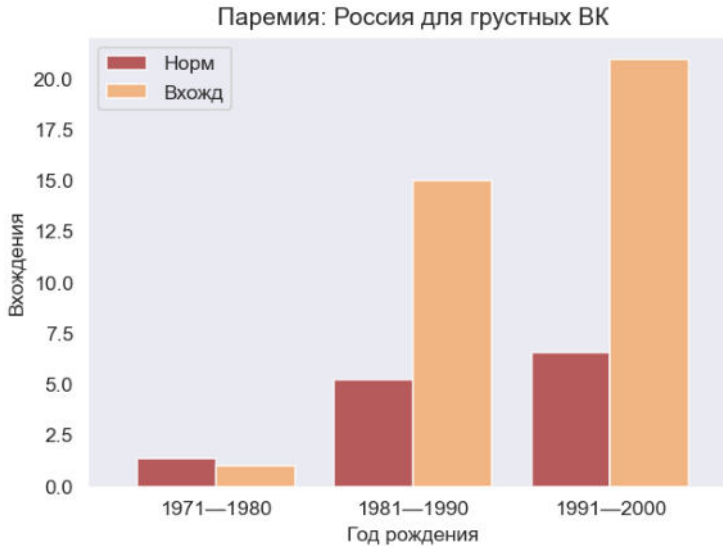


Рис. 2. **Восходящий тренд**

Рис. 3. **Отсутствие выраженного тренда**

Возрастное распределение паремий

	1971–1980	1981–1990	1991–2000
Шах и мат, господа атеисты	11,1	17,8	26,9
Надо, Федя, надо	22,2	7,3	5,3
Пруфов не будет	4,2	8	12,2
Все в сад	233	144	22,9
Есть два стула	5,5	11,5	13,8
Россия, которую мы потеряли	48,5	18,1	4,3

7. Выводы и задачи на будущее

Данная работа является заведомо постановочной. Мы только наметили основные направления исследований, основываясь на собранном проектном материале, некоторых идеях и первых результатах дифференциальных корпусных исследований лексических паремий.

Мы полагаем, что поиск, анализ и словарное описание лексических паремий очень важны как для прикладных областей типа перевода и лингвопедагогики, так и для экспликации представлений о составе и структуре Лексикона.

Команда ГИКРЯ приглашает коллег участвовать в этих исследованиях на материале новой версии корпуса 2.0.

Литература

1. *Берди М., Ланчиков В. К.* (2006), Успех и успешность. Русская классика в переводах Р. Пивера и Л. Волохонской. Мосты, № 1(9).
2. *Баранов А. Н., Добровольский Д. О.* (2021), Об одном подходе к количественной оценке идиоматичности текста как характеристике авторского стиля. Компьютерная лингвистика и интеллектуальные технологии: По материалам международной конференции «Диалог». Вып. 20, М.: РГГУ, с. 58–67.
3. *Баранов А. Н., Добровольский Д. О.* (2008), Аспекты теории фразеологии. М.: Знак.
4. *Пермяков Г. Л.* (1988), Основы структурной паремиологии. М.: Наука.
5. Словарь-тезаурус современной русской идиоматики (2008), под ред. *А. Н. Баранова, Д. О. Добровольского*. М.: Мир энциклопедий Аванта+.
6. *Шмелев А. Д.* (2020), Паремии, используемые в прозе Солженицына, и проблемы их перевода. Шаги/Steps, т. 6, № 3.

References

1. Berdi M., Lanchikov V.K. (2006), Uspekh i uspeshnost'. Russkaia klassi-ka v perevodakh R. Pivera i L. Volokhonskoi. Moscow.
2. Baranov A.N., Dobrovol'skii D.O. (2021), Ob odnom podkhode k koli-chestvennoi otsenke idiomatichnosti teksta kak kharakteristike avtorskogo stilia. *Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog»*. iss. 20, Moscow, pp. 58–67.
3. Baranov A.N., Dobrovol'skii D.O. (2008), Aspekty teorii frazeologii. Moscow.
4. Permiakov G.L. (1988), Osnovy strukturnoi paremiologii. Moscow.
5. Slovar'-tezaurus sovremennoi russkoi idiomatiki (2008), Ed. A.N. Baranov, D.O. Dobrovol'skii. Moscow.
6. Shmelev A.D. (2020), Paremii, ispol'zuemye v proze Solzhenitsyna, i problemy ikh perevoda. Moscow.

Беликов Владимир Иванович
д. филол. наук., свободный художник (Москва)
Vladimir Belikov
PhD, retired (Moscow)
E-mail: vibelikov@gmail.com

Селегей Владимир Павлович
директор по лингвистическим исследованиям компании ABBYY (Москва)
Vladimir Selegey
ABBYY, Head of Advanced Research Department (Moscow)
E-mail: vladimir.selegey@abbyy.com

Верещагина Анна Дмитриевна
магистрант ФиКЛ РГГУ (Москва)
Vereshchagina Anna
master candidate in Computer Linguistics, RSUH (Moscow)
E-mail: lorelei.aether@gmail.com

СЕКЦИОННЫЕ ДОКЛАДЫ

SECTION TALKS

L. Abalo-Dieste, J. Pérez-Guerra

PASSIVISATION AND RELATIVISATION AS COLLOQUIALISATION STRATEGIES IN PRESENT-DAY ENGLISH: A CORPUS-BASED STUDY[1]

Abstract. This corpus-based study aims to determine whether colloquialisation is a process affecting written English or a general tendency of the English language that can also be attested in speech. To that end, this study investigates two syntactic clausal processes evincing variation in twenty-first century English due to a process of colloquialisation: passivisation and relativisation. The first phenomenon is explored through the overall productivity of passive predicates, and of *be*- and *get*-passive constructions. The second phenomenon is investigated by analysing choices between, on the one hand, *who* and *whom*, and *which* versus *that*/zero relativisers. The data are collected from the Spoken BNC2014 and the BNC1994DS corpora, as well as from the F-LOB and the BE06 corpora. The findings support the hypothesis that passivisation and relativisation are determinants of colloquialisation both in spoken and in written English.
Keywords. Colloquialisation, written, spoken, passive, relative.

1. Introduction

Traditional distinctions between speech and writing have been blurred due to the demands of present-day communication. The growing proximity between writing and speech is called 'colloquialisation' and constitutes the topic of this study. Recent linguistic research on colloquialisation has evinced a growing degree of orality and simplicity techniques in written English in all varieties and text types. Specifically, this paper investigates (potentially) representative linguistic features of colloquialisation in spoken Present-Day British English (BrE) with the aim of determining whether col-

[1] We would like to acknowledge and thank the Spanish State Research Agency and the European Regional Development Fund (grant no. FFI2016-77018-P), and the Regional Government of Galicia (grant no. ED431C 2017/50) for their generous financial support.

loquialisation is a phenomenon specific of written English or a tendency affecting the English language as a whole, which, in consequence, is also shown in spoken English.

This study is organised as follows. Section 2 summarises previous work on colloquialisation. Section 3 describes the purpose of the current corpus-based research, followed by a description of the methods used to collect, retrieve and analyse the data. Section 4 reports the empirical results and discusses the main findings. Finally, Section 5 contains the summary and the concluding remarks, with a focus on the contribution of this chapter to the study of colloquialisation.

2. Colloquialisation: An overview

This section summarises previous studies on colloquialisation. Colloquialisation is defined by [Iosef 2013: v] as "the growing influence of speech on written language". In studies like [Collins, Yao 2013: 480], this process of language change is associated with the influence exerted by speech on the written medium, speech being regarded as a combination of features typical of face-to-face conversation and characteristics of written language. [Iosef 2013: 133] argues that this process of growing orality does not only involve informality in writing, but also the incorporation of "more direct speech" in written texts. [Collins, Yao 2013: 480] contend that colloquialisation spreads not only to written genres but to spoken language as well, which justifies this investigation.

Linguistic features associated with colloquialisation have been investigated through corpus analyses. In [Collins, Yao 2013] the effect of colloquialisation in non-native varieties of English is assessed by exploring data from ICE corpora through features such as contractions, semi-modals, *get*-passives and the contraction *let's*. Another investigation measuring the significance of colloquialisation in English is [Iosef 2013]. This author analyses contractions, quotative *like*, and phrasal verbs with *up* and *out*. [Collins 2008] and [Levin 2013] carry out descriptive studies of the distribution of specific linguistic features, such as modal auxiliaries and semi-modals. [Baker 2017: 237–238]. compares BrE and American English (AmE) by paying attention to spelling, vocabulary and grammar differences, and takes colloquialisation as an explanatory theory of the results obtained. Regarding colloquialisation, he investigated constructions with empty subjects *it/there*, past participles (as representative of passive constructions), modal auxiliaries and relative pronouns. [Leech et al. 2009: 19] investigate the Brown

family of corpora and check the role of prescriptive grammatical approaches to language description. Their study reveals a decrease in the frequency of modals and passives, and an increase of progressives and contractions.

3. The case study

This section deals with the purpose, methodology and data retrieval. In Section 3.1 I describe the aim of this research. Section 3.2 tackles the research design and method alongside the criteria for the selection of the linguistic features. In Section 3.3 I present the specific search queries and their limitations.

3.1. Aims and scope

A corpus-based analysis has been conducted to explore linguistic features associated with the phenomenon of colloquialisation in the recently released spoken part of the British National Corpus (BNC) 2014, alongside in its predecessor the British National Corpus. The diachronic study of spoken English is compared with the results on written English provided by [Leech et al. 2009] and [Baker 2017: 175] in an attempt to determine whether colloquialisation is a process that also affects spoken BrE or not.

3.2. Data and methodology

This investigation is based on data from the demographically-sampled (DS) part of the original BNC spoken part (BNC1994DS) and the spoken section of the BNC2014 [Love et al. 2017: 324]. The data have been retrieved through the Lancaster University's CQPweb platform [Hardie 2012].

Drawing on [Leech et al. 2009] and [Baker 2017: 176], relativisation and passivisation as investigated here as potential strategies of colloquialisation in spoken BrE. As regards passivisation, I have explored *be*- and *get*-passives. Although *be*-passives outnumber *get*-passives, [Leech et al. 2009: 244] showed that (informal) *get*-passives were gaining ground, particularly in text types closely related to conversation, whereas *be*-passives have become less frequent. Since *be*-passives are considered formal passive constructions which convey objectivity in writing, their declining tendency strengthens the hypothesis of growing colloquiality in BrE in favour of active voice and informal passive variants such as *get*-passives. From this perspective, this paper investigates not only the frequency of *be*- and *get*-passives in spoken English, but also the overall decline of passive voice as a feature not evincing colloquialisation. Relativisation phenomena have been approached in

the literature by exploring the distribution of *wh-, that-* and zero-relative clauses. [Leech et al. 2009: 288] take the declining trend of *wh*-relative clauses to be a sign against colloquialisation. Specifically, the frequency of the *who* relativiser decreases alongside other *wh*-forms in BrE, and increases in AmE. [Baker 2017: 167] also reports the decline of *wh*-relative clauses, particularly *which*-clauses, in written BrE, in favour of their more colloquial and growing *that*-relative counterparts. [Leech et al. 2009: 231] found zero-relative clauses to be increasing and, thus, competing with *that-* and *which*-relativisers, while [Baker 2017: 168] reports decreasing frequencies of this relativisation type since 1961. In this paper I investigate the variation of the different relativisation options in an attempt to discern their influence on colloquialisation in speech.

The spoken data have been retrieved from the two 1994 and 2014 BNC corpora. Previous research results have been taken as the benchmark for written BrE. When neither [Baker 2017: 151–175] nor [Leech et al. 2009] provides frequencies for a specific feature, the feature's frequency is retrieved from the F-LOB (BrE 1991) and the BE06 (BrE 2006) corpora of the Brown family.

The retrieval of the relevant constructions was carried through the CQPweb platform provided by Lancaster University. The F-LOB, BE06, BNC1994DS and Spoken BNC2014 corpora have been annotated with different CLAWS tag-sets, which required the design of specific queries for each corpus. Regarding relative clauses, queries were kept as similar as possible in order to maximise comparability (e.g., **[_AT (_{A})?_{N} (_{N})? which (I|you|he|she| it|we|they)]** for *which-, that-* and zero-relative clauses, which retrieves instances of an article followed by an optional adjective, followed by a noun, followed by an optional noun, followed by the relativiser *which,* followed by a personal nominative pronoun, likely to function as the subject of the relative clause, as in (1):

(1) apart from the different business culture which you've mentioned there's also … (Spoken BNC2014, SP2Y 2707)

A comparable query for *who* and *whom* required a complex query through CQP syntax. This query was designed to avoid prepositions before *whom* and attributive complements after *who*: **[!(pos= 'II|IO|IW')] 'who' ('I|you|he|she|it|we|they') [!(pos= 'VB.*')]**. It retrieved instances of *who* not preceded by a preposition and followed by a personal pronoun not followed by any *be* verbal form. Passive search strings were designed to grant comparison. The query **[pos= 'VB.*'] [pos=**

'XX|R.*']? [(pos= 'VVN') & !(word= 'used|married')] detected forms of the verb *be*, followed by an optional negative particle (*not/n't*) or an adverb, followed by a past participle, excluding *used* and *married*. The query for *get*-passives is practically identical, with a '**get|gets|got**' slot.

4. Results and discussion

This section is devoted to the analysis of relativisation and passivisation. Passivisation has been explored through the distribution of *be*- and *get*-passive constructions in the corpora. The raw and normalised ('Fpw', per 1,000,000 words) frequencies of the constructions are given in Table 1.

Table 1. **Passivisation**

	Spoken				Written			
	BNC1994DS		Spoken BNC2014		F-LOB		BE06	
	Raw	Fpw	Raw	Fpw	Raw	Fpw	Raw	Fpw
Be-passives	30 737	2565.02	22 921	2006.63	11 165	9768.51	9754	8503.20
Get-passives	2594	216.47	4442	388.87	79	69.11	80	69.74
Active voice	2 253 002	188 014.6	2 538 495	222 234	162 912	142 535.4	165809	144546.6

In light of the normalised frequencies in Table 1, we hypothesised a number of tendencies: (i) greater frequency of passive constructions in written English (see Figure 2), (ii) decline in the use of the passive voice from the twentieth to the twenty-first century, and (iii) greater frequency of *get*-passives in spoken English (see Figure 1).

In order to test these hypotheses, I compared the frequencies of active versus passive sentences, and *be*- versus *get*-passives, in twentieth and twenty-first written and spoken BrE. The statistical tests reveal highly significant differences between passive and active constructions in written versus spoken twentieth-century English (F-LOB vs. BNC1994DS; $\chi^2 = 22729.82$; $p < .00001$) and in twenty-first-century English (BE06 vs. Spoken BNC2014; $\chi^2 = 25231.94$; $p < .00001$). These results confirm the preference for passive voice in writing, which reflects the association of this construction to formal language. Regarding the productivity of passive voice over time, the difference between active and passive constructions was tested in twentieth- versus twenty-first-century written ($\chi^2 = 113.53$; $p < .00001$) and spoken ($\chi^2 = 1499.53$; $p < .00001$) English. The statistically significant decrease of this

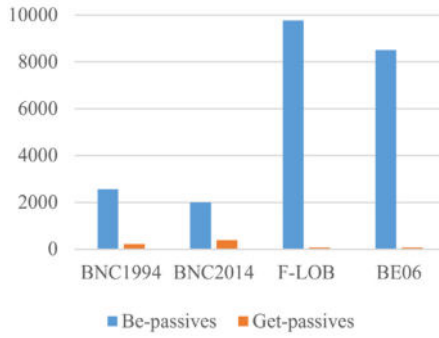


Fig. 1. ***Be-*** **and** ***get*****-passives**

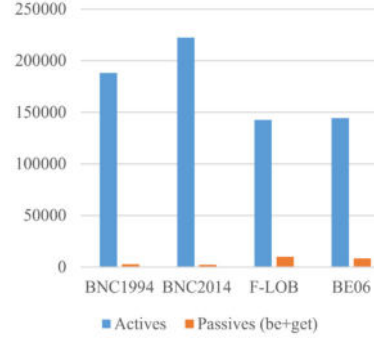


Fig. 2. **Passivisation frequencies**

feature indicates a movement away from complex constructions and towards colloquialisation. A closer look at the two variants reveals a highly significant difference between *be-* and *get*-passive constructions in written versus spoken twentieth-century English (F-LOB vs. BNC1994DS; $\chi^2 = 747.63$; $p < .00001$) and twenty-first-century English (BE06 vs. Spoken BNC2014; $\chi^2 = 1610.76$; $p < .00001$). Although the prototypical *be*-passive is still the most frequent structure, *get*-passives seem to be gaining ground as informal alternatives in speech. These findings evince that passivisation is a determinant of colloquialisation in English, not only in writing but also in speech.

Regarding relativisation, my analysis focuses on two choices between relativisation devices which differ in style: *who-* vs. *whom*-relativisers, and *which-* vs. *that-*/zero-relativisers. Attention was paid to the distribution of *who-* and *whom*-relativisers in linguistic that allow for alternation with no semantic consequences. Table 2 and, more visually, Figure 3 provide the frequencies, which reflect the greater frequency of *whom* in written English, and the decrease in the use of *whom* from twentieth- to twenty-first-century English.

Table 2. **Relativisation:** ***who/whom***

	Spoken				Written			
	BNC1994DS		Spoken BNC2014		F-LOB		BE06	
	Raw	Fpw	Raw	Fpw	Raw	Fpw	Raw	Fpw
Who	450	37.553	654	57.255	10	8.749	19	16.564
Whom	50	4.173	4	0.35	35	30.622	10	8.718

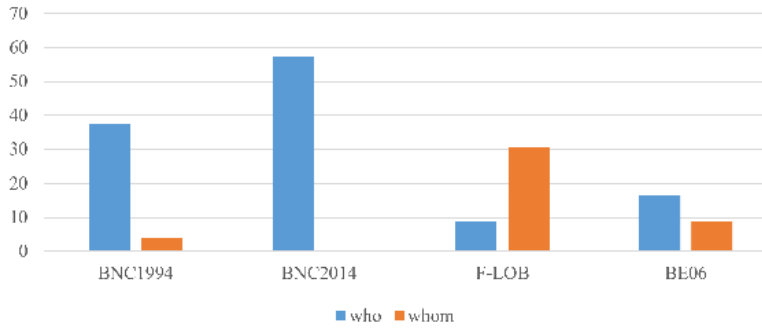


Fig. 3. **Relativisation: *who/whom***

The tests reveal significant differences for written ($\chi^2 = 13.87$; $p < .00019$) and spoken (Fisher; $p < .00001$) English. Also, differences between the use of *who* and *whom* in spoken versus written twentieth- ($\chi^2 = 114.07$; $p < .00001$) and twenty-first-century English ($\chi^2 = 159.66$; $p < .00001$) proved significant. *Who* is more productive than *whom* in both speech and writing, and this trend becomes more salient over time. The data also reflect a greater frequency of *whom* in written English, thus evincing a connection between *whom* and textual formality, and between *who* and speech. As suggested in the literature, these findings reveal that *who* vs. *whom* variation can be taken as a colloquialisation strategy in written and spoken English. The second stylistic choice affecting relativisation contrasted *which*- with *that*- and zero-relative clauses. Following the competition between *which*- and *that*-clauses suggested in the literature, I combined the frequencies of these two relativisers, both considered informal alternatives to *which*-clauses. Table 3 and Figure 4 reveal, first, a higher frequency of *that*- and zero-relative clauses in spoken English and, second, an increase of these relativisers from twentieth- to twenty-first-century spoken English and decrease in written English.

Table 3. **Relativisation: *which, that* and zero**

	Spoken				Written			
	BNC1994DS		Spoken BNC2014		F-LOB		BE06	
	Raw	Fpw	Raw	Fpw	Raw	Fpw	Raw	Fpw
Which	756	63.08	212	18.56	41	35.87	17	14.82
That	5869	489.77	4050	354.56	228	199.48	179	156.04
Zero	25 276	2109.3	30 937	2708.39	1064	930.91	1059	923.2

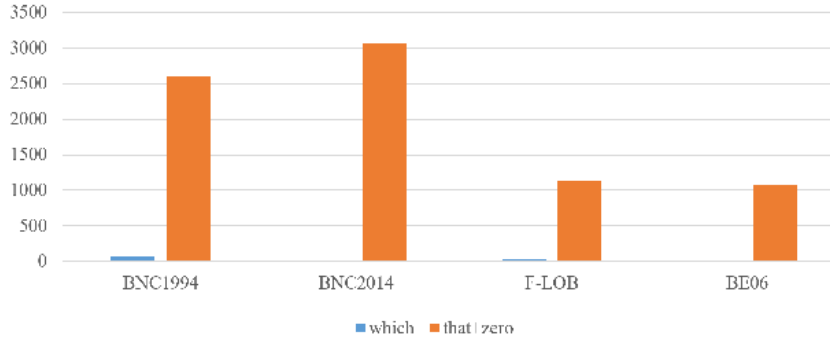


Fig. 4. **Relativisation: *which, that* and zero**

The analysis of relativisers showed that the difference between *which*- and *that*-/zero-relative clauses is not significant in twentieth-century written English ($\chi^2 = 2.72$; $p = .098833$). By contrast, this choice is statistically significant in twentieth- and twenty-first-century spoken English ($\chi^2 = 10.99$; $p < .000918$). This suggests that the alternation *which* versus *that*/zero is a relevant to colloquialisation in spoken English in both centuries, while in written English it is only significant as a colloquialisation strategy in the twenty-first century.

4. Conclusion

This study has explored two potential colloquialisation strategies in spoken BrE: passivisation and relativisation with the purpose of determining whether the same trends are observed or not in written and in spoken English. In light of data retrieved from twentieth- and twenty-first-century spoken and written English, this study corroborated the connection of passive voice with textual formality and its decline in present time, which contributes to the colloquialisation process. Secondly, the choice between the relativisation options (*who* vs. *whom*, *wh-* vs. *that*-/zero relativisers) proved to be a marker of colloquialisation in writing and speech. These results not only support the colloquialisation hypothesis in spoken BrE, but also suggest that this phenomenon started first in spoken English.

References

1. *Baker P.* (2017), American and British English: Divided by a Common Language? Cambridge: Cambridge University Press.

2. *Collins P.* (2008), The English Modals and Semi-modals: Regional and Stylistic Variation. In: T. Nevalainen, I. Taavitsainen, P. Pahta, M. Korhonen (eds.). The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present. Amsterdam: John Benjamins, pp. 129–145.
3. *Collins P., Yao X.* (2013), Colloquial features in World Englishes. In: International Journal of Corpus Linguistics. Vol. 18(4), pp. 479–505.
4. *Hardie A.* (2012), CQPweb: Combining Power, Flexibility and Usability in a Corpus Analysis Yool. In: International Journal of Corpus Linguistics. Vol. 17(3), pp. 380–409.
5. *Iosef M.* (2013), Signs of Colloquialization: Three Corpus-based Case Studies. Master's Thesis. Oslo: Universitetet i Oslo.
6. *Leech G., Hundt M., Mair Ch., Smith N.* (2009), Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press.
7. *Levin M.* (2013), The Progressive Verb in Modern American English. In: B. Aarts, J. Close, G. Leech, S. Wallis (eds.). The Verb Phrase in English: Investigating Recent Language Change with Corpora. Cambridge: Cambridge University Press, pp. 187–216.
8. *Love R., Dembry C., Hardie A., Brezina V., McEnery T.* (2017), The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations. In: International Journal of Corpus Linguistics. Vol. 22(3), pp. 319–344.

Laura Abalo-Dieste
University de Vigo (Spain)
E-mail: laura2abalo@gmail.com

Javier Pérez-Guerra
University de Vigo (Spain)
E-mail: jperez@uvigo.es

L. N. Beliaeva, O. N. Kamshilova

MT RESULTS AND PARALLEL SCIENTIFIC TEXT CORPORA FOR LEXICOGRAPHY

Abstract. The paper suggests that the use of full-text parallel corpora for lexicographic and terminographic aims may turn more effective, provided they have a "built-in" corpus of MT results, since analysis and comparison of these corpora will make it possible to identify lexical units that are to be considered as dictionary entries. A corpus of MT results, we assume, can store the history of NP transformations and modifications in the corpus. The complicated structure of multicomponent terminological NPs, their variants and modifications within the same text determine the need for a three-part text corpus, including parallel/comparable texts and their MT translation.
Keywords. Lexicography, terminology, parallel and comparable corpora, MT, multicomponent NPs

1. Introduction: Applied Lexicography and Parallel Corpora

The aim of the paper is to propose a way to optimize the use of parallel and comparable corpora as lexicographic resources by means of MT procedures and results. Being a major branch of applied linguistics, applied lexicography traditionally aims at building and updating subject oriented databases and automated/automatic dictionaries, specifically terminological ones. The level and reliability of information/knowledge extracted from texts of various composition, structure and destination is determined by the completeness and adequacy of lexicographic systems used for the purpose.

It has become almost commonplace, that much of lexicographic (terminological) job today is based on text corpora, that provide a reliable database for dealing not only with research issues, but with practical lexicographic tasks as well, such as terms identification and extraction, translation, etc. [Beliaeva 2009, 2014; Delpech, Daille 2010; Heja 2010; TTC Project[1]]. Parallel and comparable text corpora are effectively used for creating multilingual lexicons and concordances.

In this paper we dare to suggest that, if we use full-text parallel or comparable corpora as a lexicographic base, it is necessary to expand them with a corpus of MT results. Analysis and comparison of these corpora will make it possible to identify lexical units as candidates for special dictionary entries [*cf.*: Delgado et al. 2002; David, Curran 2007; Lavie et al. 2008]. The main difficulty in this identification process is to establish the boundaries and structures of these lexical units. Whenever we deal with a scientific text, we have to admit that the lexical units in question are multiword termino-

[1] http://www.ttc-project.eu/about-ttc/concept-and-objectives

logical constructions. Scientific texts abandon in simple (without preposition) noun phrases (NPs) which are usually multiword units with a number of attributive elements modifying the head noun. They form one syntactic group with its head element and have a common syntactic function in the sentence.

NPs are objects of study in both theoretical and applied aspects [Baroni, Zamparelli 2010; Bergsma, Wang 2007]. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e., they are definitely units of syntax, not lexicon. The NP dependency structure has always been a major issue for MT or human translation of scientific texts (*cf.*: [Feldman, Dagan 1995; Babych, B., Hartley 2002; Shen et al. 2008; Reiter, Frank 2010]), especially when translating from English to any inflectional language. We assume that the internal structure of an NP correlates with the internal dependencies structure of the appropriate sentence. The point is to find a procedure to recognize this structure in a convolution.

The paper focuses on multicomponent terminological NPs in scientific texts, their structure and transformation to find a procedure for matchings in parallel and comparable texts.

2. Tracking Noun Groups in a Scientific Text

An NP in its simplest form consists of a determiner and a head noun or pronoun. It may consist of one word or include a number of embedded premodifiers that make it a complex unit. NPs with a number of premodifiers are called simple if they include no preposition, no matter how many premodifiers they have [Malakhovskaya et al. 2021]. As terminological NPs actually represent a sentence compression (convolution), their internal structure must correlate with the internal structure of a corresponding sentence, thereby revealing the syntactic dependencies. Thus, finding a procedure to recognize this structure in a concise form of an NP becomes a key problem.

Since an NP is a sentence convolution, the markers of relations between its actual components normally show in inflectional languages, while a simple English NP hardly has any, except for *-'s*. The compression of sentence structure, the external simplification of both structure and form of English NPs causes semantic complication.

Our corpus findings in parallel and comparable scientific text corpora of different subject areas (medicine, space systems, seismic isolation, power plants construction, machine translation, language teaching), built for re-

search and practical translation and lexicographic aims, prove that most frequent in English are 2-component combinations with a head noun, which three times exceed three-element combinations, second frequent combinations in scientific and technical texts (see Table 1).

Table 1. **Frequency of English NP Length in Scientific Texts (Subject Area "Seismic Protection"[2])**

No	NP length	Number of different models	NP frequency
1	2	1516	3457
2	3	674	1053
3	4	207	380
4	5	51	61
5	6	20	164
6	7	2	2
7	8	5	5
Total		2475	5122

In spite of their model variety, 2-component NPs prove to exceed other multicomponent NPs in scientific texts across different subject areas. Table 2 illustrates the distribution of 2-, 3- and 4-component NPs in a special comparable corpus on web and linguistic technologies:

However, the external simplicity of the most frequent English NP structures is misleading due to the fact that this simplicity could be the result of another NP or a sentence compression, which, as it was above mentioned, leads to its semantic complication.

Pursuant to this, NPs formation in a text is either merging NPs and single (independent) lexical units in a new, more complicated nominative construction, or condensing multicomponent NPs at the expense of deleting the units which are implicitly obvious.

[2] The "Seismic Protection" corpus a 1-million-word partly parallel corpus, the size of English and Russian parts is 500,000 tokens each (19,145 English and 25,872 Russian wordforms respectively).

Table 2. **Models of English NPs in Scientific Texts (Subject Area "Web and Linguistic Technologies"[3])**

Model number	Model	Length	Frequency	Number of NPs
1	A+N	2	1474	748
2	A+PII+N	3	9	6
3	N1+N2	2	1407	530
4	N1+N2+N3	3	248	128
5	A/N1+N2	2	71	47
6	N1+G/N2	2	24	18
7	A1+A2/N1+N2	3	10	10
8	A+G/N2 +N2	3	7	4
9	A1+A2+N	3	151	104
10	A+N1+N2	3	292	172
11	PII+A+N	3	25	20
12	PII+N	2	170	73
13	A1+N1+A2/N2+N3	4	3	2
14	A1+C+A2+N	4	15	9
15	A1+A2/N1+N2+N3	4	6	7

In accordance with this we find two ways of building new NPs in a real text: either by adding a lexeme to a previously used or standard NP, thus producing a novel, more complicated nomination: *machine translation => machine translation method*, *machine translation service, machine translation program*, or by deleting implicitly obvious units, thus condensing the sentence structure to a multicomponent NP: *syntactic dependency, syntactic formalism, syntactic dependency tree annotation => dependency annotation formalism*

The first way is a step-by-step process of gradual transformation (complication), adding specific characteristics to the head element, while the second presents a transformation process of sequential convolution which is successively realized on three levels:

[3] The "Web and Linguistic Technologies" corpus is a comparable corpus, the English part including 372 English texts (3,468,000 tokens).

Level 1. Transfer from a complex NP (with a preposition) to a simple one by inversion of its elements: *phrase-structure trees from dependency annotations => syntactic phrase-structure dependency trees annotation.*

Level 2. Elimination of duplicate components in the new NP: *syntactic phrase-structure dependency trees annotation => syntactic dependency tree annotation.*

Level 3. Elimination of duplicate semes: *syntactic dependency tree annotation => dependency tree annotation.*

Comparison of 2- and multicomponent NPs within a text gives evidence of 2-component NPs being the source for longer constructions. When two 2-component NPs result in a multicomponent structure within the boundaries of one text or texts of the same subject area, one can establish a number of patterns for the resulting combination:

- a 4-component NP as a result of merging two NPs of A + N type, embedding NP1 as an attribute before the head of NP2:

  seismic analysis+ indirect method => indirect seismic analysis method
  second language + adult learner => adult second language learner

- a 3-component NP, when two initial NPs have a common component:

  mental processing + processing operation => mental processing oper-ation

- a 3-component NP, when one component in NP1 is semantically supported by a component in NP2:

  communicative method + language learning => communicative lan-guage learning

- a 3-component NP, when the semantics of the resulting NP is determined by the domain extralinguistic knowledge, for example:

  seismic stability + direct analysis => seismic stability direct analysis,
  which in the text may be convoluted up to a 3-component NP *seismic direct analysis.*

3. Translating multicomponent NPs as a part of lexicographic analysis

NP standard transformations described in section 2 do not show all possible variants of NP development in a text. However, they might be helpful when translating an NP with a high degree of structure com-pression. As the research shows, it is exactly 2-component NPs that pre-sent particular difficulties in their analysis and translation.

To overcome the difficulties, we find only two approaches which can be used both in MT and human translation.

The first approach includes modelling the knowledge base of the domain in question (within the framework of the MT system) or appeal-ing to the translator's factual knowledge. In the case of MT this approach is based on extensive research into possible relationships be-tween the basic concepts of the domain and the items of the linguistic database. Creating such a thesaurus or a semantic net is not only ex-tremely laborious, but also space-consuming. And the most serious dis-advantage is that sometimes it is impossible to achieve an unambiguous solution to the problem. For example, a semantic network for *constant amplitude deformation cycle* would show relations between the nodes *constant* and *amplitude, constant* and *deformation, constant* and *cycle*, and this information doesn't help to establish the dependencies struc-ture of the NP both in MT and human translation.

The second approach is more formal: we can use the information obtained from the analysis of the entire text. This approach seems more appropriate, since it is based on formal indicators of the author's inten-tions, which are reflected both in the text structure and in the composi-tion of different NPs with the same components.

NP contextual analysis within the text space, provided by concord-ancing in scientific text corpora, leads to establishing procedures of coining novel NPs from those featuring in the text and to recognizing the compressed sentence structure in a concise form of an NP.

To establish the procedures, we suggest to use MT results for the source part of a parallel corpus as a reference base. Thus, comparing the translations of the English part (source) and the Russian part in a parallel corpus, we can find exact matches of NPs, as well as partly coincidence for term components and their presence in full and com-pressed terms.

For instance, a 3-component NP *design equipment models* in the source English part can be variably translated as *модели расчетного оборудования* or *расчетные модели оборудования*. The English part also has a 2-component NP *design models*, its MT is *расчетные модели*, which finds an exact match in the Russian part: *расчетная модель*. But there is no variant of design equipment with an expected MT *расчетное оборудование*. Nothing similar is found in the Russian part, either (see Table 3). The comparison suggests that *design models/расчетные модели* demonstrates stronger dependences between *design* and *models* in the texts of this subject area, than between *design* and *equipment*. So, the right candidate for a dictionary entry is *расчетные модели оборудования*.

English part	MT stored results	Russian part
design models	*расчетные модели*	*расчетная модель*
design equipment models	**модели расчетного оборудования / расчетные модели оборудования*	*конструктивная модель проектная модель*

Analysis of texts across different subject areas has shown, that if an NP of more than two components appears in the text, it is generally followed by a 2-component NP in the nearest context within the limits of 2–3 sentences, or it can be found in the title, keyword list or abstract. Hence, in human translation this fact can be a clue for NP structure diagnostics. Searching parallel corpora, we may fail to fix such relations, but referring to MT results as a storage base, we can optimize term identification and translation.

4. Conclusion

We have demonstrated that a scientific text abandons in multicomponent terminological NPs, most frequently 2-component combinations, with ambiguous dependency relations. This ambiguity is caused by their syntactic compression, since an NP is normally the result of a sentence or of another NP convolution.

To establish dependency relations of a multicomponent terminological NP it is useful to seek for its variants and modifications that can be found within the same text or texts of the subject area. These modifications are results of a number of standard transformations described in the paper.

Our study of various subject area texts has shown, that if an NP of more than two components appears in the text, it is generally followed by a 2-component NP in the nearest context, or it can be found in the title, keyword list or abstract.

While context analysis and comparison of NP modifications within a text may serve a reliable clue in human translation and manual/traditional lexicographic work, searching parallel and comparable text corpora for terminological equivalents may find few exact matches and the NP modifications may show no kinship of the components. We suggest to optimize the use of full-text parallel corpora for lexicographic and terminographic aims by adding a third part — an MT results corpus — as a reference base to fix and store the history of NP transformations and modifications in the corpus.

References

1. *Babych B., Hartley A.* (2003), Improving machine translation quality with automatic named entity recognition. In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT. Association for Computational Linguistics, pp. 1–8.
2. *Baroni M., Zamparelli R.* (2010), Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1183–1193.
3. *Beliaeva L.* (2014), Applied Lexicography and Scientific Text Corpora. In: Transactions on Business and Engineering Intelligent Applications. G. Setlak, K. Markov (ed.). Rzeshov, Poland: ITHEA, pp. 55–63.
4. *Belyaeva L.* (2009), Scientific Text Corpora as a Lexicographic Source. In: SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern. Conference, November 25–27, 2009. Smolenice, Slovakia, pp. 19–25.
5. *Bergsma S., Wang Q. I.* (2007), Learning noun phrase query segmentation. In Proc. EMNLP-CoNLL, pp. 819–826.
6. *David V., Curran J.* (2007), Adding noun phrase structure to the penn treebank. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics. Prague, Czech Republic, pp. 240–247.
7. *Delgado M., Martin-Bautista M. J., Sanchez D., Vila M. A.* (2002), Mining Text Data: Special Features and Patterns. In: Lecture Notes In Computer Science, Springer-Verlag GmbH. Vol. 2442, pp. 140–151.
8. *Delpech E., Daille B.* (2010), Dealing with lexicon acquired from comparable corpora: validation and exchange. In: Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). Fiontar, Dublin City University, pp. 229–223.
9. *Feldman R., Dagan I.* (1995), Knowledge discovery in textual databases (KDT). In: Proceedings of the 1st Int. Conference on Knowledge Discovery and Data Mining (KDD-95), AAAI Press, pp. 112–117.
10. *Heja E.* (2010), The Role of Parallel Corpora in Bilingual Lexicography. In: N. Calzolari et al. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valetta: European Language Resources Association (ELRA), pp. 2798–2805.
11. *Lavie A., Parlikar A., Ambati V.* (2008), Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In Proc. 2nd SSST, Association for Computational Linguistics, pp. 87–95.
12. *Malakhovskaya M., Beliaeva L., Kamshilova O.* (2021), Teaching Noun-Phrase Composition in EAP/ESP Context: A Corpus-Assisted Approach to Overcome a Didactic Gap. In: Journal of Teaching English for Specific and Academic Purposes. Vol. 9, No. 2, pp. 257–266. doi.org/10.22190/JTESAP2102257M
13. *Reiter N., Frank A.* (2010), Identifying Generic Noun Phrases. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**,** Uppsala, Sweden: Association for Computational Linguistics, pp. 40–49.

14. *Shen L., Xu J., Weischedel R.* (2008), A new string-to-dependency machine translation algorithm with a target dependency language model. In Proceedings of ACL-08: HLT, pp. 577–585.

Беляева Лариса Николаевна
Российский государственный педагогический университет
им. А. И. Герцена (Россия)
Beliaeva Larisa
Herzen State Pedagogical University of Russia (Russia)
E-mail: lauranbel@gmail.com

Камшилова Ольга Николаевна
Российский государственный педагогический университет
им. А. И. Герцена
Санкт-Петербургский университет технологий управления и экономики
(Россия)
Kamshilova Olga
Herzen State Pedagogical University of Russia (Russia)
Saint Petersburg University of Management Technologies and Economics (Russia)
E-mail: onkamshilova@gmail.com

CORPUS PROCESSING: THE LINGUISTIC APPROACH DEVELOPING RESSOURCES FOR RUSSIAN, APPLICATIONS TO LINGUISTICS AND LANGUAGE TEACHING

Abstract. We present a set of linguistic resources developed for Russian: a morphological dictionary associated with a set of grammars to describe the inflection, a semantic component as well as a set of syntactic grammars that solve various types of ambiguities and recognize several named entities. We show how these resources can be used to automatically process corpora and how they can be used to teach Russian as a second language.
Keywords. Linguistics, Corpus linguistics, Natural Language Processing, Russian language, Second Language Teaching.

Introduction

To study Russian corpora, most linguists and language teachers make extensive and almost exclusive use of the "ruscorpora.ru" site of the Russian National Corpus[1], which contains texts with complete morphosyntactic and semantic tagging, *i. e.* completely pre-disambiguated. Russian linguists can also use the "cfrl.ruslang.ru" collection site of the Computer Fund of Russian Language, which can display occurrences of wordforms (but not lemmas) for a verified but limited set of texts.

Most users of corpora (linguists, language teachers as well as many researchers in the humanities and the social sciences) need to explore and study other texts, such as those that constitute the Computer Fund of Russian Language, and more generally, any corpus of texts that they might have collected, from any source. Sketch Engine does offer access to Russian texts published on the Internet. Unfortunately, its linguistic functionalities, while satisfactory for users who are performing global statistical analyses on graphical wordforms, are not reliable enough for any precise linguistic analysis[2].

The NooJ software is a linguistic development environment platform that allows users to formalize eight levels of linguistic phenomena, for any written language: spelling and typography, inflectional, derivational and agglutinative morphology, local, structural and dependency syntax, transformational grammar and semantics. NooJ provides users with formal tools

[1] See website: http://ruscorpora.ru.

[2] See [Kilgarriff et al. 2014; Khokhlova 2009]. [Silberztein 2021] shows that corpus processing tools that do not have access to precisely handcrafted dictionaries and grammars are not reliable enough for many linguistic analyses.

adapted to each type of phenomenon (regular, context-free, context-sensitive and unrestricted grammars), as well as software engineering tools that make it possible to develop, test, accumulate and share linguistic resources with wide coverage. NooJ linguistic resources at any level can then be combined automatically applied to large texts to perform various analyses, often in real time; hence, NooJ is being used as a corpus processor by many linguists, language teachers are more generally researchers in the digital humanities[3].

The INALCO institute[4] has been collaborating with the Vinogradov Russian Language Institute of the Russian Academy of Sciences for over thirty years, working with J. Anoshkina's Unilex and A. Baranov's Dialex software[5], which have made it possible to develop grammatical and morphosyntactic resources for NooJ in a relatively short period of time, using Zalizniak's grammatical dictionary[6]. However, these two software applications did not allow linguists to describe syntactic grammars and apply them to Russian texts.

The grammatical dictionary

We have developed three dictionaries: a dictionary of common nouns; a dictionary of proper nouns, and a dictionary of substantive adjectives. The latter dictionary was constructed to avoid the tedious description of homographic forms of adjectives and nouns; if this dictionary were not activated, words like русский [*Russian*] or новое [*new*] would always be processed as adjectives (and not as potential substantives). Together, the three dictionaries contain about 3,500,000 wordforms, associated with over 95,000 different linguistic analyses.

---

[3] NooJ is a free software and runs on several platforms (Windows, Macos, Linux and Unix) and can be downloaded at: http://www.nooj-association.org.

[4] The *Institut National des Langues et Civilisations Orientale*s (INALCO) is a French university specialized in the study and teaching of Oriental and Slavic Languages.

[5] See [Anoshkina 1993; Baranov 2001].

[6] See [Zalizniak 1977]. [Nagel 2002] describes a morphological Russian dictionary extracted from Dostoïevski's *The Gambler* novel, but only 15 % of its content is freely available. As it was developed on the INTEX/Unitex platform (see note 7), it contains very limited description (no syntax nor semantics), cannot be easily enhanced to link perfective and imperfective lexical entries and is not reversible and therefore cannot be used by a transformational engine such as NooJ's.

Following are extracts from the dictionary and the corresponding morphological grammar:

волейбол, N + m + inan + Sport + FLX = завод
волк, N + m + an + Animal + FLX = волк
газировать, V + ipf + pf + FLX = интересовать

The codes following each lexical entry indicate the category and linguistic properties of the word: N = Noun, m = Masculine, an = Animated; Sport is a semantic domain, etc. The FLX property specifies the entry's morphological paradigm. For example, following is the definition of paradigm "завод":

завод = <E>/Im+s | <E>/Vi+s | а/Ro+s | у/Da+s | ом/Tv+s | е/Pr+s | ы/Im+p | ы/Vi+p | ов/Ro+p | ам/Da+p | ами/Tv+p | ах/Pr+p ;

Inflectional paradigms associate each inflected wordform with several properties, such as Case (Im, Vi, Ro, Da, TV, Pr) and Number (s or p). NooJ uses a dozen basic morphological operators such as <B> (delete current letter), as well as other specifically defined for a certain language; for instance, there is a special reduplication operator <D> specific to Amerindian languages, a special consonant finalization operator <F> for Semitic languages, etc.[7]

NooJ's operators are different from Zalizniak's. They have been designed to be analytic so they are better adapted to the western Slavic tradition. In NooJ, all properties' names and values must be defined in a separate property definition file[8].

The Russian property definition file defines 11 categories: adjectives (A), adverbs (ADV), nouns (N), numerals (NUM), pronouns (PRO), verbs (V), prepositions (PREP), conjunctions (CONJ), interjections (INTERJ), particles (PART), and pparenthetic phrases (INTRO). These categories are further associated with several properties. For instance, adjectives are associated with seven properties:

[7] For a comparison between NooJ, INTEX and Unitex, see http://www.nooj-association.org/intex-and-unitex.html.

[8] The possibility for linguists to define their own set of properties is a crucial characteristic of NooJ; however, this freedom is sometimes unfortunate for linguists who work on several languages of a given linguistic family (*e. g.* Russian, Belarusian and Ukrainian), for which different property definition files have been designed by different teams of linguists.

Genre = m | f | n ; *(masculine, feminine and neutral)*
SGenre = an | inan ; *(animated and non-animated)*
Number = s | p; *(singular and plural)*
Case = Im | Vi | Ro | Da | Tv | Pr | Zv; *(nominative, accusative, genitive, dative, instrumental, prepositional and vocative)*
Degree = Comp | Sup ; *(comparative and superlative)*
Sem = App | Color | Body ; *(Semantic features)*

Nouns are associated with the same properties as adjectives, plus a few extra, for example:

Sem = Human | Forename | Profession | Parent | Body | Concrete | Abstract | Organization | Text | Animal | Food | Arts | Literature | Music | Sports | Topography | Country | River | City | Mount | Lake | Posit | Time | Color ;

For instance, there are 409 lexical entries associated with the class "Animal". Grammatical words (Numerals, Pronouns, Prepositions etc.) have a limited set of properties, e.g.:

Case = Im | Vi | Vip | Ro | Rop | Da | Dap | Tv | Tvp | Pr ;

The massive polysemy of grammatical words forced us to duplicate many entries (for instance: causality or origin for «из» et «от»). When we develop more and more complex syntactic grammars and apply them to corpora, we will be able to decide if it is better to process these words as highly ambiguous (*i.e.* associated with multiple different sets of properties), or not, depending on the fact that these properties might be redundant with the semantic value of the following noun or verb.

Semantic dictionary

We have associated over 5,000 lexical entries with 33 semantic features such as the "Sem" class (see above). Here is a sample of NooJ's dictionary:

барсук, N + m + an + Animal + FLX = рыбак
барсучонок, N + m + an + Animal + FLX = волчонок
барсенок, N + m + an + Animal + FLX = утенок

The semantic classes are based on Tuzov's semantic dictionary[9], which is structured as an ontology (semantic tree), and contains over 145,000 entries, in a format similar to the following one:

Entry	*Semantic Code*	*Grammatical Code*
вагон	$12132411	# {м1 12}
вагонетка	$12132411	# {ж3 168}
вагонеточный	$12132411	# {п1 36}
вагонетчик	$12413220	# {м3о 96}
вагонетчица	$12413220	# {ж5о 1304}

We have converted this file into the clearer NooJ formalism. For example, $12132411 is now represented by the sequence of features: "+PhysObj +Inanimate +Thing +Technics +Transport +Terrestrial +Noengine". We made sure that each class in Tuzov's ontology corresponds to one and only one semantic feature in the resulting NooJ dictionary.

Current limitations of our dictionary

Initially, we had chosen to ignore the "ë" and the tonic accent because they are never indicated in written texts (except for pedagogical applications); we are planning to add them. Maximova and Gulyakova's[10] online Russian grammatical dictionary does not contain any written accent but ë. Note that Hetsevich's team at the Institute of Informatics in Minsk has developed a Russian dictionary that contains the "ë" and tonic accents[11].

In Zalizniak's dictionary,[12] imperfective and perfective forms of a verb are represented as two independent lexical entries. Unfortunately, this means that NooJ cannot link them automatically and thus cannot perform transformations from one aspect to the other one. We are currently working on associating the perfective and imperfective forms of verbs (but only when they are linked semantically).

Pedagogical applications

There are numerous pedagogical applications of NooJ in the Language Teaching domain[13]. Following is an example of a lab session:

[9] See [Tuzov 2004].

[10] See [Maximova, Gulyakova 2011]

[11] See [Hetsevich, Hetsevich 2012].

[12] See [Zaliznyak 1977].

[13] See for instance [Frigière, Fuentes 2015; Rodrigo 2018].

After launching the NooJ software, select "ru" in the menu "Info→Preferences", then load a text file. Select the lexical and morphological resources to be used via the command "Info → Preferences → Lexical Analysis"; select the syntactic and semantic resources to be used via the command "Info → Preferences → Syntactic Analysis".

The first exercise consists of retrieving the list of all the wordforms associated with their linguistic analyses and their frequency, as well as the list of all "unknown" wordforms, *i. e.* those that were not recognized by any of the selected linguistic resources. Students can then see instantly the list of words that they don't know. By double-clicking any occurrence in the concordance, they can then see their wider contexts in the text.

NooJ's lexical parser displays for a given sentence all its potential analyses and ambiguities in the Text Annotation Structure (TAS), see figure 1. Annotations that are displayed in parallel correspond to linguistic ambiguities (at the lexical, morphological, syntactic or semantic levels).

One exercise is therefore to ask students to solve all the ambiguities by deleting the wrong annotations in the TAS (*e. g.* is "мне" a dative or prepositional form? Does "лет" correspond to "лето" or "год"?)[14].

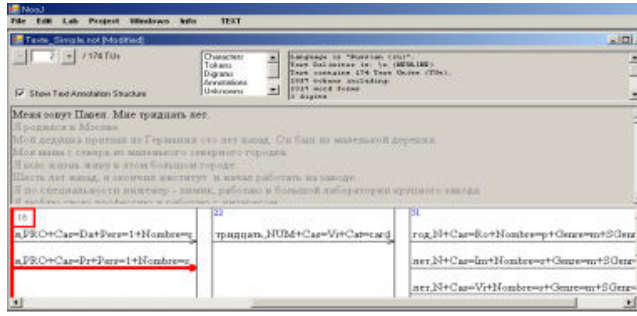


Fig. 1. **Lexical analysis**

Students can enter queries in the form of syntactic grammars and apply them to the text, to locate certain sequences of interest. For instance, grammar "Vb mvt" (movement verbs) recognizes simple sequences that contain movement verbs with no preverb. By applying this grammar to a text, one

[14] Sometimes, students have found problems in the coverage of some of our grammars: either an ambiguous word that should have been disambiguated, or a word that was wrongly disambiguated. With their feedback, we have been able to correct and enhance our disambiguation grammars.

obtains the list of all occurrences of movement verbs in the text. In the same way, grammar "Name.nog" can be used to extract from a text all the structures that correspond to "меня зовут" or "это называется". These grammars are also be used by students to solve ambiguities.

Semantic features are also used in class. Figure 2 below shows the occurrences of adjectives of color in Anton Chekhov's novel "The lady with the Dog", obtained by entering the simple query: <A+Color>. It is interesting to find out that the novel contains mostly occurrences of the colors black, grey and white, and that the color red is mentioned only once.

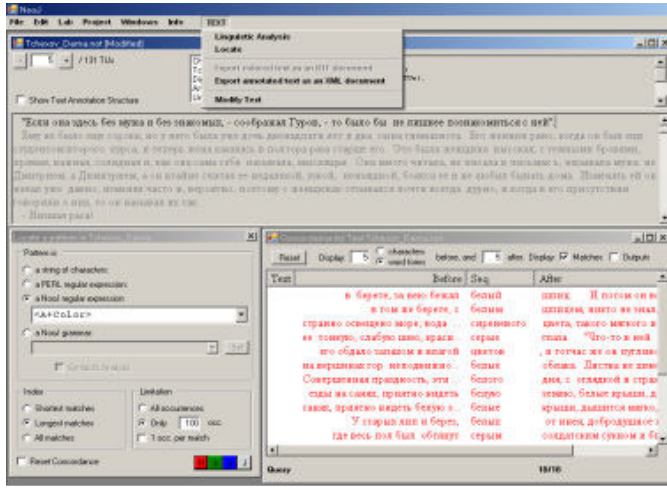


Fig. 2. **Color Adjectives**

Various thematic studies can be performed just by using the semantic codes previously mentioned. For instance, looking for "body parts" would show that hands and eyes are the ones that are the most frequent in the novel. Finally, developing more sophisticated grammars (such as disambiguation rules) constitute a goal for the most advanced students. Figure 3 displays a grammar used to automatically disambiguate the wordform "на" (in "на столе" vs. "на, возьми!").

Disambiguation grammars describe the minimal contexts needed to disambiguate certain wordforms. Figure 4 shows that the wordform "на" will be analysed as a preposition if it is followed by a word in the accusative or prepositional form, whereas it will be analysed as a particle when it is followed by a verb in the imperative.

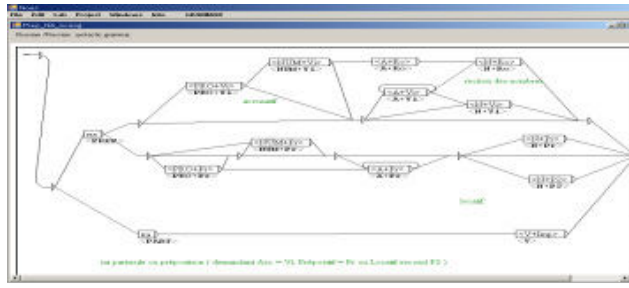


Fig. 3. **Disambiguisating grammar for "НА"**

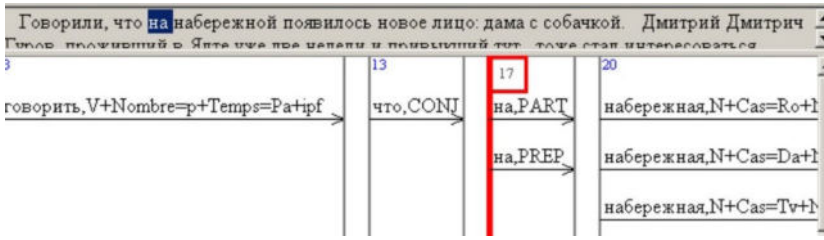


Fig. 4. **Initial Text Annotation Structure (TAS)**

Fig. 5. **TAS after the disambiguation of "НА"**

There are dozens of similar grammars that recognize Russian structures to express possession, ages, dates, durations, locations, etc. that are of interest to all learners of Russian.

Conclusion

We have seen that developing wide-coverage and precise linguistic resources is beneficial, to linguists who want to formalize a natural language, to users of corpora who want to apply sophisticated queries to their own texts, and to language teachers who want to demonstrate to their students how certain linguistic phenomena are occurring in real texts.

Our dictionary is unique because it is fully and freely available for download, and it is fully compatible with the NooJ platform, which guarantees that NooJ's more sophisticated functionalities (such as the formalization of derivational morphology, of constituent and dependency syntax, as well as transformational engine, etc.) can be used to perform more and more sophisticated analyses.

The limitations are only those of the linguistic resources one is willing to develop, and new ones are developed for Russian as well as for other languages every day by linguists of the NooJ community.

References

1. *Anoshkina J. G.* (1993), Preparing frequency dictionaries and concordances on computer with Unilex-T. Moscow, Institute of Russian Language, Computer Fund of Russian Language.
2. *Baranov A. N.* (2001), Introduction to applied linguistics: Tutorial. Moscow, Editions URSS. ISBN 5-8360-0196-0.
3. *Frigière J., Fuentes S.* (2015), Pedagogical Use of NooJ dealing with French as a Foreign Language. In: J. Monti, M. Silberztein, M. Monteleone, M. Di Buono (eds.). Formalising Natural Languages with NooJ 2014. Cambridge Scholars Publishing.
4. *Hetsevich Y., Hetsevich S.* (2012), Overview of Belarussian and Russian Electronic Dictionaries. In: Selected Papers from the 2011 International NooJ Conference, pp. 29–40.
5. *Kilgariff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.* (2014), The Sketch Engine: ten years on. In Lexicography, vol. 1, no. 1, 7–36.
6. *Khokhlova M.* (2009), Applying Word Sketches to Russian. In RASLAN, Recent Advances in Slavonic Natural Language Processing, pp. 91–98.
7. *Maksimova E. A.* (2011), Russian Grammatical Dictionary. Seelrc. Org.
8. *Nagel S.* (2002). Formenbildung im Russischen. In: Formale Beschreibung und Automatisierung für das CISLEX-Wörterbuchsystem.
9. *Rodrigo A., Reyes S., Bonino R.* (2018). Some Aspects Concerning the Automatic Treatment of Adjectives and Adverbs in Spanish: A Pedagogical Application of the NooJ Platform. In: Formalising Natural Languages with NooJ 2017 and its Natural Language Processing Applications. Springer CCIS Series #811.
10. *Silberztein M.* (2003), NooJ manual. URL: http://www.nooj4nlp.net/NooJManual.pdf (date of access: 01.07.2021).
11. *Silberztein M.* (2016), Formalizing Natural Languages: the NooJ approach. Wiley Eds: Hoboken (NJ).
12. *Silberztein M.* (2021), Linguistic Resources for Corpus Processing: the ATISHS project. Proceedings of the 56th Linguistic Colloquium. Linguistik International: Peter Lang International Academic Publishers.
13. *Tuzov B. A.* (2004), Komp'juternaja semantika russkogo jazyka [Computer Semantics of Russian Language]. St Petersburg: St Petersburg University.

14. *Zaliznyak A. A.* (1977). Grammaticheskij slovar' russkogo jazyka [Grammatical dictionary of Russian Language]. Moscow: Russkij jazyk.

Bénet Vincent
INALCO (France)
E-mail: vincent.benet@inalco.fr

Silberztein Max
University of Franche-Comté (France)
E-mail: max.silberztein@gmail.com

I. Chiari, M. Bader,
A. Salem, L. Squillante

USING CORPORA IN BUILDING A MULTILINGUAL GLOSSARY OF MIGRATION

Abstract. This paper describes the different steps in which corpora have been involved in the construction of a multilingual glossary of migration terminology that is meant to represent different stratification of the lexicon pertaining the migration domain starting from a set of corpora of parallel and comparable corpora in Italian English and Arabic. The Language on the Fly project is a monodirectional multilingual platform aimed at representing a prototype resource for glossaries that include a stratification of terminology varied in scope: international, national, and local.
Keywords. Corpus-based lexicography, multilingual glossary, Italian, English, Arabic.

1. Introduction

1.1. Terminology of Migration and its purposes and addressee

Building an updated and reliable glossary of migration is a challenge and at the same time a duty at a time when mass movements, especially towards Europe, oblige us to face daily interactions with a vast number of extensively diverse languages, whose origin countries have customs, cultures, legal and administrative systems sometimes significantly different from those shared in the European framework. Thus, a glossary of migration has both a linguistic and a civil purpose: making accessible, understandable, and comparable terminology that bears a huge impact on human lives and rights.

The challenges posed by the migration lexicon are based on its internal stratification depending on factors such as: national, regional, and local differences, diversity in possible audiences (from institutional international stakeholders to aid workers and finally migrants themselves) and from the intrinsic non correspondence of different migration responses along with rapid changes in legislation for each recipient country taken into consideration.

At a macro level the migration lexicon can be organized into three large sub-set or domains differentiated by scope, and in some cases, partially overlapping:

An international or transnational level, here identified with the institution of the European Union's regulations (legal and administrative/institutional and its migration approach principles, e. g., *Accordo di Cotonou, beneficiario di protezione internazionale, cittadino di un paese terzo, Convenzione di Dublino, discriminazione diretta, migrante di seconda generazione*).

This domain is best described since it needs to be standardized at least for all EU languages (although not taking into consideration the languages of migrants which are seldom corresponding to any of the above-mentioned languages)[1]. This level, from a translation point of view is medium-high standardized in terminology and TE within EU languages, but there is no standardization for languages outside EU languages and in TE. Available glossaries for this level contain a selection of about 500 entries at most and are not derived using corpus-based procedures. Examples of different kinds including Italian language are [European Migration Network 2018, IATE 2018, International Organization for Migration (IOM) 2019].

The second macro sub-set is national in scope, and it regards procedures, regulations, adaptations, and additions that are modified and proposed by each specific country administrative and general migration policies. Each country does in fact implement and define regulations and implementations in individual and not cross-nationally comparable ways. These regulations and implementing decrees are constantly changing depending on government direction and overturns and on public opinion stances (e.g., in Italy for example Security decrees that have deeply changed in time asylum typologies in the last few years, etc.: *Agenzia del demanio, ente territoriale, certificazione sanitaria, ufficio G.I.P., SPRAR, abuso della libera circolazione, associazioni del Terzo Settore, autonomie locali*, etc.). The second layer is country-specific presents low standardization of terminology and TE within EU languages, with partial overlapping with level 1 and no standardization for languages outside the EU.

A last sub-set of the lexicon relevant to migration management process concerns activities that are interlinked to aspects that migrants must face in their interactions with institutions for social security, health, education, administrative issues etc. (*carta di identità, stato civile, certificato di matrimonio, tessera sanitaria*, etc.).

The basic needs regarding glossary enrichment concern: methodology (corpus-based, corpus design and updating); inclusion of languages of migrants; inclusion of the overall stratification of domains pertaining migration not focusing only on institutional and regulative texts. Furthermore,

[1] An example of a multilingual glossary of migration terms at EU level is the [European Migration Network 2018] it contains contains translations of terms into 22 of the EU's 24 languages (about 520 lemmas). While the online version of the Glossary is version 7.0 and was updated in July 2020 with the addition of 28 new terms reflecting the most recent European policy on migration and asylum.

the migration domain is affected by a widespread media usage of terms in a non-technical sense generating potentially dangerous ambiguity (*illegal and legal immigration, refugee, economic migrant*, etc.). For all these sub-strata there is an urgent need to develop strategies and tools that cover significant gaps both at international and national levels for the interest of multiple stakeholders (lexicographers, international and national policy makers, aid workers and cultural mediators and migrants themselves).

1.2. The Language on the Fly resource (LoF)

Language on the Fly is a prototype resource for building corpus-based glossaries and resources in the domain of migration. It is meant to be a full-fledged free online platform which aims to provide reliable and updated linguistic, lexicographical, and documentary information for orientation language in the first reception of migrants and asylum seekers upon arrival in an EU country. The prototype of the platform, to be released by the end 2021, and covers basic local, national, and international terminology in Italian (as source language), English (as lingua franca) and Arabic (as target language)[2].

The Language on the Fly project is by design monodirectional, since it aims to consider especially the lexicon that is used in country specific national, regional, and local texts on migration and that often do not conform to international standards[3].

The goal of the project is to bring together different types of operational, scientific, and psychosocial skills in a way to produce a model of linguistic guidance that is both synthetic and characterized by precision and translation accuracy, and usable for different types of audiences.

2. LoF resource structure and corpus-based approach

2.1. The structure and examples of the LoF cycle

The LoF project is organized in three stages involving the three languages included. The phases of processing are organized in cycles to benefit from

[2] Working versions now are being developed six EU languages (Italian, English, French, Spanish, Portuguese, German) and 10 non-EU languages (Arabic, Azerbaijani, Serbian, Pashto, Russian, Persian, Albanian, Turkish, Chinese, Norwegian). Priority is given to languages that are most represented as countries of origin of migration in Italy.

[3] The prototype itself can nevertheless be applied to any source country (and language) as a methodology and procedure for collection and description of data.

the empirical approach and to consider updates in relevant documents and procedure.

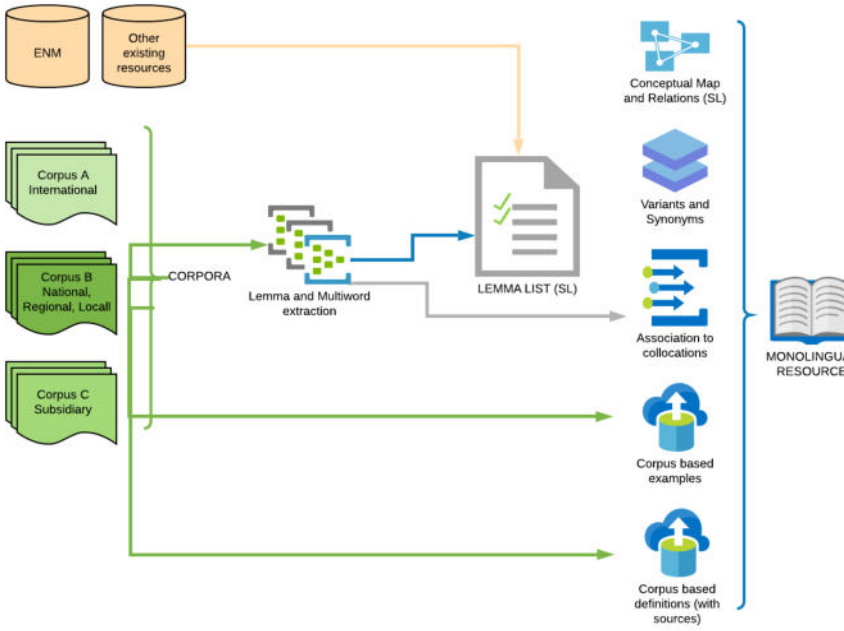


Fig. 1. **Monolingual SL resource map**

As can be observed in Fig. 1 the first step in glossary building is the definition of a starting lemma list. The lemma list is further increased by progressive cycles that enlarge the corpora for SL. As can be seen the starting point is a set of lemmas already collected in previous general glossaries including Italian as SL or TL. The starting list is made by 564 lemmas. But the core of the lemma list extraction is corpus based. Corpora representing the three levels of stratifications have been used and constructed for resource and are constantly updated. At the moment of writing, corpora are composed the following way:

Table 1. **Overview Corpora used or created for LoF (05/2021)**

Corpus Title	Area	Original	Occurrences	Text Typologies
A. 1 Corpus of International EU Migration (Italian)	EU	yes	ca 5,000,000	Documentation from the European Parliament, The Council of Europe, The Court of Justice, The European Commission on migration issues
A. 2 Corpus of International EU Migration (English)	EU	yes	ca 5,000,000	Parallel corpus to Corpus A1
A. 3 Corpus of International EU Migration 2 (Italian)	EU	yes	ca 1,000,000	Documents produced within the EU that have no official translation in English so conceived as a comparable corpus
A. 4. EUR-Lex Italian 2/2016	EU	no	606,070,097	The EUR-Lex Corpus is a multilingual corpus in all the official languages of the European Union. Thanks to the coverage of a vast area of subjects, the corpus is an excellent general-purpose resource for anyone looking for translation examples in many languages.
A. 5 EUR-Lex English 2/2016	EU	no	629,722,593	The EUR-Lex Corpus is a multilingual corpus in all the official languages of the European Union. Thanks to the coverage of a vast area of subjects, the corpus is an excellent general-purpose resource for anyone looking for translation examples in many languages.
A. 6 Comparable Migration Corpus (Arabic)	EU/Int	yes	ca 1,000,000	Documents and translations transnational in nature regarding migration in Arabic.

B. 1 Corpus of National Migration 1 (Italian)	Italy	yes	ca 5,000,000	Italian legislation and documentation of migration management specific to the Italian context
B. 2 Comparable National Migration Corpus (Arabic)	Arab countries	yes	ca 5,000,000	Documents regarding legislation and management of migration in Arab countries.
C. 1 Corpus of Subsidiary National Migration 1 (Italian)	Italy	yes	ca 5,000,000	Documentation for social security, health, education, administrative issues etc.

The first lemma list of LoF was extracted from corpora A. 1 and B. 1 by using keyword (both corpus internal TfIdF and using reference corpora comparison) and collocation extraction measures (N-grams, association measures) and further manually selected. The first release of the LoF resource is (provisionally) composed of 2,094 entries (of which 1,558 multiwords, and 89 named entities).

Following the definition of the entry list a set of procedures are applied for the description of the SL entry properties:

1) Identification of relationships (type of, hypernyms, hyponyms, related items) among lemmas by using external reference resource tree system, in our case EUROVOC descriptors, subjects and Directory code provided for the EU document database;
2) Identification of formal variants and synonyms: using both corpus internal (intra-language) and multilingual parallel corpora (inter-language) methods [A. 1-A2, A. 4-A. 5] and comparable corpora [A. 6, B-2], focusing only on meanings pertaining to the migration domain;
3) Identification and association of collocations and idioms (using N-grams, association measures and work sketches);
4) Extraction of primary source definition for key concepts with relative source: for each term concordances from corpora led us to identify technical definitions of terms, that were added with the source to the glossary. A prototypical alternative definition in a controlled language was further provided;
5) Extraction of usage examples from corpora: following guidelines that avoid editing materials, examples with named entities or dates, and length requirements, examples were selected for each entry.

The result of these operations is converted in a relational database form that produces a monolingual resource of terminology that is the starting point for the processing for all target languages.

For pure exemplification the entry for the SL has the following structure see Fig. 2.

One of the main challenges remains the absence of a common migration framework that grant good quality translations that do not introduce potentially risky ambiguities. In this respect glosses are provided for translations in cases where the cultural and legislative background demands it (some examples will be provided in the following par. 2.2).

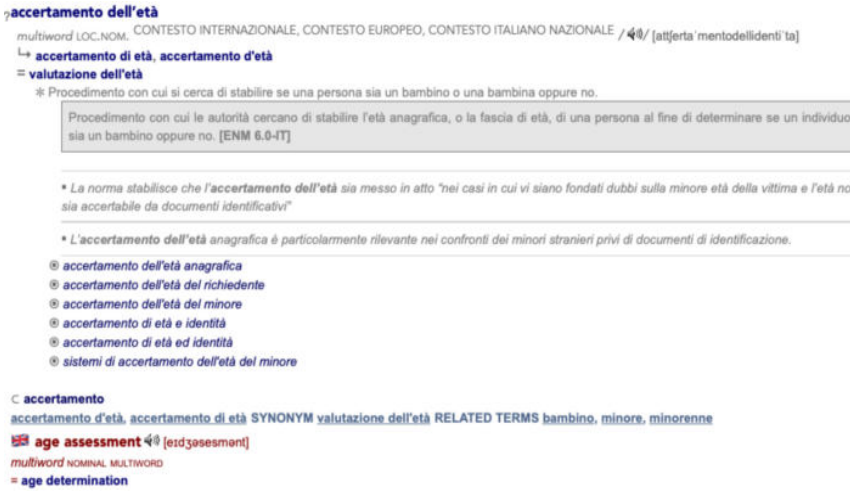


Fig. 2. **Example of Monolingual SL resource database**

We will now give an overview about quantitative data on the resource on some of the languages as it has been collected until the day of writing, see Table 2.

Table 2. **Overview of Entry list of LoF (05/2021)**

Languages (ISO codes)	# entries	# multiwords	variants	synonyms
IT	2,094	1,558	455	970
EN	2,094	1,609	519	1,196
AR	1,009	764	0	22

2.2. Challenges in using corpora for the Arabic version of the glossary

The choice of Arabic as one of the languages for the pilot study has not been casual since Italy is a country that in time has witnessed different waves of migration from Arabic speaking countries, mainly Maghreb and Egypt, but more recently since 2011 has been a final or intermediate destination of many Syrian refugees. The challenge lies in translating terminologies that point out the cultural differences in the Arab communities, to cite some examples: *migrazione* (migration), *matrimonio fittizio* (*marriage of conven-*

ience) and *adozione fittizia* (*adoption of convenience*) but also terminology referring to migration itself. The root for the Arabic word هجر means both to migrate and to abandon. In order to express the verb *to migrate, or to immigrate*, in Arabic we choose a morphological form which is هاجر which expresses a long duration of abandonment/ migration. While in Italian lant guage we can use 3 verbs to express the idea of leaving one's country (*migrare, emigrare, immigrare*), in Arabic language we use the nominal verb only to express the three cases followed by an adjective هجرة وافدة، هجرة مغادرة، هجرة. The previous distinction is reflected also on the word *migrant*, *immigrant* and *emigrant*. It is worth noting that in Italian the last two terms are used in the past participle form, while it is used in the present participle form in English. It is necessary to add an adjective to specify the direction to express the word emigrant (he/she who leaves his/her own country) مهاجر مغادر, while we add another adjective to express the idea of the immigrant, he/she who arrives, we say مهاجر وافد which changes perspective.

Matrimonio fittizio (*marriage of convenience*) is a recent phenomenon which is becoming commonly practiced within Arab communities in the European or Western context, while it is less common in the Arab countries. The closest translation would be زواج صوري ("*false marriage*") or زواج المصلحة ("*marriage of interest*"), was thus translated as an "artificial marriage" or "marriage for benefit".

Whereas *adozione fittizia* (*adoption of convenience*) In the case of Italian *adozione fittizia* (*adoption of convenience)*. The first translation results in an explanation rather than a short translation as it is necessary to provide context and an explanation to the term as it is not simply an adoption, and it means ''adoption for the sake of receiving a residence permit", in this case, the word ''adoption" is the usual concept of adoption. The second one is translated into تبني صوري fake adoption" which comes from صورة which means "an image/photo". Whereas the third translation احتضان means "embracing" a child for the sake of receiving the residence permit. While the fourth translation means "legal fostering of a child" or literally "sponsorship", the origin of the word comes from كفل which also means guaranteeing for example food and care for the child. This translation is the mostly common and used in the Arab world as the cultural and Islamic context of the word adoption تبني is not commonly used by Arab Muslims because the implications of the two words "adoption" تبني and "legal fostering" كفالة are significantly and legally different. Adoption تبني involves giving the child the father's name and family name, and the same child can fully inherit from his father. While in the fourth translation كفالة, which is the legal fostering,

means taking care of the child fully as if he/she's a biological child, however, the child cannot inherit the father's fortune and at the same time cannot take the full name of the father (name and family name), the child can be given the first name of the father or his last name, but not the full name. Christian Arab families can fully adopt children granting them their full names and those children can inherit. The closest translation of the world adoption in a European context is تبني in which a child can be named after his father and can inherit. It is worth pointing out to the fact that a child can still inherit from his parents even within kafala كفالة through leaving a will.

It is possible to identify two main critical areas of translation that pose challenges in the case of building glossaries involving EU languages and Arabic: on one side, there is the language use and connotations, this area includes the terminology or word-usage whether in Modern Standard Arabic or in different spoken dialects across the Arab world. Depending on the word's root or verbal form, the meaning can change, or it can bear different or involving controversial connotations. In addition, there are cases in which the is a missing translation equivalent or multiple equivalents or even translations that refer only to narrower concepts that need to consider. On the other hand, the social and cultural differences are also a significant area to be taken into consideration, for the legislative backgrounds that differ even within different Arab countries and in addition to the different practices based on religious or cultural habits that are not share in host countries.

References

1. European Migration Network (2018), Asylum and Migration. Glossary 6.0. European Migration Network (EMN), Belgium.
2. IATE (2018), Interactive Terminology for Europe. European Union, Belgium.
3. International Organization for Migration (IOM) (2019), International Migration Law N°34 — Glossary on Migration. International Organization for Migration (IOM), Interactive Terminology for Europe.

Isabella Chiari
Sapienza Università di Roma (Italy)
E-mail: isabella.chiari@uniroma1.it

Maha Bader
University of Bergamo (Italy)
E-mail: maha.bader@guest.unibg.it

Alma Salem
Pontifical Institute for Arabic and Islamic Studies (Italy)
E-mail: alma.salem@pisai.it

Luigi Squillante
Sapienza University of Rome (Italy)
E-mail: luigi.squillante@uniroma1.it

N. Cortegoso Vissio, V. Zakharov

A RULE-STOCHASTIC HYBRID POS-TAGGER FOR SRANAN TONGO WITH MINIMAL LEXICON AND TRAINING DATASET

Abstract. This article presents the results of an experiment designed to evaluate the performance of a part-of-speech (POS) tagger for Sranan Tongo. The tagger combines a rule-based approach and a minimal lexicon with a trigram tag model to disambiguate the attributed tags. Since Sranan Tongo has no corpora, the trigram model was trained on sentences extracted from the APiCS dataset (329 sentences, 2853 tokens) and a description of the language (221 sentences, 1660 tokens), that were manually annotated with POS tags for this purpose.
Keywords. Sranan Tongo, rule-based, POS-tagger, trigram tag model, low-resource.

1. Introduction

Sranan Tongo is a Creole language spoken in urban areas of Suriname and by the Surinamese diaspora in The Netherlands. Although since 1950 some Surinamese writers have been publishing their works in Sranan Tongo, it is still primarily a spoken language, used especially in informal communication. The official language in Suriname is Dutch and for this reason, there are very few texts written in Sranan Tongo. As far as natural language processing tools are concerned, there are no automatic morphoanalyzers for Sranan Tongo.

In Sranan Tongo the semantics of words do not determine part-of-speech affiliation and grammatical relationships are expressed through word order. This language shows also a high degree of multifunctionality, so that lexical elements can function as members of different grammatical categories without any change in form [Sebba, 1997: 119]. For example:

- transitive verbs can be employed as nous;
- stative verbs can function as predicative adjectives;
- attributives adjectives are used as nouns;
- most attributives adjectives could perform as adverbs and vice-versa.

Words such as verbs of motion, time and aspect markers or conjunctions are less likely to take other functions in the syntax, although they may have homonyms in other classes. Anyway, in most of the cases, the class of a word is determined by examining the possibilities of a combination of words in the context. For example, given the following sentence:

mi nen Juwan [Voorhoeve, 1956: 191]

The word form "nen" can function as the noun "name" or as the verb "to be called". In addition, "mi" can refer to both the personal pronoun "I"

and the possessive "my". Although the interpretation of the pair "mi nen" as "my name" is correct, it is unlikely to read the whole sentence as "my name *is* Juvan", because the linking copula "na" is missing. Consequently, in the context of Voorhoeve's example, the word form "nen" can only be a verb.

Hidden Markov models (HMM) are probabilistic sequence classifiers that have been widely used in part-of-speech tagging and word class disambiguation. Being a purely stochastic method, HMM need to be trained on corpora but, unfortunately, there is no corpus available yet for Sranan Tongo. Since, manual tagging is an expensive and time-consuming task, the experiment presented here proposes a workaround to train a model on a very small amount of data. The resulting tagger is a hybrid that disambiguates pre assigned POS tags using trigrams.

2. The hidden Markov model

HMM is a stochastic model in which it is assumed that the system to be modeled is a Markov process with unknown parameters (hidden states). The task of the model is to determine the hidden parameters of a sequence, in this case the part-of-speech tags, from the observable parameters, the word forms.

To train the model, it is necessary to compute two parameters in a labeled corpus with POS tags, the emission and transition probabilities:

- the emission probability $p(w_i|t_i)$ is the conditional probability that a word w_i corresponds to the tag t_i. This assumes that the probability of an output observation w_i depends only on the state that produced the observation t_i and not on any other state or observation;
- the transition probability (for trigrams) $p(t_i|t_{i-2}, t_{i-1})$ is the probability that the tag t_i occurs, provided that is preceded by the tags t_{i-1} and t_{i-2}. This relies on the assumption that the probability of a particular tag depends only on the previous two.

Once the model is trained, to predict the most probable sequence of tags for a given a word sequence the following formula is applied:

$$\operatorname{argmax}\left(x_{1\ldots}x_n, y_{1\ldots}y_n\right) \approx \operatorname{argmax}\prod_{i=1}^{n+1} q\left(y_i \vee y_{i-2}, y_{i-1}\right)\prod_{i=1}^{n} e\left(x_i \vee y_i\right), \quad (1)$$

where the left side of the equation is the joint probability for a sequence of word forms and a sequence of tags. The right side of the formula equals the product of two factors: q the probability that tag y_i occurs given before it appeared tags y_{i-1} and y_{i-2} and e, the probability that tag y_i corresponds to

the word form x_i. The most likely tag sequence is the one that maximizes the product of on both sides of the equation [Jurafsky, 2009: 139–141].

Returning to the task of building a POS-tagger for Sranan Tongo, the starting hypothesis is that the size of the training data is not critical to calculate the transition probabilities, provided it covers most of the syntactical features of the language. Another assumed principle is that the sentences extracted from the APiCS database and those given by Nickel et al. are good sources to train the model. Since they are part of works that describe most of the morpho-syntactic features of the language, it is very likely that they will also cover most of the common POS combination sequences in Sranan Tongo.

However, if the training data is not big enough, the lexical items and their respective part-of-speech could not be learned from the data and must therefore be provided from another source to the tagging algorithm. For this matter, a lexicon was extracted from the online version of the Sranan Tongo-English dictionary "Wortubuku fu Sranan Tongo" [Wilner 2007].

In the proposed experiment, instead of calculating the emission probabilities, the model only counts tags during the training process. These counts are then used by the tagging algorithm to estimate "on the fly" a replacement for the emission probabilities. Transition probabilities are computed as usual.

3. Tag set, lexicon and tagging algorithm

Despite the fact that the "Wortubuku fu Sranan Tongo" includes part-of-speech tags for the entries, it does so regarding the grammatical categories of the target languages (English or Dutch). Consequently, some manual editing was required to adjust the entries to the set of tags used in this experiment.

The part-of-speech tags were defined mainly on the basis of the language description provided by [Nickel et al. 1984] with some minor changes, for a total of 31 tags. Along with the parts-of-speech that are commonly listed (noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, numeral, article, etc), other more specific POS tag have also been introduced (proper name, wh-question, copula, determiner, modal, tense and aspect marker, etc).

The compiled lexicon is trusted to cover most (if not all) words of the closed class like articles, pronouns, modals, etc. However, regardless of size, a lexicon cannot contain all open class words in a language like nouns, verbs,

adjectives and adverbs, interjections. To make things worse, Sranan Tongo shows variations in spelling (despite standardization efforts) and borrows extensively words from Dutch. For this reason, the tagging algorithm will come across many words that are not included in the lexicon. To avoid handling unknown words, each word is pre-marked with a list that contains the tags "noun", "verb", "adjective" and "adverb". The preassignment of these tags is an attempt (albeit very simplistic) to simulate the previous described phenomenon of word multifunctionality.

This general assumption of multifunctionality is then restricted by the lexicon following some simple conditions. Once the pre-tagging is done, the algorithm looks for the word form in the lexicon, if it is found, then the preassigned tag list is replaced by the tag list given in the lexicon. Words of the closed class that have a structuring role in the grammar are expected to be identified by the lexicon. If the word is not in the lexicon, the algorithm checks if it starts with a capital letter. If this is not the case, the tags from the preassigned list are considered the best guess. But if the word does start with a capital letter and it is not the first word in the sentence, then the algorithm assumes that it is a proper noun and the preassigned list is overwritten with this tag. However, if the word does occupy the starting position of the sentence, then the algorithm cannot accept the uppercase letter as conclusive proof that the word is a proper noun, so it returns the preassigned list with the addition of the tag corresponding to the proper noun.

After these rules were applied, if the mapping word form/part-of-speech is unambiguous, its probability equals 1 (tags that are not listed for a given word form have probability 0 and, consequently, they cannot occur). If a word form has more than one tag assigned, then the probability for each tag is calculated using one of the three metrics described in the next section.

4. The metrics

All three metrics share the same principle: they use the tag counts from the training set to calculate the probability of the tags for a given word form.

The first metric (A) translates the proportion of the tags in the training data in the context of the assigned tags for a word form. Tags with higher counts in the training data get higher probabilities:

$$p(t_i) = \frac{tf_{i,j}}{\sum_{i=1}^{n} tf_i} \qquad \text{(A)}$$

where *tf* is the number of times the tag *i* appears in the training data *j*. The count is normalized by adding all the frequencies of the tags in the list. For example, if the list for a given word form *w* contains t_1, t_2 and t_3 and t_1 appears 100 times in the training data, while t_2 and t_3 just 20 and 10 respectively, the probabilities of the tags will be 0.77, 0.15 and 0.08.

The metric (B) is basically the same as (A) with the addition of the natural logarithm to smooth out the difference between tag counts:

$$p(t_i) = \frac{\ln(tf_{i,j})}{\sum_{i=1}^{n} \ln(tf_i)} \tag{B}$$

as result, the probabilities of the tags t_1, t_2 and t_3 from the previous example are now closer to each other: 0.47, 0.30, 0.23.

The third metric (C) penalizes frequency, making those tags with lower counts more likely:

$$p(t_i) = \frac{\sum_{i=1}^{n} tf_{i,j} - tf_{i,j}}{\sum_{i=1}^{n} tf_i n - 1} \tag{C}$$

where the numerator is the difference between the total frequencies of the tags in the list and the count for a given tag, so that tags with lower counts get higher values. The denominator contains the normalizing constant. The tags t_1, t_2 and t_3 from the two previous examples will get now the probabilities of 0.12, 0.42 and 0.46 respectively.

Table 1 shows the POS tags given to the word form "moro" (more) according to the lexicon, the frequencies of the POS tag in a training set with 3890 tags and the probabilities calculated on the basis of the three described metrics:

Table 1. Tag **probability for "moro" (more)**

tag	f	part-of-speech	(A)	(B)	(C)
RB	129	adverb	0.66	0.42	0.16
AB	48	quantifier	0.24	0.33	0.37
COMP	16	comparative	0.08	0.24	0.45

5. Experiment and testing results

For the experiment the HMM was trained in four stages to assess the impact of the amount of training data on the results. Consequently, the training data was divided into four samples. The first two contain the examples from the APiCS database [Winford et al. 2013] and the last two, those from [Nicket et al. 1984]:

- T1: APiCS database: 164 sentences, 1472 tokens;
- T2: APiCS database: 165 sentences, 1381 tokens;
- T3: Nickel et al.: 111 sentences, 827 tokens;
- T4: Nickel et al.: 110 sentences, 833 tokens.

During the first stage of training, only sample T1 was used. The other samples (T2, T3, T4) were added one by one in the following instances, which means that in the fourth stage the model was trained with all of them. The performance of the model was evaluated on the test data after each phase. The values presented for precision, recall and f-score are the average of the individual measurements for the 31 tags.

The test data is intended to represent different syntactic constructs. It consists of 70 manually POS-tagged sentences extracted from the "Wortubuku fu Sranan Tongo" [Wilner 2007] (612 tokens).

The employed lexicon was minimal, including only 346 word forms, although it comprised all the closed class words in the online dictionary "Wortubuku fu Sranan Tongo" [Wilner 2007].

Table 2. **Testing results**

		x = T1	x + T2	x + T3	x + T4
(s) cumulative sentences		164	329	440	550
(t) cumulative tokens		1472	2853	3680	4513
(K)	precision	0.75	0.79	0.80	0.78
	recall	0.74	0.78	0.80	0.76
	f-score	0.72	0.77	0.78	0.76
(A)	precision	0.79	0.79	0.81	0.80
	recall	0.72	0.73	0.74	0.72
	f-score	0.73	0.73	0.75	0.73

		x = T1	x + T2	x + T3	x + T4
(B)	precision	0.78	0.80	0.79	0.78
	recall	0.73	0.76	0.77	0.75
	f-score	0.73	0.77	0.76	0.75
(C)	precision	0.71	0.79	0.81	0.81
	recall	0.74	0.80	0.81	0.81
	f-score	0.70	0.77	0.78	0.79

- x is the training data. initialized in the column (x = T1) with the first training sample and expanded in the successive stages;
- the columns (x = T1). (x + T2). (x + T3) and (x + T4) contain the values and results for each of the stages of the experiment;
- (s) shows the cumulative number of sentences and (t) the cumulative tokens the model is trained on after adding a new sample to the training data;
- (K) is a constant = 1 replacing the emission probabilities. so the model predictions depend exclusively on the transition probabilities. This represents the baseline that is expected to be improved upon;
- (A) (B) and (C) show the performance of the model when applying the respective metrics to replace the emission probabilities;

The experiment did not take into account the results of tagging the punctuation signs that would have artificially improved the previous numbers.

5. Discussion

The remarks presented here are drawn primarily from observing the f-score values. As expected. the size of the training data has a significant impact on the model's predictions. However. the baseline (K). which only shows the disambiguation power of the trigrams applied directly on the lexicon constraints. does not improve from (x + T3) to (x + T4) after adding 110 sentences and 883 tokens. Future experiments should explore the threshold. where more training data ceases to significantly alter the transition probabilities.

Metric (A) has generally good precision but low recall. This happens because this metric favors tags with higher counts in the training data. Nouns and verbs are likely to be correctly identified (increasing overall precision). while less frequent parts-of-speech are misclassified (reducing the recall).

The natural logarithm from metric (B) reduces the negative effect of the extreme counts from metric (A) while maintaining the relationship between more and less frequent tags in the training data. Nevertheless (B). just like metric (A). does not improve over the baseline.

Metric (C). that penalizes tags with higher counts in the training data. is the only method to perform above the baseline. It seems to indicate that a metric that promotes less frequent tags has a positive impact on the classifications when combined with the transition probabilities. Further experiments should be conducted to analyze the behavior of this metric with more training data.

7. Conclusions

This paper shows how far a POS-tagger for Sranan Tongo can go with very limited resources. The presented tagger is a hybrid that combines a rule-based and a stochastic approach. It relies on a broad assumption of the language (word multifunctionality), a minimal lexicon (containing almost only closed class words) and a couple of simple rules for assigning possible POS tags to a word form. The probabilities of the attributed tags are estimated by simple tag counts and then disambiguated using a trigram tag model trained with just 550 sentences.

Despite the fact that the performance achieved by the tagger is low, it serves as a baseline for future developments. The stochastic part of the tagger could improve with the addition of more training data. A broader lexicon will certainly limit the incidence of the preassigned tags, easing the disambiguation task and yielding better results.

Finally. although the article revolves around the case of Sranan Tongo. the method drawn here can be extrapolated to languages with a similar structure.

References

1. *Bocharov V. V., Mitrenina O. V.* (2017). Komp'juternaja morfologija [Computer morphology]. In: I. S. Nikolaev. O. V. Mitrenina. T. M. Lando (eds.) Prikladnaya i komp'yuternaya lingvistika [Applied and Computer Linguistics]. 2nd ed. Saint Petersburg. pp. 14–33.
2. *Jurafsky D., Martin J.* (2019). Speech and language processing: An introduction to natural language processing. computional linguistics. and speech recognition. 2nd ed. Pearson. Upper Saddle River. New Jersey. pp. 139–151.

3. *Nickel M.. Wilner J.* (1984). Papers on Sranan Tongo. Summer Institute of Linguistics. URL: https://archive.org/details/rosettaproject_srn_morsyn-1 (date of access: 21.05.2021).
4. *Sebba M.* (1981). Derivational regularities in a Creole lexicon: the case of Sranan. Linguistics an Interdisciplinary Journal of the Language Sciences. De Gruyter Mouton. Vol. 22(4), pp. 719–763.
5. *Sebba M.* (1997). Contact languages pidgins and creoles. St. Martins's Press. New York. Vol. 22(4).
6. *Voorhoeve J.* (1956). Structureel onderzoek van het Sranan. De West-Indische Gids. 37ste Jaarg. pp. 189–211.
7. *Wilner J.* (2007). Wortubuku fu Sranan Tongo. Sranan Tongo English Dictionary. SIL International. Fifth edition.
8. *Winford D., Plag I.* (2013). Sranan structure dataset. In: S. M. Michaelis. P. Maurer. M. Haspelmath. M. Huber (eds.). Atlas of Pidgin and Creole Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL://apics-online.info/contributions/2 (date of access: 21.05.2021).

Victor Zakharov
Saint Petersburg State University (Russia)
E-mail: vz1311@yandex.ru

Nicolás Cortegoso-Vissio
Saint Petersburg State University (Russia)
E-mail: st082534@student.spbu.ru

A. R. Gatiatullin., D. S. Suleymanov,
N. A. Prokopyev, M. M. Saifullin

"TURKIC MORPHEME" PORTAL AS A TOOL FOR UNIFICATION OF ANNOTATION SYSTEM FOR TURKIC ELECTRONIC CORPORA

Abstract. In the present time, in connection with the urgent task of creating electronic corpora of natural languages for their preservation and development, there is a surge in the number of electronic corpora for the Turkic languages. Considering cooperation with corpora developers, as well as with native speaker experts, we propose a single conceptual space for Turkic corpora development through unification of the annotation system using the "Turkic morpheme" portal presented in this paper.
Keywords. Turkology, Corpus Annotation, Linguistic Resource, Frame Ontology.

1. Introduction

Relevance and extreme importance of creating electronic corpora of natural languages for their preservation and development, as well as for a resource base of NLP technologies and language research, is beyond doubt. In the present time, due to the growing interest, understanding and support from government, international institutions, foundations with great attention to the problems of endangered and low-resource languages from UNESCO, there is a surge in the number of electronic corpora for many languages, including Turkic languages [UNESCO LT4All 2019, TurkLang 2020]. For many of languages from the Turkic family several electronic corpora are already developed. The following is a non-exhaustive list of Turkic languages corpora of those resources that we refer to in our research: Turkish National Corpus (TNC) — www.tnc.org.tr; Almaty Corpus of Kazakh language (NCKL) — web-corpora.net/KazakhCorpus; Corpus of the Altai language — altay2.gasu.ru; National Corpus of the Bashkir language — bashcorpus.ru; Bashkir Poetic Corpus — web-corpora.net/bashcorpus; "Tugan Tel" Tatar National Corpus — tugantel.tatar; Corpus of Written Tatar — www.corpus.tatar; Corpus of the Khakass language — khakas.altaica.ru; Corpus of the Yakut language — adictsakha.nsu.ru/corpora/corp; Corpus of the Uzbek language — corpus-uz.herokuapp.com; Digital Corpora in Siberian Minority Languages (Teleut and Shor) — corpora.iea.ras.ru/corpora.

In cooperation with the corpora developers [Abduraxmonova 2021], as well as with native speaker experts, we propose a single conceptual space for Turkic corpora development through unification of the annotation system using the "Turkic morpheme" portal presented in this paper.

2. Analysis of Turkic languages electronic corpora

The development of electronic linguistic corpora presents developers with a wide range of problems, successful solution of which requires combining the results of linguistic research and modern computer methods of linguistic data analysis. Capabilities of a corpus are largely determined by annotation system used in it.

In the context of globalization and integration of scientific research, the issues of linguistic data representation unification in corpora acquire special significance — systems for grammatical categories annotation in particular; this is especially important for groups of related languages. Analysis of the current situation in Turkic corpora (according to articles [Aksan et al. 2018, Kubedinova et al. 2019]) shows that in Turkic linguistics, despite genetic and structural-typological commonality of the Turkic languages, general principles and approaches to linguistic annotation of texts have not yet been formed. Obviously, this will lead to significant difficulties in comparative studies in the future, as well as in development of Turkic parallel corpora, multilingual text processing systems, and in solving of other theoretical or applied problems.

Differences in annotation systems concern both grammatical categories inventory with meta-language of their description, and composition of the required data representation layers. The same morphological categories in different studies on the Turkic languages are annotated differently. Corpus developers often use annotation systems created for Indo-European languages, which do not always adequately and fully reflect the specific features of the Turkic languages, therefore, the development of an annotation system for this particular language group is a very relevant problem.

The lack of uniformity in corpora annotation is associated with objective scientific and organizational problems. The organizational form for the creation of a unified annotation system is to hold joint conferences and seminars, to discuss the unification issues for the Turkic languages, and to implement joint projects. One of such events is the workshop on unification of grammatical annotation systems in the Turkic languages corpora (UniTurk), which is regularly held within the framework of international conferences TEL and TurkLang. At UniTurk (Kazan 2014, Istanbul 2014, Kazan 2015, Bishkek 2016), the problems of developing a unified morphological annotation of Turkic languages texts for use in corpora and automatic text processing systems were discussed.

Despite the organizational measures, practice has shown that these alone are not enough, the problem of unification is being solved rather slowly and ineffectively. Unified linguistic resources and tools are needed in addition with online platforms for discussing these issues. "Turkic Morpheme" portal (modmorph.turklang.net, [Gatiatullin et al. 2020]) was proposed as linguistic resource with a set of services and as a platform for communication on the problem.

3. Description of the "Turkic Morpheme" portal

The first stage is the unification of grammatical categories expressed with morphemes in Turkic languages, both affixal and root morphemes. A glossary of grammemes which express grammatical values has been developed for this purpose (Fig. 1). This glossary is available to all processing annotation data processing software for Turkic languages, in particular, to morphological analyzers that will be used for Turkic electronic corpora unified annotation. A detailed description is given for each grammeme in several languages (Fig. 2). This description is necessary for language experts and developers to unambiguously perceive the annotation system.

Grammemes

Typological name (English)	Typological name (Russian)	National name (Tatar)
1-st person	1-е лицо	1-нче зат
2-st person	2-е лицо	2-нче зат
3-st person	3-е лицо	3-нче зат
Ablative	Исходный падеж	Чыгыш килеше
Accusative	Винительный падеж	Төшем килеше
Active	Основной залог	Төп юнәлеш
Adjective	Имя прилагательное	Сыйфат
Adverb	Наречие	Рәвеш

Fig. 1. **Fragment of grammeme glossary on portal**

The portal has a set of pivot tables that provide functions of database overview for all languages. In particular, these tables make it possible to

assess the degree of grammatical values expression in linguistic units in different Turkic languages and to compare Turkic languages according to the degree of grammatical affinity.

The assignment of grammatical values to linguistic units (affixal and root morphemes, postpositions and postpositions) is actually based on the language expert's intuition. One linguistic unit can be assigned to several of the most frequent grammatical values. For example, the morpheme -GA in the Tatar language expresses two values: dative and directive. In the Yakut language, these grammemes are expressed by two different affixal morphemes.

Common part

Digital identifier	8
Case marking identifier	ACC
Typological name (Russian)	Винительный падеж
Description and source of typological name (Russian)	
Typological name (English)	Accusative
Description and source of typological name (English)	
Grammatical category	Case : Падеж

Language part: Tatar

National name	Төшем килеше
Description and source of national name	Кушымчасы – -ны/-не. III зат тартым кушымчалы исемнәргә -ын/-ен/-н кушымчасы ялгана: кул-ны, сөлге-не; ку-л-ын; сөлге-се-н. Төшем килеше туры объектны белдерә. Бу – аның төп мәгънәсе. Җөмләдә төшем килешендәге исем – тәмамлык. // Татар грамматикасы: өч томда / проект җит. М.З. Зәкиев. – Тулыландырылган 2 нче басма. – Казан: ТӘһСИ, 2016. – Т. II. – 432 б.

Fig. 2. **Description of the grammeme on portal**

In order to see a more complete picture of grammatical values representation, statistical studies on corpus data are needed. However, morphological analyzers and statistical analysis software do not have the appropriate functions for obtaining such data; they only allow to part out

the linguistic units from word-forms. To implement these functions, it is necessary to apply analyzers of semantic-syntactic level.

It is not possible to use the currently popular machine learning technologies for these problems, since most of the Turkic languages do not yet have sufficient linguistic resources to teach them. Therefore, it is important to create the appropriate necessary resources. Moreover, when creating such resources, it is also necessary to observe the principle of annotation system universality within the group of Turkic languages. Obviously, the principle of universality for a language group is easier to observe, having uniform multilingual resources and services for these languages. The proposed "Turkic morpheme" portal can be used as such multilingual resource and set of services.

4. Frame ontologies

Frame ontologies have recently been quite actively used in various subject areas for knowledge processing problems. They serve, on the one hand, as a knowledge base for semantic-syntactic analyzers that are used to annotate the electronic corpora [Yelibayeva 2020], and on the other hand, they are used for language comparison. For example, the Turkic languages have different sets of affixal morphemes, which partially coincide, and the percentage of this coincidence allows expert to determine the grammatical proximity of languages and to classify them into subgroups: Kypchak subgroup, Oguz subgroup, Karluk subgroup, etc.

Some of the most complete and well-known electronic resources that provide semantic role models are FrameNet, VerbNet [Kipper et al. 2006] and PropBank [Palmer et al. 2005]. A similar resource developed for the Russian language is the FrameBank [Lyashevskaya et al. 2015]. It combines a dictionary of lexical constructs with annotated corpus of their expression in texts of the National Corpus of Russian language. FrameBank constructs include predicate-argument structures for verbs, nouns, adjectives, adverbs, and predicatives. The only similar resource for the Turkic languages known to authors is Turkish PropBank [Sahin 2016].

To define a frame ontology, the model proposed in [Avdeenko et al. 2013] is used:

$O_F = <C, R, S, G, T, D_S, D_G, E>$

C — set of frame classes describing concepts of ontology domain;

R — a set of binary relations on frame classes, $R = \{R_{ISA}\} \cup \{R_{ASS}\}$, where R_{ISA} – is a set of "class-subclass" hierarchical relations; R_{ASS} – is a set of associative relations;
S — a set of slots (class attributes);
G — a set of facets (slot attributes);
T — a set for controlled dictionary of ontology domain terms;
D_S – a set of slot types;
D_G —a set of facet types;
E — a set of class individuals.

The peculiarity of frame ontology, implemented in "Turkic Morpheme" portal, is that it is initially multilingual, therefore the model is built similarly to FrameNet around situations. The situation is represented in form of a frame class, which is common to all languages. The structure of the portal database part that implements this frame ontology is shown in Fig. 3 and Fig. 4.

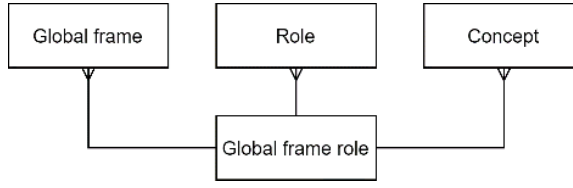


Fig. 3. **Common part of the frame ontology database**

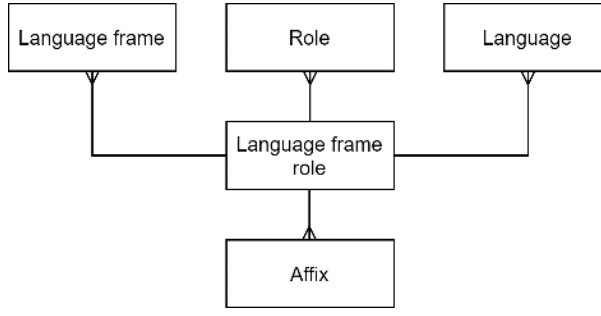


Fig. 4. **Language-specific part of the frame ontology database**

Thus, the complete database of situational frames has a number of applications for Turkic languages corpora development. These include:

1. System of standards for semantic-syntactic annotation;
2. Linguistic resource for semantic-syntactic analyzers development, including the subsequent use in machine learning.

5. Conclusion

The multilingual linguistic database, presented in the "Turkic morpheme" portal, should become an electronic resource that will allow unifying of linguistic resources for Turkic languages. The presented language constructs are to be linked with examples in the Turkic languages electronic corpora. Also, a set of linguistic services will be developed on the basis of the portal according to unified NLP pipeline standard.

References

1. *Abduraxmonova N.* (2021), O'zbek tili korpusini yaratishda lingvistik annotatsiyalash tamoyillari [Principles of linguistic annotation in the creation of the Uzbek language corpus annotation]. In: International Journal of Word Art. Vol. 4, is. 1, pp. 164–174.
2. *Aksan M., Aksan Y.* (2018), Linguistic corpora: A view from Turkish, Studies in Turkish Language Processing. Turkish Natural Language Processing. Theory and Applications of Natural Language Processing. Springer, Cham, pp. 301–327.
3. *Avdeenko T. V., Bakaev M. A.* (2013), Gibridnaya model' predstavleniya znaniy dlya realizatsii vyvoda vo freymovoy ontologii [Hybrid model of knowledge representation for inference in frame ontology]. In: Nauchnyy vestnik NGTU [Scientific Bulletin of NSTU]. No. 3, pp. 84–90.
4. European Language Resources Association (2019), Collection of Research Papers of the 1st International Conference on Language Technologies for All (UNESCO LT4All 2019).
5. *Gatiatullin A., Suleymanov D., Prokopyev N., Khakimov B.* (2020), About turkic morpheme portal. Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020), CEUR-WS.org, pp. 226–243.
6. Institute for history, language and literature, Ufa scientific center of Russian Academy of Sciences (2020), Proceedings of the VIII International Conference on Computer Processing of Turkic Languages (TurkLang 2020).
7. *Kipper K., Korhonen A., Ryant N., Palmer M.* (2006), Extending VerbNet with novel verb classes. Fifth International Conference on Language Resources and Evaluation (LREC 2006), pp. 1027–1032.
8. *Kubedinova L., Adali E.* (2019), The Crimean Tatar Electronic Corpus vs the Electronic Corpus of the Turkish Language (Grammatical Tagging of Noun, Verb). Proceedings of 4th International Conference on Computer Science and Engineering (UBMK-2019), pp. 783–788.
9. *Lyashevskaya O., Kashkin E.* (2015), FrameBank: A Database of Russian Lexical Constructions, Analysis of Images, Social Networks and Texts. AIST 2015.

Communications in Computer and Information Science. Springer, Cham. Vol. 542, pp. 350–360.
10. *Palmer M., Gildea D., Kingsbury P.* (2004), The Proposition Bank: A corpus annotated with semantic roles. In: Computational Linguistics Journal. Vol. 31(1), pp. 71–105.
11. *Sahin G. G.* (2016), Framing of verbs for Turkish PropBank. Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016).
12. *Yelibayeva G., Sharipbay A., Mukanova A., Razakhova B.* (2020), Applied ontology for the automatic classification of simple sentences of the Kazakh language. Proceedings of 5th International Conference on Computer Science and Engineering (UBMK-2020), pp. 13–18.

Gatiatullin Ayrat
Institute of Applied Semiotics of Tatarstan Academy of Sciences (Russia)
E-mail: ayrat.gatiatullin@gmail.com

Suleymanov Dzhavdet
Institute of Applied Semiotics of Tatarstan Academy of Sciences (Russia)
E-mail: dvdt.slt@gmail.com

Prokopyev Nikolai
Institute of Applied Semiotics of Tatarstan Academy of Sciences (Russia)
E-mail: nikolai.prokopyev@gmail.com

Saifullin Miras
Kazan Federal University (Russia)
E-mail: mimsajfullin@stud.kpfu.ru

M.A. Klimova, V.K. Smilga, D.A. Overnikova

USING AN ERROR-ANNOTATED LEARNER CORPUS (REALEC) IN DDL LESSONS[1]

Abstract. The paper describes the experience of introducing the Russian Error-Annotated English Learner Corpus in data-driven learning with regard to error-prone lexical items. The corpus data on confusables were selected and used for preparing teaching materials of two types: the traditional deductive approach and the DDL inductive instruction. Russian L1 first-year university students participated in the corresponding lessons. The results obtained from their pre/post-tests and questionnaires show improvement in the use of most of the target items as well as a positive attitude towards DDL but also point to issues related to DDL materials design and the development of the REALEC corpus.
Keywords. Learner corpora, data-driven learning, confusables, REALEC.

1. Introduction

Data-driven learning, i. e. using the tools and techniques of corpus linguistics for pedagogical purposes, is known to present several advantages: it exposes learners to real examples, performs a corrective function and includes an element of discovery [Gilquin, Granger 2010]. Unlike corpus-based language teaching, DDL not only uses corpus data in the preparation of learning materials, but gives learners access to substantial amounts of corpus data, either indirectly by allowing them to learn about language use by studying concordances prepared in advance by the teacher, or directly by allowing them access to corpora and concordancing software to carry out their own searches [Chambers 2010]. Although the direct hands-on approach provides greater learners' autonomy, prepared materials are known to lead to immediate benefits for learners and teachers with little or no experience in corpus linguistics [Boulton 2010]. DDL activities have mainly addressed the fields of lexis and lexico-grammar with concordances being the major way of presenting the material [Smirnova 2017].

Learner corpora can be considered an important source for DDL as they present students with typical interlanguage features, especially when the data were produced by learners from the same mother tongue background as the students' [Gilquin, Granger 2010]. The existing empirical DDL studies involving learner corpora have shown positive results for the study of error-prone items [Cotos 2014; Moon, Oh 2018].

[1] The research was carried out within the project of the Higher School of Economics Research Foundation "2020 — Automated Analysis of text written in English by learners with Russian L1 (ADWISER)".

English learner corpora containing output from Russian L1 learners, such as SBbEFL, are known to be a valuable source for interlanguage analysis that could potentially contribute to the development of teaching materials [Kamshilova 2012]. REALEC (Russian Error-Annotated Learner English Corpus) has served as a basis for developing an automated testing tool [Vinogradova 2019] but has not yet been introduced into teaching practice using DDL, which is the purpose of the current study.

This paper aims to describe the experience of introducing DDL induction based on the data from the REALEC. As teacher and learner participants have little or no experience in both DDL and using the REALEC corpus in particular, an indirect approach is chosen in order to make DDL immediately accessible. With common learner mistakes in confusables as the target language feature, the study provides learners with guided indirect DDL in the test groups and traditional instruction in the control groups. As a result, both the first experience in DDL and the REALEC as a didactic tool are assessed using pre- and post-tests as well as feedback from the teachers and learners.

2. The REALEC corpus

The Russian Error-Annotated English Learner Corpus (REALEC) is a collection of essays and learner texts that consists of 13,569 pieces of writing, the absolute majority of which are annotated. The corpus contains essays by students of HSE University and is first introduced into language teaching at the same institution, which makes the materials suited for the needs of the learners. As a learner texts corpus, REALEC implements a complex system of error annotation: each error is assigned a correction and one of 98 error tags belonging to 6 major groups — Punctuation, Spelling, Capitalisation, Grammar, Vocabulary and Discourse errors. The REALEC data has several advantages that could make it usable for teaching English as a foreign language, especially to Russian native speakers. Firstly, it can provides an insight into the mistakes most common among Russian L1 speakers, including language-specific ones. Secondly, the system of tags allows for a selection of suitable material even for most specific topics. The sentences used in pre-tests, post-tests, example-based explanations and exercises were obtained from the REALEC corpus.

3. The target language feature

When choosing the target language feature, we relied on previous research, which was concerned with using patterns of error distribution to determine which genre of academic essay featured in the REALEC a certain text is more likely to belong to [Vinogradova et al. 2020]. This allowed us to narrow down the most common mistakes in the corpus that, nevertheless, provided enough clear patterns of misuse.

When analysing the errors of all types in the corpus, we noted a large number — approximately 14,000 — of errors related to lexical choice. This particular group of tags included all REALEC tags grouped under "Word choice", which were further divided into two main tags — "Choice of a part of a lexical item" and "Choice of a lexical item", which contains a single subtag "Words often confused". This subtag, covering both incidences of paronyms and near synonyms, interested us, as it covered a specific word choice mistake, yet at the same time had enough examples to establish patterns and provide the necessary material. However, we soon found that many examples which interested us were actually annotated with the tag "Choice of a lexical item". This can be explained by the fact that, while the tags are considered by the BRAT software, on which REALEC operates, to be separate, one of them is also a subset of the other; therefore, the annotators seem to opt for a more general tag when they are unsure if the tag "Words often confused" is suitable.

After extracting all the instances of this tag from REALEC, we manually analyzed the errors and established a list of clusters of words most commonly confused for each other by English learners whose texts were submitted to REALEC. As a result, we were able to select the three clusters of confusables to be used in our lessons — near-synonymous numerical nouns (*amount, number, quantity*), near-synonymous nouns related to possibility (*possibility, opportunity, ability, potential*), and a pair of paronyms in the form of the verbs *note* and *notice*.

4. Participants

All the participants of the experiment were first-year Linguistics and Philology students of Higher School of Economics in Moscow and Nizhny Novgorod. 41 students attended the corpus-based lesson; 35 students attended the traditional one. In each case students and teachers from both of the HSE campuses were involved. The participants came from 8 learner groups; the lessons were given by 5 teachers. The specializations

of the students made the experience especially useful for them as linguists and philologists will inevitably face corpora while studying or working in the future. Moreover, REALEC is primarily a collection of students' texts written after the completion of a 2-year English course at Higher School of Economics, which is why the material was immediately relevant to the participants.

5. Teaching materials and DDL intervention

Both lesson plans followed a similar structure, containing a pre-test, a three-part introduction of the target language feature, and the post-test. In the control group, this introduction was done using the deductive method: first, the rule outlining the proper ways of distinguishing between near-synonyms was presented explicitly to the students, followed by examples of sentences where the target confusables were utilized.

The test groups, on the other hand, were not presented with the rules explicitly. Instead, they received several concordance lines presenting instances of correct usage. The concordances were followed by a list of questions designed to encourage students to independently arrive at the rules controlling the correct usage of the target lexis.

After either familiarizing themselves with a rule or inducing it from the provided information, students in both groups completed 1–2 exercises testing their understanding of the rule. The exercises came in three types — multiple choice, where students had to choose the correct word out of several options, gap filling, which allowed students to write in their own suggestions, and error correction. The latter type of exercise contained 5 sentences, of which 4 contained mistakes and one did not, which students were explicitly told about. After each block of exercises was checked, the group moved on to the next rule.

While both lessons utilized sentences obtained from REALEC as content for exercises, the test lesson introduced elements of DDL. The first element came in the usage of concordances, extracted from REALEC texts in advance using AntConc, which was emphasized by the use of screenshots including elements of its interface. This helped familiarize students with the concept of a concordance as well as corpus tools. Another way corpus data was integrated into the lesson was a task concerning the cluster of confusables related to possibility (*possibility, potential, ability, opportunity*). In this task, students were asked to analyse several examples of incorrect usage of these words, which were explicitly shown to come from REALEC

2.1. Look at the cases of students using words note/notice **correctly** in their essays.

1. Note

some indicators, which help us to note an interesting tendency in post-school

In conclusion we can note that often difference between proportion of women

, laws. To begin with, I want to note that we live in literal society where

2. Notice

approximately from 35% to 75%. We can notice that France and Sweden have the common

and the USA. Overall, it easy to notice that the number of people older tha 65

because other students or teachers do not notice them and they show off in order

*NB: It must be mentioned that most of the abstracts for notice were taken from **graph description** essays.*

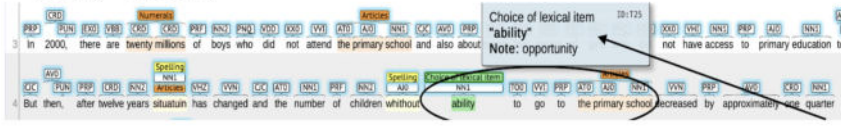
a) What words are you most likely to find after these verbs? Do these two verbs share the same patterns? What are these patterns?
b) Consider animacy/inanimacy of the objects of these verbs. Does it tell you anything about their meaning?
c) Could you think of any difference in the meanings of constructions 'notice that' and 'note that'?

Fig. 1. **An example of the rule being induced in the DDL lesson**

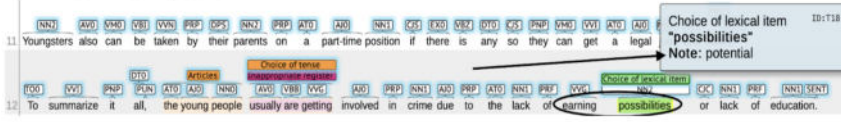
3.2. Look at the cases of students using words ability/possibility/opportunity/potential **incorrectly** in their essays. There is a correction offered by a corpus annotator for each error. Explain **why the words were used incorrectly** and **why the correction is better** in this context.

*Hint: the area important for understanding the context is circled. The correction made by annotator is indicated by an arrow and placed after the word **Note:**.*

1. Ability → opportunity



2. Possibilities → potential



3. Possibility of → opportunity for

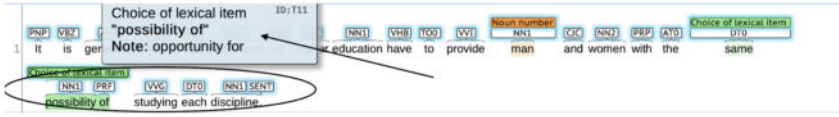


Fig. 2. **An exercise in the DDL lesson showing explicit usage of corpus data**

and included error tags and corrections suggested by the annotators, as well as explanations regarding the interface of the corpus. This exercise was accompanied by an error-correction task, which closely followed the process of tagging in an error-annotated corpus.

In providing instructions for the DDL lesson, we had to consider our target audience, who came from programs centered around philology and linguistics, which usually do incorporate corpus usage in their curriculum. However, we assumed that while the concept of the corpora was not completely new to them, their knowledge as first-year students most likely would not be advanced. The survey conducted after the experiment confirmed that, while most of them did use corpora at some point, the majority used corpora only occasionally, and mainly used Russian corpora. Therefore, while we aimed to introduce students to the concept of using corpora in L2 studies, we also tried to familiarize them with the layout and inner workings of corpus tools, which we hope will be useful for their exploration of corpora in the future.

6. Pre- and post-tests

In order to evaluate the students' progress, we created a pre-test and a post-test. Both of them were similar in their content and consisted of two parts containing five questions each. The first part consisted of multiple-choice questions with four options, containing a combination of target units and other words of similar grammatical form and meaning. The second part was an error correction task, containing five sentences in which the target word was highlighted.

7. Results and discussion

As it was expected, in all groups students' performance increased after they got acquainted with the rules of word usage. However, the evaluation of students' performance before and after the lesson showed no major difference in the two approaches. In the traditional groups the average student performance improved by 13.03 %, while in the corpus-based ones the improvement was 12.18 %.

By the design of the experiment our observations for corpus-based and traditional groups were independent and the resulting data was homogeneous as Levene's test yielded the value of 0.5278592809. Logarithmic transformation was applied to data to make its distribution closer to normal.

Table 1. **Student performance in groups with traditional and DDL-based approaches before and after the lessons**

	Traditional lessons	Corpus-based lessons
Pre-test	70 %	71.25 %
Post-test	83.03 %	83.43 %
p-value	0.0001698674	0.0011476102

Table 2. **Traditional groups' performance**

	Amount, number, quantity	Possibility, ability, opportunity, potential	Note, notice
Pre-test	56.29 %	80.07 %	69.66 %
Post-test	90.09 %	72.29 %	82.42 %

Table 3. **Corpus-based groups' performance**

	Amount, number, quantity	Possibility, ability, opportunity, potential	Note, notice
Pre-test	49.52 %	79.99 %	71.43 %
Post-test	91.67 %	74.25 %	86.37 %

These three factors allowed for the use of two-way Analysis of Variance (ANOVA) to understand how two independent variables (*student's attending a corpus-based or traditional lesson* and *the test being pre- or post-test*) affected the students' performance (*the number of mistakes made in a test*). The resulting p-value was less than 0.05 for the *test* variable only. That lets us draw the conclusion that students' performance differed significantly on pre- and post-tests (p-value = 0.0000000705), while the difference in performance between those who attended the corpus-based lesson and those who attended the traditional one was not statistically significant (p-value = 0.9413137431). p-value was also calculated separately for the participants of traditional and corpus-based lessons using one-way ANOVA, as shown in Table 1. In both cases p-value was less than 0.05, meaning that the difference in performance on pre- and post-test was statistically significant in both groups.

Having proven that the difference in student performance on pre- and post-test, unlike the difference in student performance on corpus-based and traditional lessons, was statistically significant, we can now take a closer look at the percentages indicating student performance on different sets of words (tables 2 and 3).

If we take a closer look at the student performance, we will see that the most significant improvement was achieved for groups of words with unambiguous rules, such as *amount/number/quantity*, whose usage is entirely defined by the dependent noun's characteristics, and *note/notice*, which have unambiguously different Russian translations *отметить/заметить*. All groups' performance for the last set of words, *possibility/opportunity/ability/potential*, worsened after the lesson. This situation may be accounted for by the size of the word group and our decision to include the word 'potential', which caused most confusion among the students during the lesson, as one of the teachers observed. These words have no clear rules of usage motivated by grammar and are usually translated into Russian by the same word (*возможность* or *возможности*). While the approach was effective overall, more examples and explanations were needed in some cases, as one of the students noted in the survey.

The survey answers point out that the major feature of the DDL-based approach was it being unusual and interesting both for those who teach and those who are taught. The experience was generally evaluated as positive: the average willingness of students and teachers to participate in and conduct such classes in the future was 4 out of 5. They were primarily interested in topics similar to the one offered during the lesson, but also showed interest in classes on articles, prepositional phrases and set expressions with the use of corpus material.

Several students commended the use of corpora in the learning process for giving them an opportunity to 'identify the patterns of word usage by themselves' (5 students) and get acquainted with 'real contexts of using the words' (4 students). Another advantage of the DDL-based lesson was it being 'illustrative' (4 students) and implementing more practice than theory (3 students). As well as students, teachers noted that the material was illustrative and touched on important topics that tend to cause much confusion among students.

All in all, the lesson implementing DDL method was as effective in terms of rules acquisition as the one conducted in traditional style. What is more, students' and teachers' comments have shown that the use of corpus in

learning may help to enliven English lessons and make the learning process more dynamic.

We can assume that the difference between the lesson implementing the DDL approach and the traditional one will be more visible in the long-term. That is why the future directions for our work may include conducting a similar experiment with delayed post-testing in order to reveal the retention of vocabulary knowledge, as suggested by T. Cobb [Cobb 1999]. What is more, in our further experiments we should introduce less information in one lesson, as it turned out to be rather difficult for the students to acquire three different groups of words within a relatively short time.

8. Conclusion

Being the first step in the introduction of both the DDL approach and REALEC as a didactic tool in first-year university EFL lessons, the study compared guided DDL induction with traditional deductive learning. The results suggest that the potential of exposing learners to DDL activities based on the REALEC corpus is as significant as using the same corpus data in traditional instruction with regard to confusables as the target language feature.

Although the findings for particular groups of lexis are controversial, the results of the survey confirm the pedagogical value of the DDL approach in terms of the motivation and attitude of both teachers and students. However, more examples in concordances and post-testing should be added to improve the design of further experiments. While the present study, like most DDL interventions so far, was focused on lexis, it has demonstrated that there is a call for data-driven learning related to a wide range of topics, including grammar and discourse. The range of application can be also broadened by combining REALEC data with the use of native speaker corpora.

Some implications can be drawn concerning the annotation of REALEC and its interface. Although the error annotation of the corpus proved usable for EFL teaching, tags for lexis, in particular "Choice of a lexical item" and its subtag "Words often confused", need to be better differentiated in the annotation scheme. Following the indirect approach as a lead-in to DDL, the current study did not include independent use of REALEC by the students. Using more open-ended exploration in EFL lessons would require a user-friendly concordancing tool that would make it possible for students to carry their own searches.

References

1. *Boulton A.* (2010), Data-Driven Learning: Taking the Computer out of the Equation. In: Language Learning. Vol. 60(3), pp. 534–572.
2. *Chambers A.* (2010), What is Data-Driven learning. In: The Routledge handbook of corpus linguistics. Routledge, pp. 345–358.
3. *Cobb T.* (1999), Breadth and Depth of Lexical Acquisition with Hands-On Concordancing. Computer Assisted Language Learning. Vol. 12(4), pp. 345–360.
4. *Cotos E.* (2014), Enhancing writing pedagogy with learner corpus data. In: ReCALL. Vol. 26(2), pp. 202–224.
5. *Gilquin G., Granger S.* (2010), How can data-driven learning be used in language teaching. In: The Routledge handbook of corpus linguistics. Routledge, pp. 359–370.
6. *Kamshilova O.* (2012), Learner Language analysis in SPbEFL Learner Corpus. In: LLLC 2012 Abstracts, p. 40.
7. *Moon S., Oh S.* (2018), Unlearning overgenerated be through data-driven learning in the secondary EFL classroom. In: ReCALL. Vol. 30(1), pp. 48–67.
8. *Smirnova E. A.* (2017), Using corpora in EFL classrooms: The case study of IELTS preparation. In: RELC Journal. Vol. 48(3), pp. 302–310.
9. *Vinogradova O.* (2019), To Automated Generation of Test Questions on the Basis of Error Annotations in EFL Essays: a Time-Saving Tool? In: Learner Corpora and Language Teaching. Vol. 92, pp. 29–48.
10. *Vinogradova O., Lyashevskaya O., Smilga V.* (2020), Correlations between Accuracy, Complexity, and Task Type: Learner Corpus Research (in press).

Klimova Margarita
HSE University (Russia)
E-mail: mfokina@hse.ru

Smilga Veronika
HSE University (Russia)
E-mail: smilgaveronika@gmail.com

Overnikova Daria
HSE University (Russia)
E-mail: daovernikova@edu.hse.ru

A. Orenha-Ottaiano,
M. E. O. de Oliveira Silva

A CORPUS-BASED PLATFORM OF MULTILINGUAL COLLOCATIONS DICTIONARIES (PLATCOL): SOME LEXICOGRAPHICAL ASPECTS AIMING AT PRE- AND IN-SERVICE TEACHERS[1]

Abstract. This paper aims at describing an Online Platform for Multilingual Collocations Dictionaries (PLATCOL), highlighting its relevance to FL teaching and learning. We discuss some lexicographical aspects to develop a customized platform to meet pre- and in-service teachers' needs. Its design, layout and part of the methodological procedures are based on the Bilingual Online Collocations Dictionary Platform. The methodology relies on the combination of automatic methods to extract candidate collocations (Garcia et al., 2019a). Statistical measures, and distributional semantics strategies are applied to select the candidates, and extract examples .
Keywords. Collocations, collocations dictionary, corpus, corpus linguistics, lexicography, pre-service teachers, in-service teachers, customized platform.

1. Introduction

Corpus Linguistics has been successfully applied in different fields of Linguistics, especially in Foreign Language (FL) Teaching and Learning. More specifically, learner corpora have also contributed to a better understanding of the second language acquisition process as well as L2 vocabulary learning and teaching [Granger 1998, 2015]. Besides that, Corpus Linguistics has also been regarded as a powerful tool as well as a suitable match to Lexicography, mainly concerning the retrieval of phraseological units and collocations.

Numerous studies have highlighted the challenges in the FL teaching and learning of collocations [Laufer 2011; Nesselhauf 2005; Martelli 2007; Orenha-Ottaiano 2013, 2020; Torner Castells and Bernal 2017] and a lot of proposals have been put forward to surmount these obstacles.

With a view to contribute to the teaching and learning of collocations and fill one of the gaps regarding the availability of specific lexicographical work, such as a collocations dictionary, we proposed the creation of an *Online Platform for Multilingual Collocations Dictionaries* (*PLATCOL*), in English, Portuguese, French, Spanish and Chinese, whose lexicographical aspects and features are here briefly described.

In Section 1, we address the issue of raising collocational awareness in pre- and in-service teachers to develop collocational competence and the

[1] We gratefully acknowledge the financial support provided by The São Paulo Research Foundation (FAPESP), Process ner 2020/01783-2.

importance of collocations dictionaries use to achieve that. Section 2 describes PLATCOL and focuses on some lexicographical aspects and other features necessary to develop a customized platform suitable to meet pre- and in-service teachers' needs. Section 3 describes our research corpora and briefly outlines the methodological aspects for automatic extraction of corpus data. Finally, Section 4 presents concluding remarks and highlights some ideas for further work.

2. Developing collocational competence in pre- and in-service teachers: the importance of collocations dictionaries use

Developing collocational competence[2] is one of the most challenging tasks for FL learners as well as pre-service teachers. This challenge can also be applied to in-service teachers who still face the same problem even after having graduated, according to our 20-year experience as university professors both in graduate (B. A. in Languages and B. A. in Translation) and post-graduate courses.

In this study, also as a major motivational factor for the creation of PLATCOL, we claim that, among the many ways of helping FL learners, pre- and in-service teachers to achieve collocational competence, the use of online collocations dictionaries can be regarded an important pedagogical tool. In order to get the most out of them, professors should raise pre-service teachers' awareness of collocations and in-service teachers should recognize their relevance to FL proficiency and fluency development. This way, they will be able to better identify the collocational patterns they are looking up and enhance retention.

In general, collocations are mentioned, directly or indirectly, as an object of study in different researches that address the topic of pedagogical lexicography [Higueras 2005, 2006; Pérez Serrano 2014; Torner and Bernal 2017) which confirms the relevance of this type of unit for the teaching and learning of foreign languages.

According to [Neshkovska 2018], training students to regularly consult collocation dictionaries is also highly recommended by FL researchers in general. She adds that collocation dictionaries can be "a significant tool when it comes to mastering collocations". The author also addresses the is-

[2] By collocational competence we understand the ability to identify, understand and, mainly, to produce collocations in a given language context, for translational purposes, for teaching or communicating in a given language.

sue of FL learners being encouraged to look up collocations in dictionaries of collocations on a regular basis, not only in class but also outside the classroom environment. Teachers are also advised to develop activities based on collocation dictionaries.

Rezaeil and Davoudi [Rezaeil, Davoudi 2016] carried out a research on the influence of electronic dictionaries on vocabulary knowledge. Even though the investigation does not focus on collocations, its results can also be extended or applied to collocation knowledge, regarding that a high proportion of language is formulaic in nature and is formed by the co-occurrence of lexical items that, in turn, make up the vocabulary of a language. They concluded that the use of electronic dictionaries can indeed improve vocabulary learning.

The authors also pointed out that students reported more interest and motivation to learn new words when they use electronic dictionaries — and we claim that it can also be extended to online dictionaries or collocation dictionaries. They added that "owing to the multiple modes of presentation and interesting and eye-catching nature of electronic dictionaries, they are embraced more easily by language learners in their attempt to learn new vocabularies in a second of foreign language". Besides these findings, it is worth mentioning the special issue of *RILE* (*Revista Internacional de Lenguas Extrangeras*) on digital pedagogical lexicography, whose presentation stresses the importance of digital dictionaries for language teaching (Nomdedeu 2019). It can be inferred that this type of work must be accessible, reliable and up-to-date and that the concept of dictionary is also a serious issue, as it must offer careful lexicographic treatment (good definitions, adequate lexical selection, inclusion of pragmatic and cultural data, etc.).

Another important aspect to be highlighted here is that electronic, and especially online dictionaries, which is the proposal of this paper, have changed the way users interact with the data they are looking up and that contributes to one of PLATCOL's goals: help users master collocations. A clearer and eye-catching structure for the entries, with a neat and uncluttered page layout, with visually obvious features so that users do not need to understand how it is organized, etc., are all aspects this project lexicographers must have in mind if they aim at compiling an online customized dictionary.

We thus acknowledge the useful role of collocations dictionaries in raising pre- and in-service teachers' awareness of collocations and we recognize them as a valuable pedagogical tool.

3. A Customized Online Platform for Multilingual Collocations Dictionaries for pre- and in-service teachers

Researches that highlight the importance of collocations in the FL teaching-learning process undoubtedly have a long and solid tradition. Many of them are based on neurolinguistic studies that confirm the relevance of the prefabricated elements in the construction of the oral and written discourse of speakers of any language [Lewis 1993]. However, most researches focus on aspects related to students, their needs and their learning experience; on rare occasions, attention is paid to teachers. Thus, in most situations, the previous knowledge that teachers have or should have on the subject, for example, are not considered, even though this may be considered a crucial issue. López Jiménez [López Jiménez 2017], in a comparative study on lexical collocations in English and Spanish as L2 textbooks, found that many of the 208 teachers who participated in the research were unaware of the concept of lexical collocation. On the other hand, the information which teachers must have access to in order to be able to properly conduct the teaching process of collocations or the resources they usually use for their didactics is also not considered.

Taking into account this complex context, it is necessary, when developing a dictionary aimed at pre- and in-service teachers, to inquire about the specific needs of this type of user, in order to adapt its design and the lexicographic treatment given to collocations according to these needs.

We follow the recommendations of the functional theory [Fuertes-Olivera, Tarp 2008, 2014; Tarp 2004] to define, on the one hand, the profile of users and their potential needs, and, on the other hand, to select and present the corresponding lexicographic data.

Hence, users' profile to whom this proposal is directed is:

- in the case of pre-service teachers; language learners (student teachers) from higher education institutions trained to become professional language teachers.
- with respect to in-service teachers: additional language teachers, native or non-native ones, with specific training or degree in Languages.

Based on the definition of the profile, we establish an initial delimitation, of general character, of the typology of needs these users may have. For this, we draw on our experience as professors, researchers and lexicographers. We understand that pre- and in-service teachers will need to use the platform in tasks such as: proofreading and correction of written or oral texts or to

expand their knowledge on the topic, when preparing teaching materials. These needs are linked, thus, to two of the specific extra-lexicographic or social situations identified, by the functional theory: 1. Communicative, in which a user can try to solve a problem related to production, reception, translation, proofreading and correction of written or oral texts; and 2. Cognitive, when the user needs or wants to expand their knowledge of something.

These situations have a direct influence on the configuration of the dictionary. Furthermore, following Tarp's recommendations (Tarp 2004: 233–234), in order to define this configuration, it is convenient to consider the lexicographic data that dictionaries generally offer about collocations. This reflection makes it possible to judge which data is useful, which ones must be improved and which new data must be considered.

On the one hand, we understand that the platform should provide a wide range of examples of usage of collocations, as well as data on their frequency of use. Thus, as shown in figures 1 and 2, each collocation is accompanied by an illustrative example. At the end of the example, users can click on "See more", which will allow them to consult other usage examples. This way, we do not clutter the article with information that may demotivate the dictionary consultation. Data on collocation frequency and collocation statistics are also offered in "Advanced options".

The data available on the platform should allow users to develop pedagogical proposals, according to the methodological perspective they want to adopt. Hence, it is hoped that, after consulting the platform, they can:

a) propose pedagogical materials that serve both for the explicit and implicit teaching of collocations;
b) create activities that provide and encourage:
    i) the development of autonomous learning;
    ii) memorization of collocations.
    iii) working with the different communicative skills.
    iv) linking the collocations with lexical fields or communicative functions.

We must not lose sight of the fact that it is essential to present this information in a clear, attractive and accessible way. Let us remember the results presented by [Chen 2011], in a research about the effectiveness of several computer assisted English collocation tools. The author showed that teachers prefer to use resources that provide quick and easy access to colloquial information and offer many examples of usage.

On the other hand, it is also necessary to pinpoint that teachers do not always have the necessary training to enable them to use dictionaries properly. As a consequence, these works end up being underused [Hernández 2005; Ureña, Penadés 2020]. For this reason, it is essential to include a consultation guide specially designed for teachers, in order to instruct them in its correct use and show them their pedagogical applications.

The entries of the multilingual collocation dictionaries consist of the following elements:

- a headword, which corresponds to the basis of the collocations. This way, the headwords are made up of lexical items such as nouns, verbs and adjectives;
- a word class: a word class is placed right after the headword (the base of the collocation). In the case of these collocation dictionaries, they will be either a noun (n.), a verb (v.) or an adjective (adj.). If a word belongs to more than one word class, such as abstract (n.), abstract (v.) and abstract (adj.), they are shown in separate entries, so that the collocations, collocations structures and other pieces of information are easily organized;
- frequency of each headword;
- a definition — a brief definition of the different senses of the base will be provided. The decision of including a base definition is that the collocations can be duly organized according to each sense of the headword;
- usage examples: to illustrate how collocations are used, based on a specific meaning. Users will have the chance to choose from displaying from 1 to 5 examples.

Hence, users will be able to have a quicker access to the collocations they are searching for (see Fig. 1).

Besides the basic microstructure, *Advanced options* will be available if a user opts to sign in. Hence, according to a users' profile, new dictionary structure will be available as follows (see Fig. 2).

As this is an ongoing research, the lexicographical aspects and other features here reported are still under review and there are other issues that should be considered and tested.

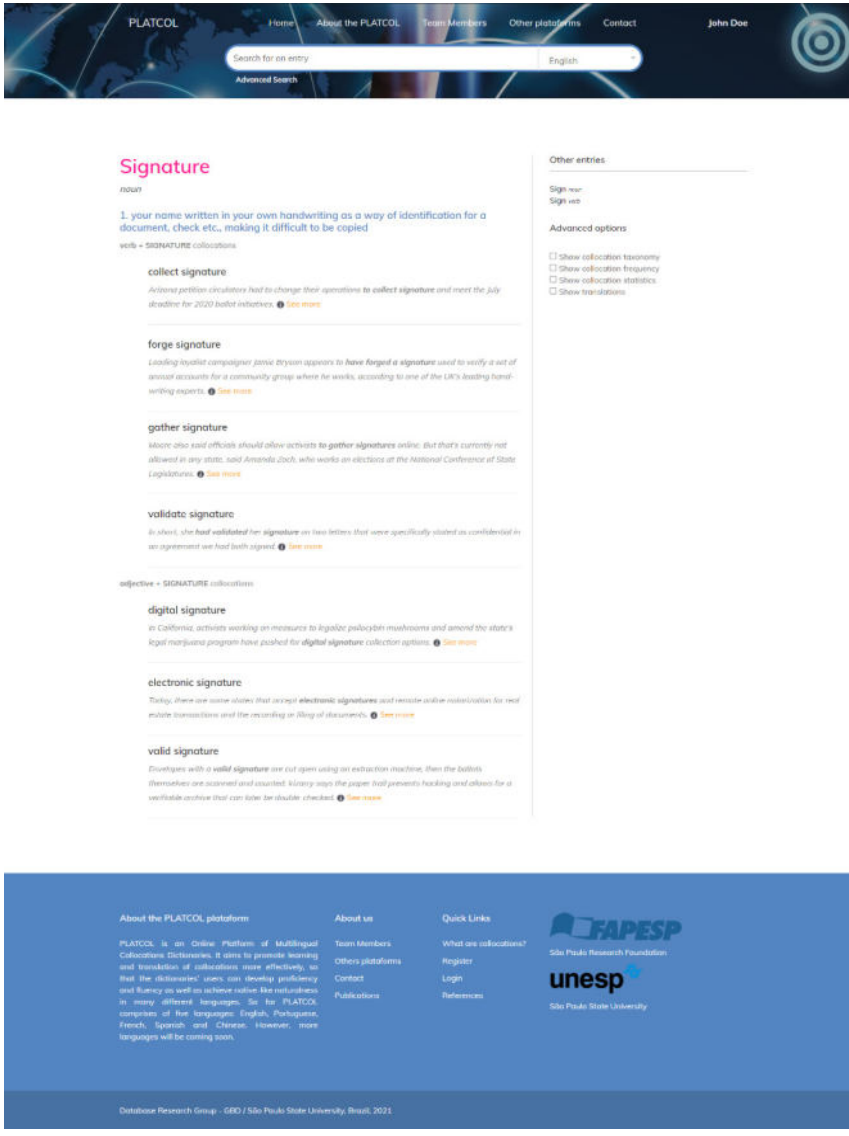


Fig. 1. **Screenshot of PLATCOL'S basic structure of an Entry**

Fig. 2. **Screenshot of Advanced Option Microstructure (Entry is a verb)**

4. Automatic retrieval of bases, collocations and corpus-based examples

A large corpus for each of the five languages of the platform was compiled using different source data, as the Fig. 3 below illustrates.

This research methodology relies on the combination of automatic methods to extract candidate collocations [Garcia et al. 2019a] and takes advantage of NLP tools to annotate large corpora with lemmas, PoS-tags and dependency relations in the five languages.

Statistical measures [Evert et al. 2017; Garcia et al. 2019b], and distributional semantics strategies are applied to select the candidates [Garcia et al. 2019c], and retrieve corpus-based examples [Kilgarriff et al., 2008].

Language	Sources	Size (tokens)
Portuguese	Jornal do Brasil, Wikipedia/Wikibooks, Paracrawl, CHAVE (Santos & Rocha, 2004), CBras, BrWaC (Wagner Filho et al., 2018)	4B
Spanish	EuroParl (Kohen, 2005), Literature (short stories/romances) (Garcia et al., 2019a), Wikipedia/Wikibooks	1,2B
English	EuroParl, Wikipedia/Wikibooks	1.6B
French	FrWaC (Baroni et al., 2009), Wikipedia/Wikibooks	2.5 B
Chinese	Wikipedia, Wikibooks, and literary texts	600M

Fig. 3. **Corpora Sizes and Sources**

5. Concluding remarks and future research

PLATCOL aims to be user-friendly and user-based, hoping to make a contribution to the teaching and learning of collocations and to boost pre- and in-service teachers' text production and collocational competence. When encouraging them to develop collocational competence, we automatically stimulate them to improve their fluency in a given foreign language.

Thanks to the interest collocations have aroused in the past decades, we have advanced in the knowledge of the characteristics of this type of lexical unit as well as of the aspects that must be considered in its lexicographic treatment. The great challenge facing researchers is, above all, to put this knowledge into practice and generate lexicographic works more suitable to the specificities of the object of study, the idiosyncrasies of its users and their lexicographic needs. This is undoubtedly an ambitious goal; however, we believe that the guidelines we have investigated in the development of PLATCOL will make it possible to achieve it.

References

1. *Chen H.-J. H.* (2011), Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. In Computer Assisted Language Learning 24. Vol. 1, pp. 59–76. doi 10.1080/09588221.2010.526945
2. *Evert S., Uhrig P., Bartsch S., Proisl T.* (2017), E-VIEW-affilation — A large-scale evaluation study of association measures for collocation identification. In: Proceedings of eLex 2017 — Electronic lexicography in the 21st century: Lexicography from Scratch, pp. 531–549.
3. *Fuertes-Olivera P. A., Tarp S.* (2008), La Teoría Funcional de la Lexicografía y sus consecuencias para los diccionarios de economía del español [Functional theory of Lexi-

cography and its consequences for Spanish dictionaries of Economy]. In: Revista de Lexicografía. Vol. XIV, pp. 75–95.
4. *Fuertes-Olivera P. A., Tarp S.* (2014), Theory and Practice of Specialised Dictionaries. Lexicography versus Terminography. Berlín/Boston: Walter de Gruyter.
5. *Garcia M., García-Salido M., Alonso-Ramos M.* (2019a), Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics. In: Proceedings of eLex 2019: Smart Lexicography, Sintra, pp. 747–762.
6. *Garcia M., García-Salido M., Alonso-Ramos M.* (2019b), A comparison of statistical association measures for identifying dependency-based collocations in various languages. In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019) at the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, pp. 49–59.
7. *Garcia M., García-Salido M., Alonso-Ramos M.* (2019c), Weighted compositional vectors for translating collocations using monolingual corpora. In: Computational and Corpus-Based Phraseology (EUROPHRAS 2019). Lecture Notes in Artificial Intelligence, 11755, Springer, pp. 113–128.
8. *Granger S.* (ed.) (1998), Learner English on Computer. London and New York: Longman.
9. *Granger S., Gilquin G., Meunier F.* (eds.) (2015), The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press, 761 p.
10. *Hernández Hernández H.* (2005), Quince años después: estado actual y perspectivas de la lexicografía del español para extranjeros. In: Castillo, M.ª A. et al. (eds.). Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua: deseo y realidad. Actas del XV Congreso Internacional de ASELE. Sevilla: Universidad de Sevilla, pp. 465–472.
11. *Higueras M.* (2005), Necesidad de un diccionario de colocaciones para aprendientes de ELE. In: Castillo, M.ª A. et al. (eds.). Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua: deseo y realidad. Actas del XV Congreso Internacional de ASELE. Sevilla: Universidad de Sevilla, pp. 480–490.
12. *Higueras M.* (2006), *Las colocaciones y su enseñanza en la clase de ELE.* Madrid: Arco Libros.
13. *Laufer B.* (2011), The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. In: International Journal of Lexicography. Vol. 24, no. 1, pp. 29–49.
14. *Lewis M.* (1993), The Lexical Approach. The State of ELT and a Way Forward. Londres: Language Teaching Publication.
15. *López Jiménez M. D.* (2017), Colocaciones léxicas: Estudio comparativo de libros de texto de inglés y de español como L2. In: Odisea. Vol. 14, pp. 101–120.
16. *Martelli A.* (2007), Lexical Collocations in Learner English: a corpus-based approach. Alessandria: Edizioni dell'Orso.
17. *Nesselhauf N.* (2005), Collocations in a Learner Corpus. Amsterdam & Philadelphia: John Benjamins.
18. *Neshkovska S.* (2018), What do advanced ESL\EFL students' need to know to overcome 'collocational' hurdles? Thesis. Pristina: AAB College. Vol. 7, iss. 2, pp. 53–74.

19. Nomdedeu Rull A. (coord.), (2019), Lexicografía pedagógica digital [Digital pedagogical lexicography]. In: Revista Internacional de Lenguas Extranjeras. Vol. 10. https://doi.org/10.17345/rile10
20. *Orenha-Ottaiano A.* (2013), Collocations and the design of teaching materials for second language learners. In: Proceedings of the 10th International Conference on Teaching and Language Corpora. Varsovia, Polony: Warsaw: Institute of Applied Linguistics, pp. 93–103.
21. *Orenha-Ottaiano A.* (2017), The compilation of an Online Corpus-Based Bilingual Collocations Dictionary: motivations, obstacles and achievements. In: Proceedings of E-Lex Conference 2017. Leiden, Netherlands, pp. 458–473.
22. *Orenha-Ottaiano A.* (2020), The creation of an Online English Collocations Platform to help develop collocational competence. In: PHRASIS. Rivista di Studi Fraseologici e Paremiologici, pp. 59–81.
23. *Pérez Serrano M.* (2014), ¿Son indispensables los diccionarios combinatorios? [Are combinatory dictionaries indispensable?] In: Revista de Lexicografía. Vol. 20, pp. 121–145.
24. *Tarp S.* (2004), Basic Problems of Learner's Lexicography. In: Lexikos. Vol. 14, pp. 222–252.
25. *Torner Castells S., Bernal E.* (2017), Collocations and Other Lexical Combinations in Spanish: Theoretical, Lexicographical and Applied Perspectives. New York: Routledge.
26. *Ureña Tormo C., Penadés Martínez I.* (2020), Análisis del uso del diccionario en L2. In: Logos: Revista de Lingüística, Filosofía y Literatura. Vol. 30, no. 1, pp. 154–170.

Adriane Orenha-Ottaiano
São Paulo State University — UNESP (Brazil)
E-mail: adriane.ottaiano@unesp.br

Maria Eugênia Olímpio de Oliveira Silva
University of Alcalá (Spain)
E-mail: eugenia.olimpio@uah.es

Natalia Perkova, Dmitri Sitchinava

ON THE SWEDISH-RUSSIAN PARALLEL CORPUS AND ITS POSSIBLE APPLICATIONS (WITH THE FOCUS ON SEVERAL SWEDISH CONSTRUCTIONS)[1]

Abstract. The paper presents some possible applications of the Swedish-Russian parallel corpus within the Russian National Corpus, which is the third largest language pair of the RNC. They are exemplified by the study of several specific Swedish constructions. Last, but not least, the perspective of multilingual constructicons is considered.
Keywords. Swedish, Russian, parallel corpora, construction grammar, constructicon.

1. Introduction

The Swedish-Russian parallel corpus has been developed since 2016 [Sitchinava, Perkova 2019] and has already become one of the biggest parallel corpora within the RNC: the version of November 2021 comprises 16 million tokens, which puts it right after the biggest English-Russian and German-Russian corpora. The following data refer to the version of July 2020 accessed in March 2021 (12 million tokens).

The paper presents several case studies based on the cross-linguistic comparison of constructions, which could be seen as a potential step towards building multilingual construction databases known as constructicons [Lyngfelt et al. 2018; Boas, Höder 2018] implementing key ideas of Construction grammar. Meanwhile, bilingual text pairs provide a more fine-grained optics for comparing constructions that are more language-specific than various more grammatical categories.

2. Possible applications: several case studies

In this section, we will focus on several cases of what can be analysed as constructions in terms of Construction Grammar [Goldberg 1995, 2006]. Here, we decided to choose Swedish constructions, as the corpus is currently not well balanced, and its Swedish part is bigger and more representative.

Both Swedish and Russian are languages that have corpus-based constructicons: they are developed by Benjamin Lyngfeldt's group in Göteborg and Ekaterina Rakhilina's group in Moscow (which recently went into a

[1] Работа Д. В. Сичинавы над статьей поддержана грантом РФФИ 17-29-09154 офи_м «Динамика языковой системы: корпусное исследование синхронной вариативности и диахронических изменений в текстах разных типов».

more international collaboration with the UiT The Arctic University of Norway)[2]. Using the parallel corpus for the purpose of linking these databases to a multilingual contructicon (on these, see [Lyngfelt et al. 2018]) seems to be a very helpful instrument.

In the next section, we show how the corpus can be used in the analysis of two Swedish constructions and their equivalents in Russian.

2.1. V_{imper} lagom

The word *lagom* 'just enough, just right' has recently gained its international popularity as an unofficial Swedish trademark, the symbol of Swedish (and wider, Scandinavian or Nordic) lifestyle. Russian has several equivalents that render this meaning rather well: в *меру*, *как раз*, *как(ой) надо*. However, used with imperatives [Olsson 2013] it can form the construction with the following meaning (translated from Swedish; the definition comes from the constructicon): "it is used to ask someone (Actor) to stop doing or pretending to do a certain thing (Activity)". Here we have a lexically filled slot (*lagom* is obligatory), while the verb slot allows for some variation. The construction has an evaluative semantic component: the attitude of the speaker is critical, rather negative, which can be reduced neither to just the imperative form or the word *lagom* alone. Moreover, as Olsson [Olsson 2013] notices, it means not only that one should act in just a moderate *lagom* manner, but rather should stop doing something at all.

A query can be built as follows: **V & imper на расстоянии от 1 до 3 от** (at the distance 1 to 3 from) **lagom**. The distance parameter helps to capture examples with particle verbs or a postposed addressee *du* 'you'; without this parameter we would get 4 examples, but the extended query returns 7 examples, which is not that bad, considering the relatively low frequency of this construction. The example from Lindgren's Karlsson story illustrates rather a more compositional reading: *lägg för dej lagom* 'put yourself at a more moderate extent'. Other examples seem to fit the construction semantics. Interestingly, one of the examples (4) is of the Russian-Swedish direction: the translator used the construction to render Gogol's specific expressive colloquial phrases.

[2] Static versions of these constructicons are available at the Språkbanken's page with lexical resources: https://spraakbanken.gu.se/karp/#?mode=konstruktikon&lang=swe (the Swedish constructicon) and https://spraakbanken.gu.se/karp/#?mode=konstruktikon-rus (the Russian constructicon). A more recent version of the Russian constructicon, still work-in-progress, being extended by more constructions, examples and features, can be currently accessed here: https://constructicon.github.io/russian/.

(1) — ***Fjäska lagom****, fnyser hon.*

— Хватит подлизываться, — фыркнула она. [Annika Thor. En ö i havet (1997) | Анника Тор. Остров в море (М. С. Конобеева, 2006)]

(2) — ***Hyckla lagom!***

— Нечего притворяться!
[Maria Gripe. …och de vita skuggorna i skogen (1984) | Мария Грипе. …и белые тени в лесу (Анна Зайцева, Ксения Коваленко, 2005)]

(3) …åja, åja, Gunilla, **lägg** för dej **lagom**, JAG ska väl också ha lite tårta?

Эй, Гунилла, Гунилла, ты слишком много накладываешь себе на тарелку! Я ведь тоже хочу пирога… [Astrid Lindgren. Lillebror och Karlsson på taket (1955) | Астрид Линдгрен. Малыш и Карлсон (Л. Лунгина, 1957–1973)]

(4) Но из угрюмых уст слышны были на сей раз одни однообразно неприятные восклицания: «Ну же, ну, ворона! зевай! зевай!» — и больше ничего.

Men från den buttra munnen hördes denna gång bara enformigt obehagliga rop: ”Så ja, så din kråka! **lata dig lagom**!” och det var allt. [Н. В. Гоголь. Мертвые души. Том 1 (1842) | Nikolaj Gogol. Döda själar (Staffan Skott, 2014)]

According to [Olsson 2013], the most prominent exemplars of this constructions are *skratta lagom* (*skratta* ‘to laugh’) with a synonym *garva lagom*, *skryt lagom* (skryta ‘to boast’) and *skrik lagom* (*skrika* ‘to shout’). Among other examples, verbs *raljera* ‘to joke ironically’, *jubla* ‘to rejoice’, *smöra* ‘to fawn on’, *håna* ‘to make fun of’, *skälla* ‘to scold’, *whina* ‘to whine’, *mobba* ‘to mob’, *klaga* ‘to complain’, *stirra* ‘to stare’, *kaxa* ‘to be cocky’, *tjata* ‘to nag’, *retas* ‘to tease’, *hyckla* ‘to be hypocritical’, *stila* ‘to show off’, *hasta* ‘to hurry up’, *gnälla* ‘to whine’ are mentioned. It can be seen that most of these verbs denote certain speech-related actions or some undesirable behaviour. In the comprehensive Swedish-Russian dictionary [Marklund-Sharapova 2007] only *skrik lagom!* ‘не кричи (ори) так!’ is mentioned in the entry for *lagom*. Examples from the parallel corpus include verbs *skratta*, *fjäska* ‘to fawn on’, *hyckla*, *skryta*, *tjoa* ‘to yell’. The example from Gogol with *lata sig* ‘to be lazy, to idle’ stands out, but still it seems that the negative connotation has been captured by the translator, and the choice of the Swedish construction is not random due to the shared evaluative semantic component.

In our sample, the Russian equivalents represent several constructions. First of all, reduplication of imperative forms (with a specific intonation) has a similar component of negative evaluation (*смейся-смейся*; *зевай, зевай*). Second, the construction **хватит VP-Imp.Inf**, listed in the Russian constructicon[3], is attested in example 1. Its synonym, **нечего VP-Inf**[4], is illustrated by example 2. Other expressive analogues are *Да ладно тебе!* (related to the construction **ладно Pron-2.Dat VP-Inf**) and *не больно VP-Imper*, where *больно* is metaphorically used as intensifier (='Don't do VP too extensively').

2.2. i ADJ$_{superl}$ laget

The next construction in our list is again rather specific due to its evaluative component. Its form-meaning relations are more idiomatic, as the fixed lexical component, the word *laget* (*lag*) 'team' has a desemantisized meaning here, and the superlative form of the adjective does not fully correspond to the meaning of the construction: "A phenomenon (Theme) has a property (Property) of unreasonable proportions with respect to the (implicit) standard". In other words, the evaluation related to this implicit standard is what defines the resulting semantics of this construction.

In this case, Marklund-Sharapova [Marklund-Sharapova 2007] gives more information with several examples: *kjolen är i kortaste laget* 'юбка коротковата', *500 kronor är i mesta (minsta) laget* '500 крон — это многовато (маловато)', *i senaste laget* 'поздновато; в последний момент'. Based on these examples, one might draw a conclusion that Russian attenuative adjectives with the suffix *-оват-* render the desired meaning. Indeed, they are, besides their properly attenuative function, widely used for understatement expressing negative qualities (cf. *плоховатый, глуповатый, трусоватый, скучноватый* lit. 'slightly bad, silly, cowardly, dull' vs. **хорошеватый, *умноватый, *храброватый, *интересноватый* lit. 'slightly good, intelligent, courageous, interesting' [Kagan, Alexeyenko 2011: 322]).

---

[3] Its meaning is described as follows: "The construction is used when the speaker wants the interlocutor to stop [some action]Action that is currently taking place. The speaker evaluates this action negatively, as it causes him/her discomfort or seems to the speaker too long".

[4] Its meaning is described as follows: "This construction is used when someone is [doing something]Action they should not be doing, e. g. баловать детей 'spoil children'. The subject tells them to stop. The construction has a negative connotation".

Let's look at the examples from the parallel corpus. The query **i на расстоянии 1 от A & supr на расстоянии 1 от laget** is the first option to come with; one can also filter examples with *i senaste laget* as the most frequent collocation, adding **-sen** to the adjective lexeme field. The former query retrieves 12 examples from 10 texts. The three examples with *i senaste laget* are rendered as в *последний момент* 'at the last moment'/ *поздновато* 'somewhat late'/ *опоздали* 'are too late'. The equivalents for several other examples are as follows: ***i högsta laget*** (*hög* 'high') — *запредельный*, ***i vekaste laget*** (*vek* 'weak') — *слабоватый*, ***i äldsta laget*** (*gammal* 'old') — *староватый*, ***i futtigaste laget*** (*futtig* 'ridiculously small') — *чепуха*, ***i minsta laget*** (*liten* 'small, little') — *совсем маленький*, *меньше некуда*. Finally, two trickier examples are given below: here, the collocation *явная натяжка* 'an obvious stretch' and the idiom *шито белыми нитками* 'completely transparent' are used to render constructions with the superlatives of *pretentiös* 'pretentious' and *genomskinlig* 'transparent'.

(5) Det vore **i pretentiösaste laget**.

Это было бы явной натяжкой.
[Maria Gripe. …och de vita skuggorna i skogen (1984) | Мария Грипе. …и белые тени в лесу (Анна Зайцева, Ксения Коваленко, 2005)]

(6) Nej, det var **i genomskinligaste laget**.

Тут все шито белыми нитками.
[Мария Грипе. Тень на каменной скамейке (Елена Ермалинская, Елена Серебро, Ирина Матыцина, Мария Хохлова, 2005)]

It can be seen from the examples that attenuative adjectives are not the only strategy that can be chosen by the translators. However, it should be noticed that in some cases the choice is defined by the context, so the semantics of the adjective plays a crucial role here.

3. Conclusions

We have presented the Swedish-Russian parallel corpus, a valuable tool for various linguistic studies, showing that it can be successfully used in exploration of language-specific constructions. The analysis of the Swedish constructions V_{imper} + *lagom*, and ADJ_{superl} *laget* shows the perspectives of further exploration of equivalence between constructions from the two languages. The crosslinguistic comparison of constructions shows that a multi-token ("syntactic") construction can be represented in another lan-

guage by a derivational model: for example, the Swedish (and other Germanic) compounds are well known to correspond to syntactic units in Russian, whereas the analytical *i* ADJ_{superl} *laget* construction has among its equivalents the *-оват-* derivational model. These examples underline transparency of the boundary between constructions and derivation (cf. the idea of Construction Morphology in [Booij 2010]) and call for a wider range of items included in constructicons.

References

1. *Boas H., Höder S.* (2018), Construction Grammar and language contact: an introduction. Constructions in contact: Constructional perspectives on contact phenomena in Germanic languages. Amsterdam/Philadeliphia: Benjamins, pp. 5–36.
2. *Booij G.* (2010), Construction Morphology. Oxford: Oxford University Press.
3. *Goldberg A.* (1995), Constructions: A Construction Grammar Approach to Argument structure. Chicago.
4. *Goldberg A.* (2006), Constructions at Work. The nature of generalization in grammar. Oxford.
5. *Kagan O., Alexeyenko S.* (2011), Degree modification in Russian morphology: The case of the suffix *-ovat*. In: I. Reich et al. (eds.). Proceedings of Sinn & Bedeutung 15. Saarbrücken, Universaar — Saarland University Press, pp. 321–335.
6. *Levshina N.* (2015), European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. In: Folia Linguistica. Vol. 49(2), pp. 487–520.
7. *Lyngfelt B. et al.* (eds.) (2018), Constructicography: Constructicon development across languages. Amsterdam/Philadelphia: Benjamins.
8. *Lyngfelt B., Torrent T., Laviola A., Bäckström L., Hannesdottir A., Matos E.* (2018), Aligning constructicons across languages: A trilingual comparison between English, Swedish, and Brazilian Portuguese. B. Lyngfelt et al. (eds.), pp. 255–302.
9. *Marklund-Sharapova E.* (2007), Novyi bol'shoi shvedsko-russkii slovar' [New Great Russian-Swedish dictionary]. Moscow: Zhivoi jazyk.
10. *Olsson E.* (2013), "Garva lagom!" Imperativa fraser med *lagom*. L2-uppsats, Göteborgs universitet/Institutionen för svenska språket.

Перкова Наталья Викторовна
Упсальский университет (Швеция)
Perkova Natalia
Uppsala University (Sweden)
E-mail: nofpernat@gmail.com

Сичинава Дмитрий Владимирович
Институт русского языка им. В. В. Виноградова РАН (Россия)
Школа лингвистики, Национальный исследовательский университет «Высшая школа экономики» (Россия)

Sitchinava Dmitri
Vinogradov Russian Language Institute of
the Russian Academy of Sciences (Russia)
School of Linguistics, Higher School of Economics (Russia)
E-mail: mitrius@gmail.com

SIMPLICITY BEATS SOPHISTICATION: AN EVALUATION OF ADJUSTED FREQUENCY MEASURES

Abstract. Adjusted frequency measures are required to generalize frequency counts obtained from a corpus to the whole population it represents. However, no systematic evaluation of such measures has ever been made. In this paper, I describe a way of testing whether an adjusted frequency measure is good at predicting the frequency ranking of words in unseen data. 11 adjusted frequency measures are compared using different-sized corpora of eight languages. The results show that Range, one of the simplest adjusted frequency measures, combined with Plain Frequency at the second level of sorting, provides the best ranking. Average Reduced Frequency (*ARF*) outperforms all other measures, except for Range.
Keywords. Adjusted frequencies, range, Average Reduced Frequency, evaluation, cross-validation.

1. Introduction

Compiling a frequency list does not seem to be a complicated task: take a tokenized corpus, calculate how many times each word type occurs, and sort the list in descending order. This approach is unproblematic if the corpus one uses is also one's object of interest, i. e. if the corpus is equal to the population. However, in most cases we are interested not in the corpus itself, but rather in the population it represents. This means that a frequency list compiled from a corpus is mostly not a frequency list of the population, but a frequency list of the sample that approximates the frequency distribution in the population more or less successfully. One might say that building a frequency list of a language is absolutely impossible and that we can only construct a frequency list for a certain corpus, but this is not what public expects from lexicographers. For instance, Routledge has been publishing frequency dictionaries since 2008, starting with Portuguese [Davies, Preto-Bay 2008]; the series includes frequency dictionaries of 14 languages (American English, Arabic, Czech, Dutch, French, German, Japanese, Korean, Mandarin Chinese, Persian, Portuguese, Russian, Spanish, Turkish). These dictionaries would be of little interest for language learners if they were called "A Frequency Dictionary of the Turkish National Corpus" or "A Frequency Dictionary of Russian Internet Corpus" rather than "A Frequency Dictionary of Turkish" [Aksan et al. 2017], "A Frequency Dictionary of Russian" [Sharoff et al. 2013], etc. Thus, even though it is impossible to compile a frequency dictionary of a language, this is an aim that lexicographers and corpus linguists have always been trying to approach.

It is never an easy question whether a corpus is truly representative of the language as a whole or not, cf. the seminal paper by [Biber 1993]. Even if we assume it to be representative, there is another unavoidable problem: each text in the corpus has a topic, and the words relating to these topics occur in the corpus more often that one would expect in the language in general. Since [Kilgarriff 1997], this issue is known as the *whelk* problem: if a corpus contains a text on whelks, this word is going to be much more frequent in this corpus than in the language. For lexicographic purposes, e. g., in order to compile a frequency dictionary, it is important to know the "true" ranking of the words rather than a corpus-specific ranking, which means that we need to find a way to construct this "true" ranking based on a corpus we have at hand.

For some applications, it might be necessary to find out not only the ranking, but also the "true" frequencies of the words in a language; a similar task has been often addressed in Natural Language Processing where it is important to estimate frequencies of higher-order ngrams even though many of them remain unseen [Jurafsky, Martin 2009]. However, it is outside the scope of this paper to discuss how one can arrive at "true" frequencies, since humans using frequency dictionaries are only interested in rankings rather than exact counts; I am going to focus on adjusted frequency measures that can be used to obtain a robust ranking of words in a frequency list.

2. Adjusted frequency measures

The most comprehensive survey of adjusted frequency measures was compiled by [Gries 2008]. This paper was primarily focused on measures of dispersion, especially the new measure DP introduced by the author, but it contains two sections (2.2 and 2.3) on adjusted frequencies.

Adjusted frequency measures are classified into two groups. The first group includes measures that are based on dividing a corpus into equally-sized parts. The simplest adjusted frequency measure, which is also a measure of dispersion, is called Range (R). If we divide a corpus into n equally-sized parts, R is the number of parts that contain at least one instance of the word whose frequency is being computed. Other measures take into account the frequencies of a word in different parts of the corpus (v_1, v_2, … , v_n; their mean is denoted as $\overline{v}$ and their sum is f, the total frequency of the word) and mostly rely upon some measure of dispersion, e.g., standard deviation. Here is the list of these measures based on [Gries 2008],

but also on the primary sources where some clarification was required [Carroll 1970; Kromer 2003]:

Juilland's U: $f \cdot \left(1 - \frac{\sigma}{\overline{v}\sqrt{n}}\right)$, where $\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(v_i - \overline{v}\right)^2}{n-1}}$

Rosengren's AF: $\frac{1}{n^2}\left(\sum_{i=1}^{n}\sqrt{v_i}\right)^2$

Carrol's U_m: $f \cdot D_2 + \left(1 - D_2\right) \cdot \frac{f}{n}$,

where $D_2 = \left(\log_2 f - \left(\frac{\sum_{i=1}^{n} v_i \log_2 v_i}{f}\right)\right) \cdot \frac{1}{\log_2 n}$

(if $v_i = 0$, assume $v_i \log_2 v_i = 0$)

Engvall's measure: $f \cdot \frac{R}{n}$.

Kromer's U_R: $\sum_{i=1}^{n}\left(\psi\left(v_i + 1\right) + C\right)$, where ψ is the digamma function, and C ≈ 0.58 is the Euler-Mascheroni constant.

Median (M) is another simple measure that can be mentioned here. If the median of the values v_1, v_2, … , v_n is multiplied by n, this results in an estimate of the frequency of the word in question.

Among these seven measures, Range and Median have a drawback with respect to ranking. The application of these measures results in numerous ties, since R has only n possible values, and M will very likely be the same for many low-frequency items. If one intends to use these measures for ranking, one can introduce a second level of sorting. Namely, for words with the same R or M, Plain Frequency f serves as a tie breaker. Technically, this can be implemented by adding f / l to R and to M for each word, where l is the length of the corpus. For instance, if a word with f = 650 occurs in 90 out of 100 parts in a corpus that contains 1,000,000 words, its R is assumed to be 90.00065 rather than 90. Such a word would be sorted above a word with f = 500 that also occurs in 90 parts of the corpus, because its R would be 90.0005.

Another group of measures is based on analyzing the distances between the occurrences of a word w in a corpus of size l. These distances are denoted as d_1, d_2, … , d_n, where d_i is the interval between the $(I - 1)$-th and

the *i*-th occurrence of *w* and d_1 is the distance between the last and the first occurrence under the assumption that the corpus is periodically repeated. This group of measures was introduced by [Savický, Hlaváčová 2002] and includes the following three measures:

Average Reduced Frequency (*ARF*): $\frac{f}{l}\sum_{i=1}^{f}\min\left(d_i, \frac{l}{f}\right)$

Average Waiting Time Frequency (f_{AWT}): $\frac{l^2}{\sum_{i=1}^{f} d_i^2}$

Average Logarithmic Distance Frequency (f_{ALD}): $e^{-\sum_{i=1}^{f}\frac{d_i}{N}ln\frac{d_i}{N}}$

In spite of the fact that there are so many adjusted frequency measures, a rigorous comparative evaluation of these measures has never been conducted, the most notable exception being a paper by Gries [Gries 2010] that studies correlations between these measures. [Savický, Hlaváčová 2002] compare the stability of their three measures across different corpora and come to a conclusion that f_{AWT} is the least stable, whereas the stability of *ARF* and f_{ALD} depends on how variable the frequency of a word is (f_{ALD} should be used for words with substantial variation).

3. Evaluating adjusted frequency measures

To make an evaluation of frequency measures, we need to find out how well a frequency dictionary compiled from a corpus represents the population from which this corpus was taken. Obviously, we have no access to the population as a whole, but we can check how well the ranking we obtained using some frequency measure on a training set from a corpus describes an unseen sample from the same corpus.

For the experiment, eight corpora of eight languages were taken. They are listed in Table 1.

Each corpus was split into five parts in order to perform 5-fold cross-validation. Four parts were used as training set, and the remaining part served as test set. For the training set, 46 frequency lists were compiled using different frequency measures:

1) **Plain frequency**; **Distance-Based measures;**
2) Average Reduced Frequency (*ARF*);
3) Average Waiting Time Frequency (f_{AWT});
4) Average Logarithmic Distance Frequency (f_{ALD}); **Part-based measures:**

5) Juilland's U;
6) Rosengren's AF;
7) Carroll's U_m;
8) Engvall's measure;
9) Kromer's U_R;
10) Median frequency;
11) Range.

Table 1. **Corpora used for the experiment**

Language	Corpus name	Tokens
Arabic	NYUAD Arabic UD	841,460
Catalan	AnCora	533,150
Czech	Prague Dependency Treebank 3.0	1,509,236
English	Brown	1,161,192
German	Hamburg Dependency Treebank	3,055,010
Icelandic	Icelandic Parsed Historical Corpus	1,016,527
Russian	SynTagRus	1,107,741
Spanish	AnCora	551,456

The ranking in the resulting frequency lists was compared to the plain frequency list of the test set. To perform the comparison, we take those types that were attested in the training set at least five times, and calculate Spearman's rank-order correlation coefficient ρ for the 46 frequency lists obtained from the training set against the plain frequency list of the test set.

For example, if we consider Czech words *být*, *v*, *a*, *se*, *na*, *ten*, *že*, *z*, *s*, *který* and use the first four parts of the Czech corpus as training set and the last one as test set, we get the following rankings on our training data:

ARF: *být*, *v*, *a*, *se*, *na*, *ten*, *z*, *s*, *který*, *že*;
f_{AWT}: *být*, *v*, *a*, *se*, *na*, *s*, *který*, *z*, *ten*, *že*.
The same types in the test set are ordered as follows:
Test: *být*, *v*, *a*, *se*, *na*, *že*, *ten*, *z*, *s*, *který*.

In this case, ρ(*ARF*, Test) = 0.88 and ρ(f_{AWT}, Test) = 0.77, which shows that *ARF* is better at predicting the actual ranking of these words in the unseen

data that f_{AWT}. The actual calculation in this case was based on 15,124 lemmas that occurred in the training set at least five times (out of 51,703 lemma types in total), and *ARF* also outperformed f_{AWT}, the two measures scoring 0.745 and 0.736 respectively.

Each of the seven part-based measures was computed using 5, 10, 20, 50, 100, and 200 equally-sized parts. It is worth establishing the best number of parts for each measure and then proceed with comparing 11 rather than 46 measures. For each measure, we find with what number of parts this measure performs better than with any other number of parts (i. e., gets a higher ρ in 21 or more cases out of 8 × 5 = 40). Table 2 illustrates this for Juilland's *U*.

Table 2. **Pairwise comparison of Juilland's *U* with different number of parts**

Number of parts	5	10	20	50	100	200
5		5	6	10	20	22
10	35		15	19	23	26
20	34	25		28	30	32
50	30	21	12		35	37
100	20	17	10	5		35
200	18	14	8	3	5	

For instance, this table shows that Juilland's *U* with 10 parts outperforms Juilland's *U* with 100 parts 23 times out of 40, but it outperforms Juilland's *U* with 20 parts only 15 times out of 40. The table makes clear that the best number of parts for Juilland's *U* is 20, because all five numbers in the corresponding row are greater than 20.

The optimal numbers of parts for part-based measures are as follows:

Juilland's U	20	Engvall	50	Range	100
Rosengren's AF	50	Kromer's UR	50		
Carroll's U_m 5	0	Median	5		

Further comparison between the 11 measures follows the same procedure. The results are shown in Table 4.

Table 4. **Pairwise comparison of 11 frequency measures**

Rank	Measure	Range-100	*ARF*	Kromer's U_R-50	Rosengren's *AF*-50	Juilland's *U*-20	Carroll's U_m-50	f_{ALD}	Engvall-50	f_{AWT}	Median-5	Plain
1	Range-100	0	23	26	29	29	29	34	29	37	38	40
2	*ARF*	17	0	25	27	30	29	31	27	38	40	40
3	Kromer's U_R-50	14	15	0	34	34	35	25	37	36	37	40
4	Rosengren's *AF*-50	11	13	6	0	24	29	22	36	35	37	40
5	Juilland's *U*-20	11	10	6	16	0	21	23	22	36	37	40
6	Carroll's U_m-50	11	11	5	11	19	0	21	22	34	37	40
7–8	f_{ALD}	6	9	15	18	17	19	0	20	38	33	38
7–8	Engvall-50	11	13	3	4	18	18	20	0	34	37	40
9–10	f_{AWT}	3	2	4	5	4	6	2	6	0	20	33
9–10	Median-5	2	0	3	3	3	3	7	3	20	0	38
11	Plain	0	0	0	0	0	0	2	0	7	2	0

4. Discussion

The results presented in Table 4 are to some extent surprising. Many years of research into adjusted frequencies have given rise to many sophisticated adjusted frequency measures, but the best measure turns out to be a very simple one, namely Range with Plain Frequency used as tie breaker. The superiority of *ARF* to other distance-based measures was already hinted at by [Savický, Hlaváčová 2002] in the paper where these measures were introduced, and it is confirmed by our experiment. A high place occupied by Kromer's U_R is also worth noting. [Gries 2008] criticizes this measure and says that Kromer's claim [Kromer 2003] to its psycholinguistic appropriateness is not corroborated by any evidence. However, the fact that this measure is so good at predicting unseen data speaks in its favor. The criticisms against Juilland's *D* as a dispersion measure [Biber et al. 2016] are supported by the fact that Juilland's *U*, which is based on *D*, does not turn out to be among the best adjusted frequency measures.

Obviously, Range has some drawbacks that prevent one from recommending it as an ultimate adjusted frequency measure. Our experiment shows that it is good for ranking purposes, but it is not the best measure if we need to estimate not only the ranking, but also frequencies, e.g. for keyword extraction. *ARF* is much better suited to this purpose, and it is probably not surprising that this measure is used in the Czech National Corpus as well as in SketchEngine rather than Range. The inferiority of other measures is not as dramatic as it may seem; in fact, even using Plain Frequency results in reasonably good frequency lists.

There are also some limitations of the experiment that I must address. First, it is not clear how the performance of adjusted frequency measures relates to corpus size; this issue requires further investigation. Second, the experiment does not tell anything about adjusted frequency measures in case where a training corpus is divided into different-sized categories. Third, the order of the texts in a corpus may be of critical importance for adjusted frequency measures, and this issue was not touched upon. However, these concerns do not undermine the validity of the experiment as a whole. One may conclude that Range combined with Plain Frequency as the next level of sorting provides the best ranking, which makes it questionable whether other part-based distance measures are actually needed for lexicographic purposes. As for distance-based measures, *ARF* fares almost as well as Range, and it is advisable to use this measure in cases where not only rankings, but also frequency estimates are required.

References

1. *Aksan Y., Aksan M., Mersinli Ü., Demirhan U. U.* (2017), A frequency dictionary of Turkish: Core vocabulary for learners. Routledge, Taylor & Francis Group.
2. *Biber D.* (1993), Representativeness in corpus design. In: Literary and Linguistic Computing. Vol. 8(4), pp. 243–257. https://doi.org/10.1093/llc/8.4.243
3. *Biber D., Reppen R., Schnur E., Ghanem R.* (2016), On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. In: International Journal of Corpus Linguistics. Vol. 21(4), pp. 439–464. https://doi.org/10.1075/ijcl.21.4.01bib
4. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python. O'Reilly Media, Inc.
5. *Carroll J. B.* (1970), An Alternative to Juilland's Usage Coefficient for Lexical Frequencies. In: ETS Research Bulletin Series, pp. 1–15. https://doi.org/10.1002/j.2333-8504.1970.tb00778.x
6. *Davies M., Preto-Bay A. M. R.* (2008), A frequency dictionary of Portuguese: Core vocabulary for learners. Routledge.

7. *Gries S. Th.* (2008), Dispersions and adjusted frequencies in corpora. In: International Journal of Corpus Linguistics. Vol. 13(4), pp. 403–437. https://doi.org/10.1075/ijcl.13.4.02gri
8. *Gries S. Th.* (2010), Dispersions and adjusted frequencies in corpora: Further explorations. In: S. Th. Gries, S. Wulff, M. Davies (eds.). Corpus-linguistic applications, pp. 197–212. Brill. https://doi.org/10.1163/9789042028012_014
9. *Jurafsky D., Martin J. H.* (2009), Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. Pearson Prentice Hall.
10. *Kilgarriff A.* (1997), Putting frequencies in the dictionary. In: International Journal of Lexicography. Vol. 10(2), pp. 135–155. https://doi.org/10.1093/ijl/10.2.135
11. *Kromer V.* (2003), A usage measure based on psychophysical relations. In: Journal of Quantitative Linguistics. Vol. 10(2), pp. 177–186. https://doi.org/10.1076/jqul.10.2.177.16718
12. *Savický P., Hlaváčová J.* (2002), Measures of word commonness. In: Journal of Quantitative Linguistics. Vol. 9(3), pp. 215–231. https://doi.org/10.1076/jqul.9.3.215.14124
13. *Sharoff S., Umanskaya E., Wilson J.* (2013), A frequency dictionary of Russian: Core vocabulary for learners. Routledge.

Alexander Piperski
Russian State University for the Humanities / HSE University (Russia)
E-mail: apiperski@gmail.com

Н.Л. Аванесян, А.М. Чеповский, Т.Ю. Шерстинова,
Ф.Н. Соловьев, Д.Ю. Чуйкин

N.L. Avanesyan, A.M. Chepovskiy, T. Yu. Sherstinova,
F.N. Soloviev, D. Yu. Chuikin

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ЧАСТОТНЫХ СЛОВАРЕЙ ЛИНГВИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ РУССКОЙ ПРОЗЫ 1900–1930 гг. В ДИНАМИКЕ[1]

CORRELATION ANALYSIS OF FREQUENCY DICTIONARIES OF LINGUISTIC INDICATORS OF RUSSIAN PROSE 1900–1930 IN DYNAMICS

Аннотация. Проведен сравнительный анализ текстовых подкорпусов четырех периодов русской литературы 1900–1930 гг. с использованием коэффициентов попарной ранговой корреляции частотных словарей лексических характеристик. Показано, что социальные катаклизмы в меньшей степени повлияли на употребление глаголов, чем прилагательных и существительных. Установлено, что другими показателями, отличающими рассматриваемые временные подкорпусы художественной прозы, являются именные и глагольные группы, а также частотные последовательности символов.
Ключевые слова. Ранговая корреляция, корпусная лингвистика, именные группы (ИГ), глагольные группы (ГГ), литературные тексты.

Abstract. The paper describes a comparative analysis of text subcorpuses referring to four periods of Russian literature 1900–1930, which was carried out using the coefficients of pairwise rank correlation of frequency dictionaries of different lexical characteristics. It is shown that social cataclysms influenced the use of verbs to a lesser extent than that of adjectives and nouns. It has been also established that other indicators that distinguish the considered temporal subcorpuses of fiction are noun and verb groups, as well as frequency sequences of symbols.
Keywords. Rank correlation, corpus linguistics, noun phrases (NP), verb phrases (VP), literary texts.

1. Введение

Корреляционный анализ традиционно используется в качестве меры, позволяющей количественно оценить связь между двумя непрерывными переменными. Предлагаемое исследование направлено на изучение различных исторических периодов русской прозы 1900–1930 гг., а именно на сопоставление трех последовательных периодов

[1] Создание корпуса русских рассказов поддержано грантом РФФИ № 17-29-09173 офи_м, описанная статистика получена в рамках гранта РФФИ № 19-07-00806.

русской литературы. Оно основано на корреляционном анализе количественных текстовых показателей, представленных в виде частотных словарей различных лингвистических данных (в частности, частей речи (ЧР), именных и глагольных групп, псевдооснов слов и т. д.). Целью исследования является оценка эффективности этих количественных методов анализа корпуса художественных текстов.

Материалом послужил корпус русских рассказов первой трети XX века, разрабатываемый на филологическом факультете СПбГУ при участии департамента филологии НИУ ВШЭ в Санкт-Петербурге [Мартыненко и др., 2018; 2018а; Martynenko, Sherstinova, 2019; Sherstinova et al., 2020]. Корпус предназначен для проведения стилеметрических исследований русской прозы, а также для изучения изменений, произошедших в русском языке в эпоху революций.

Литературные произведения как нельзя лучше отражают события и изменения во всех сферах жизни человека: в культурной, политической, социальной и даже бытовой. В отличие от традиционных корпусов, ориентированных на тексты, написанные выдающимися писателями, корпус русских рассказов 1900–1930 гг. предназначен для изучения языка всей литературной системы рассматриваемой эпохи, поэтому в него включены рассказы, написанные многими забытыми и второстепенными писателями. Таким образом, он позволяет в полной мере использовать весь потенциал художественной литературы для исследований в области корпусной лингвистики [Martynenko, Sherstinova, 2019].

Для аннотированной части корпуса было отобрано 300 российских писателей из общего списка авторов, насчитывающего более 2800 персоналий. Общий объем аннотированной части составляет 310 рассказов, всего 1 млн словоупотреблений. Корпус разделен на 4 временных периода: 1) начало XX века до 1913 г., 2) Первая мировая война до революций (1914–1916), 3) революционные годы (1917–1922) и 4) раннесоветский период (1923–1930). Корпус охватывает рассказы представительного числа авторов для каждого периода времени, что позволяет проводить лингвистический и статистический анализ языка и стиля.

В работе [Лаврентьев и др., 2020] для исследования корпуса русских рассказов был использован метод сравнительного анализа корпусов текстов, который позволяет выявить неявные связи между корпусами разнородных текстов. В основе данной методики лежат методы корпусного анализа, реализованные в корпусной платформе ТХМ: анализ специфичности, позволяющий создать своеобразный «профиль»

подкорпуса на основе определенного свойства (тематика текста, психологическая направленность текста) путем выявления наиболее характерных или нехарактерных для него признаков (начальных форм слова, псевдооснов). В данном исследовании мы применяем метод ранговой корреляции для сравнения частотных словарей различных лексических характеристик подкорпусов.

Представленная методология корпусного анализа может быть использована для сравнения любых текстовых репрезентативных наборов данных.

2. Метод. Техники анализа текста

Характеристики текстов в данном исследовании определялись процедурами автоматизированной обработки текстов на естественных языках, описанными в [Чеповский, 2015; Соловьев, 2020].

Автоматический морфологический анализ словоформ проводился на основе компьютерных методов морфологии. Используемая для анализа морфологическая модель [Чеповский, 2015] относит каждое слово к одному из 24 морфологических классов, которые помимо «традиционных» частей речи включают такие категории, как «неизменяемое слово», «аббревиатура», «топоним» и т. д. Каждый из этих морфологических классов характеризуется набором грамматических характеристик: род, падеж, число, наклонение и т. д. Каждая словоформа содержит свои грамматические характеристики и свою каноническую (начальную) форму.

Кроме того, в тексте выделялись именные и глагольные группы слов. Под именной группой (ИГ) мы понимаем группу слов, в которой главным словом является существительное, а другие слова связаны с ним подчиненными синтаксическими связями. При выявлении именных групп решалась задача снятия омонимической неопределенности, возникающая в результате множественности морфологических анализов отдельных употреблений слов. Методика определения именных групп основана на рассмотрении всего набора возможных морфологических интерпретаций каждого слова.

Глагольные группы (ГГ) — это словосочетания, основным словом которых является глагол. Связи именных групп с глаголами построены на синтаксическом анализе предложения. Глагольное управление определяется как разновидность синтаксической подчинительной связи типа управления, в которой главным словом является глагол.

Анализируя глагольное управление основного слова (глагола), накладываются ограничения на употребление зависимого словосочетания в виде набора вариантов допустимых комбинаций грамматических характеристик зависимого словосочетания. Анализ глагольного употребления основан на датировке электронного словаря глагольного управления, который включает первые две тысячи наиболее часто встречающихся глаголов русского языка. В отличие от отдельных слов, выделенные именные и глагольные группы несут информацию о конкретных аспектах текстового содержания.

В качестве еще одной лингвистической характеристики текста рассматривалась псевдооснова, т. е. часть слова, не содержащая ни суффиксов, ни префиксов. Метод автоматического выбора псевдооснов заключается в сравнении рассматриваемой словоформы с допустимым в языке набором структур некорневой части слова [Чеповский, 2015]. Псевдооснова слова выделяется отбрасыванием всех аффиксов, соответствующих определенной структурной схеме, которые описывают максимальное количество комбинаций префиксов и суффиксов, разрешенное в данном языке. Метод псевдооснов позволяет анализировать текстовые структуры, а не только точные словоформы.

Сравнительный анализ подкорпусов проводился преимущественно путем попарного сравнения частотных словарей различных лексических характеристик, составленных для исследуемых подкорпусов. Чтобы оценить близость частотных словарей, ранги записей словаря устанавливаются после сортировки характеристики, внесенной в словарь, по частоте встречаемости. Сравнение словарей производится путем расчета коэффициента попарной ранговой корреляции для каждой пары словарей разных подкорпусов. Словарные записи считаются случайными величинами. Связь между наборами таких элементов различных словарей определяется как коэффициент попарной ранговой корреляции значений этих случайных величин.

Коэффициенты попарной ранговой корреляции для любых двух частотных словарей определялись с учетом элементов с одинаковыми частотами. Поскольку размеры словарей могут быть довольно большими, учитывались только первые (в порядке убывания частоты) 10 000 записей в каждом из словарей. Если какая-то лексическая характеристика встречается только в одном словаре и не встречается в другом, то при вычислении коэффициента корреляции мы принимаем ее частоту во втором словаре равной 0.

Коэффициент попарной ранговой корреляции принимает значения из интервала [–1; 1]. Значения, близкие к 1, указывают на монотонную непротиворечивость словарей: если в одном словаре в паре слов одно из них имеет частоту выше, чем другое, то во втором словаре это слово также имеет частоту выше, чем другое. И это правило действует для всех пар слов. Значения, близкие к 1, указывают на обратный эффект: если в одном частотном словаре какое-то слово имеет более высокую частоту, чем какое-то другое слово, то во втором частотном словаре его частота, напротив, будет ниже; это правило также действует для всех пар слов. Если значение близко к 0, то словари несовместимы: ранговой связи между частотами слов в двух словарях нет.

3. Сравнительный анализ частотных словарей

Сравнительный анализ хронологических подкорпусов корпуса рассказов был проведен с использованием частотных словарей лексических характеристик, полученных методами, описанными в предыдущем разделе. Методами автоматического анализа были составлены частотные словари частей речи (существительные, глаголы, прилагательные) и словосочетаний (именные и глагольные группы), размеры которых составляют от 10 000 до 60 000.

Общая характеристика частотных словарей частей для изучаемого временного интервала приведена в статье [Sherstinova et al., 2020]. Результаты сравнения частотных словарей существительных, прилагательных и глаголов с помощью коэффициента попарной ранговой корреляции показывают, что с точки зрения сравнения рангов слов различных частей речи в разных подкорпусах исследуемого корпуса художественных текстов наблюдается слабая положительная ранговая корреляция для существительных (0,4–0,5) и несколько более выраженная корреляция между рангами прилагательных (0,5–0,6) и глаголов (0,6–0,7). Таким образом, можно сделать вывод, что социальные катаклизмы в меньшей степени повлияли на частоту употребления глаголов, в то время как существительные показывают большие изменения. Это подтверждается выводами, сделанными ранее на этом материале [там же].

Интересно, что значение коэффициента ранговой корреляции в случае именных и глагольных групп достаточно мало и находится в интервале [–0,1; 0,5]. Для именных групп его значение относительно стабильно на уровне слабой положительной корреляции (0,2–0,4), в то

время как корреляция между глагольными группами отдельных периодов фактически нулевая (минимум наблюдается для двух групп значений — между советским периодом и началом века, а также между военным и революционным периодом). Интерпретация этих данных требует более детального анализа. Однако можно предположить, что это указывает на возможность определения конкретных словосочетаний для каждого подкорпуса и возможность рассмотрения словосочетаний в качестве отличительных признаков при идентификации рассказов для определенного исторического периода, особенно для глагольных групп.

Сравнение частотных словарей буквосочетаний проводилось для последовательностей из 3–6 букв. Словари буквенных сочетаний длиной 3 практически совпадают по метрике ранговой корреляции, что вполне естественно и ожидаемо: трехбуквенные сочетания характеризуют язык и все подкорпуса русского литературного языка.

Полученные результаты подтверждают утверждение: частота использования буквосочетаний длиной от 1 до 3 определяет язык, на котором написан текст. Следовательно, ранговые порядки этих частотных словарей близки, что и показывает коэффициент корреляции.

Сравнение словарей сочетаний букв с длиной 4, 5 и 6 букв показало уменьшение совпадения частотных словарей с увеличением длины сочетаний букв. Возможность разделения подкорпусов по содержанию и эмоциональной ориентации на основе сравнения словарей сочетаний букв не подтверждается в явном виде этими экспериментами. Но полученные нами результаты указывают на возможность использования длинных буквенных комбинаций для обозначения подкорпусов рассказов.

4. Заключение

Сравнительный анализ подкорпусов текстов проводился с использованием методики расчета коэффициентов попарной ранговой корреляции частотных словарей, полученных для различных лексических характеристик текста (для отдельных частей речи, именных групп, глагольных групп, n-грамм на уровне букв).

На основе полученных данных было показано, что социальные катаклизмы рассматриваемого исторического периода в меньшей степени повлияли на частоту употребления глаголов, в то время как прилагательные и особенно существительные показывают большие

изменения. Далее было установлено, что словосочетания (особенно глагольные группы) в большей степени, чем отдельные части речи, могут быть использованы для сравнительного анализа и идентификации кратких литературных текстов для данных исторических периодов. Важным наблюдением является то, что наиболее показательными характеристиками, отличающими разные подкорпусы текстов, являются частотные последовательности символов.

Литература

1. *Лаврентьев А. М., Рябова Д. М., Тихомирова Е. А., Фокина А. И., Чеповский А. М., Шерстинова Т. Ю.* (2020), Сравнительный анализ специальных корпусов текстов для задач безопасности. Вопросы кибербезопасности. № 3(37), с. 58–65. DOI: 10.681/2311-3456-2020-03-58-65.
2. *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.
3. *Мартыненко Г. Я., Шерстинова Т. Ю., Мельник А. Г., Попова Т. И.* (2018), 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 Международной объединенной конференции «Интернет и современное общество», IMS-2018, Санкт-Петербург, 30 мая — 2 июня 2018 г. Сборник научных статей. СПб: Университет ИТМО, с. 99–104.
4. *Соловьев Ф. Н.* (2020), Автоматическая обработка текстов на основе платформы TXM с учетом анализа структурных единиц текста. Вестник НГУ. Серия: Информационные технологии. Т. 18, № 1, с. 74–82. DOI 10.25205/1818-7900-2020-18-1-74-82
5. *Чеповский А. М.* (2015), Информационные модели в задачах обработки текстов на естественных языках. Издание второе, переработанное. М.: Национальный открытый университет «ИНТУИТ».
6. *Martynenko G. Y., Sherstinova T. Y.* (2019), Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, PRLEAL-2019, Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552, pp. 105–120.
7. *Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Gukasian M., Egoshina I., Turygina M.* (2020), Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900–1930). In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy, pp. 366–373. URL: https://fruct.org/publications/acm27/files/She.pdf (дата обращения: 01.07.2021).

8. *Sherstinova T. Yu, Ushakova E. O., Melnik A. G.* (2020), Measures of Syntactic Complexity and their Change over Time (the Case of Russian). In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy, pp. 221–229.

References

1. *Martynenko G., Sherstinova T., Melnik A., Popova T., Zamirailova E.* (2018), O principakh 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 20th century]. In: Trudy XV Mezhdunarodnoy konferentsii po komp'yuternoy i kognitivnoy lingvistike «TEL 2018». [Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics "TEL 2018"]. Kazan, pp. 180–197.
2. *Martynenko G. Ya., Sherstinova T. Yu., Melnik A. G., Popova T. I.* (2018), Metodologicheskie problemy sozdaniya Komp'yuternoj antologii russkogo rasskaza kak yazykovogo resursa dlya issledovaniya yazyka i stilya russkoj khudozhestvennoj prozy v ehpokhu revolyucionnykh peremen (pervoj treti XX veka) [Methodological problems of the creation of the computer anthology of Russian short stories as a language resource designed to study the language and style of Russian fiction in the era of revolutionary changes (in the first third of the 20th century)]. In: Komp'yuternaya lingvistika i vychislitel'nyye ontologii. Vyp. 2. Trudy XXI Mezhdunarodnoy ob"yedinennoy konferentsii «Internet i sovremennoye obshchestvo», IMS-2018,. SPb: Universitet ITMO Computational linguistics and computational ontologies. Issue 2. [Proc. of the XXI Int. United Conf. The Internet and Modern Society, IMS-2018] St. Petersburg, pp. 99–104.
3. *Martynenko G. Y., Sherstinova T. Y.* (2019), Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, PRLEAL-2019, St. Petersburg, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552, pp. 105–120.
4. *Sherstinova T. Yu., Ushakova E. O., Melnik A. G.* (2020), Measures of Syntactic Complexity and their Change over Time (the Case of Russian). In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy, pp. 221–229.
5. *Lavrentyev A. M., Raybova D. M., Tikhomirova E. A., Fokina A. I., Chepovskiy A. M., Sherstinova T. Yu.* (2020), Sravnitelniy analis specialnikh korpusov tekstov dlay zadach bezopasnosty [Comparative analysis of special text corpora for security-related tasks]. In: Voprosi kiberbezopasnosti [Cybersecurity issues], No. 3(37), pp. 58–65. DOI: 10.681/2311-3456-2020-03-58-65
6. *Solovev F. N.* (2020), Avtomaticheskaya obrabotka tekstov na osnove platformy TXM s uchetom analiza strukturnykh yedinits teksta [Embedding Additional Natural Language Processing Tools into the TXM Platform]. In: Vestnik NGU. Seriya: Informatsionnyye tekhnologii [Vestnik NSU. Series: Information Technologies]. Vol. 18, No. 1, pp. 74–82.
7. *Chepovskiy A. M.* (2015), Informatsionnyye modeli v zadachah obrabotki tekstov na yestestvennyh yazykah. Izdaniye vtoroye, pererabotannoye [Information models in NLP tasks]. Moscow: Natsional'nyy otkrytyy universitet "INTUIT" [National Open University "INTUIT"].

8. *Sherstinova T., Grebennikov A., Skrebtsova T., Guseva A., Gukasian M., Egoshina I., Turygina M.* (2020), Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900–1930). In: 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy, pp. 366–373. URL: https://fruct.org/publications/acm27/files/She.pdf (date of access: 01.07.2021).

Аванесян Нина Левоновна
Национальный исследовательский университет
«Высшая школа экономики» (НИУ ВШЭ) (Россия)
Avanesyan Nina
National Research University Higher School of Economics (Russia)
E-mail: nlavanesyan@edu.hse.ru

Чеповский Андрей Михайлович
Российский университет дружбы народов (РУДН) (Россия)
Национальный исследовательский университет
«Высшая школа экономики» (НИУ ВШЭ) (Россия)
Chepovskiy Andrey
Peoples' Friendship University of Russia (RUDN University), (Russia)
National Research University Higher School of Economics (Russia)
E-mail: chepovskiy-am@rudn.ru

Шерстинова Татьяна Юрьевна
Национальный исследовательский университет
«Высшая школа экономики» (НИУ ВШЭ) (Россия)
Санкт-Петербургский государственный университет (Россия)
Sherstinova Tatiana
National Research University Higher School of Economics, (Russia)
St. Petersburg State University (Russia)
E-mail: tsherstinova@hse.ru

Соловьев Федор Николаевич
Федеральный исследовательский центр
«Информатика и менеджмент» (Россия)
Soloviev Fedor
Federal Research Center "Informatics and Management" (Russia)
E-mail: the0@yandex.ru

Чуйкин Даниил Юрьевич
МИРЭА — Российский технологический университет (РТУ МИРЭА), (Россия)
Chuikin Daniil
MIREA — Russian Technological University (RTU MIREA), (Russia)
E-mail: chuykin.d.y@edu.mirea.ru

Л. Д. Бадмаева

L. D. Badmaeva

ПРОБЛЕМЫ МОРФОЛОГИЧЕСКОГО ОПИСАНИЯ ЛЕКСЕМ СТАРОМОНГОЛЬСКИХ ТЕКСТОВ[1]

PROBLEMS OF MORPHOLOGICAL DESCRIPTION OF LEXEMES OF THE OLD-MONGOLIAN TEXTS

Аннотация. Бурятские летописи, составленные на старомонгольской письменности более ста лет назад, получают свое второе рождение в виде публикаций в романизированном виде, что позволяет формировать на их текстовой базе корпусный ресурс. В работе показаны предварительные разработки для формализованного описания словоизменения старомонгольского языка для нужд морфологического аннотирования, созданных на нем летописных текстов бурят.
Ключевые слова. Бурятский язык, старописьменные летописи, корпус, морфологическое описание.

Abstract. The Buryat chronicles, compiled in the Old-Mongolian Writing more than a hundred years ago, are given their rebirth in the form of publications in a Romanized form from the beginning of the 3rd millennium, which makes it possible to form a corpus resource on their text base, which will multiply the directions of their research by various specialists, primarily linguists. The work will show preliminary developments for a formalized description of the inflection of the Old-Mongolian language for the needs of morphological annotation of the chronicle texts of Buryats created on it. The work performed should serve the formation of a grammatical dictionary of the Old-Mongolian language, the development of a morphological analyzer for texts on this language, as well as a corpus search system for language data.
Keywords. The Buryat language, old-written chronicles, corpus, morphological description.

1. Введение

Труды по проблемам разметки и грамматического описания корпусных материалов освещают различные их аспекты. В российской корпусной лингвистике можно отметить работы А. А. Пичхадзе, С. О. Савчук, С. А. Крылова, А. Е. Полякова, Д. В. Сичинавы, М. А. Даниеля, Т. А. Архангельского, В. В. Кукановой и др. Коллектив разработчиков диахронического корпуса бурятского языка полагается на опыт проведения разметки текстов, излагающийся вышеперечисленными авторами. В начальной стадии формирования корпуса по тому или иному языку в первую очередь проводятся работы по сбору, составлению различных баз данных, как текстовых, так и сугубо лингвистиче-

[1] Работа выполнена в рамках государственного задания (проект «Мир человека в монгольских языках: анализ средств выражения эмотивности, № 121031000258-9»).

ских. Последние бывают тесно связаны с теоретическими описаниями самого языка, создаваемого ресурса, на основе которых далее строятся программные средства [Орехов 2014: 135]. В работе по проблемам морфологической разметки, напр., башкирских текстов, справедливо указывается, что «в основе любого автоматического анализа морфологии лежат две составляющие: грамматический словарь и формализованное описание словоизменительной системы языка» [Там же].

Целью данной работы является освещение проблем, связанных с подготовительными работами для проведения разметки при составлении лингвистического диахронического корпуса на материале старописьменных монгольских текстов. Задачами проведенного первичного этапа являлись выбор и подготовка текстовых источников корпуса, формирование частотного словаря на их лексическом материале (выполнено Ринчиновым О. С.), разработка грамматических помет (тегов) для описания словоизменительной системы старомонгольского языка, морфологическое описание лексем указанного словаря, выявление особенностей плана выражения элементов словника частотного словаря.

Для организации названного корпуса выбраны пять транслитерированных электронных версий летописных бурятских текстов, опубликованных на CD-ROM: Летопись баргузинских бурят; Хроника Вандана Юмсунова 1875 г. Летописи хоринских бурят; Хроника Тугултур Тобоева 1863 г. Летописи хоринских бурят; Летописи селенгинских бурят. Хроника Убаши Дамби Джалцан Ломбо Цэрэнова, 1868 г. [Бадмаева, 2009] и Летопись хоринских бурят. Хроника Шираб-Нимбо Хобитуева в [Бадмаева и др. 2018: 139–188]. Оригиналы данных летописей создавались задолго до начала кириллического периода функционирования письменности у бурят, в пору бытования у них старописьменного монгольского языка. Письменная графика названного языка использовалась бурятами официально до ее отмены в 1931 г. В научный оборот данные тексты введены в 1-й половине XX века А. И. Востриковым, Н. Н. Поппе и В. А. Казакевичем.

2. Основы и некоторые особенности морфологического описания старомонгольских текстов

В основе грамматического описания старомонгольского языка лежит, в первую очередь, работа Н. Н. Поппе [Поппе 1935]. Автором использованы почти все основные труды классиков монголоведения

XIX в. — начала XX в., представляющие собой не только грамматические, но художественные сочинения, а также прессу его времени. Сама работа Н. Н. Поппе состоит из введения и трех частей: Фонетика и письмо, Морфология и Синтаксис. Для составления аннотации, представляющей собой таблицу помет/тегов с дальнейшим их толкованием в общепринятой лингвистической терминологии, использованы материалы второй части, Морфологии, причем это материалы сугубо старомонгольского словоизменения. Кроме названных материалов использованы в небольшом объеме материалы из третьей части книги, Синтаксиса, посвященные послелогам, которые сам автор называет словами, выполняющими служебную функцию [Поппе 1935: 164].

Если сравнить описания послелогов у Н. Н. Поппе с их описаниями в грамматических трудах бурятского языка, то в-последних — они получают описание в морфологическом разделе, называясь служебными словами или служебной частью речи. Поскольку мы озадачены описанием морфологической информации старописьменного монгольского языка и опираемся на конкретную работу Н. Н. Поппе, в нашей разработанной системе помет для морфологической разметки послелоги обозначены как элементы синтаксиса и получают соответствующий гипероним SYNT (Syntax), т. е. обозначение своей принадлежности к синтаксису. Из синтаксического раздела книги Н. Н. Поппе в нашу аннотационную таблицу включены знаки препинания[2] в старомонгольском обозначении, такие как *čeg* — «точка», использующаяся в нем в качестве запятой/обозначения конца предложения, *dörbelǰin čeg* — «четыре точки» для обозначения конца главы, крупного раздела или строф стиха, *dabqur čeg* — «двойные/две точки» как запятая/конец предложения. Как было обозначено выше, на данном этапе подготовки лингвистических материалов для программных инструментариев дихронического корпуса бурятского языка в них не входят материалы старомонгольского словообразования.

Техническое оформление самой системы помет выполнено в таблице Excel (она доступна в формате PDF на, разрабатываемом нами, коллективом составителей, сайте диахронического корпуса бурятского

[2] В системе латинской транслитерации, общепринятой в международном монголоведении, знаки препинания вертикальной старомонгольской графики при ее романизации передаются в виде запятой или двоеточий, в зависимости от их количества.

языка: http://annals.imbtarchive.ru/[3]). В таблице помет отражается в основном идентичная последовательность представления грамматических категорий, наблюдающаяся в книге Н. Н. Поппе. Данная таблица содержит несколько полей (в скобках даны обозначения полей с примерами): A (тег: POS), B (русский термин: Часть речи), C (английский термин: Part of speech), D (гипероним: MORPH, Morphology). В общей сложности сформулированы 122 тега, включая три вида знаков препинания старомонгольского текста. В данной таблице отражены такие грамматические категории старомонгольского языка, как части речи, склонение (падеж, двойной падеж), притяжательное склонение, число, время, лицо, наклонение, спряжение.

В таблице выделена отдельная группа существительных, состоящая из семи видов, соответственно со своими пометами: имя собственное; личное имя; фамилия; отчество; топоним; организация; существительные, выражающие одушевленность/неодушевленность; аббревиатуры. Следует сказать, что состав видов существительных, входящих в названную выше группу, продиктован соответствующим лексическим материалом летописных текстов, включаемых в разрабатываемый корпус, напр., *ilaysangdarovisi* «Александрович» — отчество, *erküü* — «Иркутск», *eǰen,* «владелец, хозяин» — одушевленное имя существительное, *dalai,* «море» — неодушевленное имя существительное.

Внутри каждой грамматической категории даны при необходимости дополнительные пометы, представляющие их специфику, напр., помета NG (Noun groups, Группы существительных) и помета DC (Double case, Двойной падеж) имеют свои дополнительные пометы отдельных подгрупп существительных и конкретных двойных падежей соответственно. Примерами могут быть следующие: PROPN — имя собственное — Proper name, PERSN — личное имя — Person name, ORGN — организация — Organization, COM.INSTR — совместно-орудный падеж — Comitative.Instrumental, COM.ACC — совместно-винительный — Comitative.Accusative.

Таким образом, сформулированы пометы по всем частям речи, таким как существительные (N — Noun), прилагательные (A — Adjective), глаголы (V — Verb), наречия (ADV — Adverb), числительные (NUM — Numeral), местоимения (PRON — Pronoun), частицы (PCL — Particle), союзы (CONJ — Conjunction), междометия (INTRJ — Interjection),

[3] Разработчик сайта — О. С. Ринчинов.

послелоги (PSTP — Postposition)[4], фиксирующимся в бурятских летописях. Пометы имеют свои соответствия на русском и английском языках, напр., по числительным, местоимениям, числу, частицам, глагольным формам (финитным, нефинитным), послелогам и т. д. Данная аннотация (пометы, их соответствия) полагалась на общепринятые традиции в соответствии с Лейпцигскими правилами глоссирования, лингвистическим словарем О. С. Ахмановой, а также с учетом, уже использованных и представленных в соответствующих трудах. В случае отсутствия в них помет для специфических словоформ старомонгольской письменности формулировались новые в соответствии с названиями/обозначениями, принятыми в монголистике.

Разработанная таблица тегов/помет используется для приписывания грамматических характеристик лексемам старомонгольского языка, которое производится вручную на данном этапе (в перспективе предусмотрено решение задачи по разработке инструментария для автоматической разметки старомонгольских транслитерированных/романизированных текстов). Для данной процедуры предварительно составленный частотный словарь словоформ включает 10,3 тыс. языковых единиц по пяти вышеназванным опубликованным летописям бурят, в которых текст наряду с оригинальной старомонгольской версией представлен в латинской транслитерации. Лексико-грамматическое описание словоформ составленного словаря выполнено в следующей последовательности признаков: словоформа — частотность — лемма — русский перевод — часть речи — одушевленность/неодушевленность — грамматические характеристики. Примерами могут быть следующие: *alba* — 60 — *alba* — дань, повинность; обязанность, служба — N (существительное) — INANIM (неодушевленность) — NOM (именительный падеж)[5]; *barayun* — 53 — *barayun* — правый — A (прилагательное); *basa* — 52 — *basa* — опять, снова — ADV (наречие); *degürgebe* — 1 — *degürgekü* — наполнить; заканчивать — V (глагол) — IND.PST1 (изъявительное наклонение. Прошедшее время 1), CAUS (побудительный залог); *dergede* — 53 — *dergede* — возле, около — PSTP (послелог); *ǰayun* — 150 — *ǰayun* — сто — NUM (числительное); *ügei* — 271 — *ügei* — нет — PCL (частица); *ba* — 648 — *ba* — *и* —

[4] Выше дано объяснение о разнице описания послелогов в книге Н. Н. Поппе и современных грамматиках бурятского языка.

[5] Описание грамматических характеристик лексем выполнено Аюшеевой М. В. и автором статьи.

CONJ (союз); *bide* — 35 — *bide* — мы — PRON (местоимение) — ANIM (одушевленность) — NOM.

Ручное приписывание грамматических характеристик словоформам названного частотного словаря выявляет их особенности, характерные в той или иной степени для подобных текстов, созданных до устоявшихся норм литературных языков, напр., русского [Гаврилова и др. 2016: 7]. Можно указать особенности лексики текстовой базы данных — вариативность написания иноязычных слов, личных имен, заимствований в виде русизмов (тибетизмы и санскритизмы встречаются, но редко). Совершенно справедливо в вышеприведенной работе указывается, что «чем больше вариантов написания будет учтено при разработке ресурсов для аннотации, тем более полно и точно будут проанализированы тексты» [Гаврилова и др. 2016: 10]. Появление русизмов/интернационализмов в бурятских летописях обусловлено их лингвоспецифичностью, т. е. отсутствием в бурятской культуре в период создания соответствующих текстов тех понятий, которые они выражали.

Примерами вариативности написания иноязычной лексики могут быть следующие. Для слова «императрица» в бурятских летописях на старомонгольской письменности встречается десять различных написаний: *imperatarsa, imperaturca, impiiratarica, impiratarica, impiratariica, impiratrica, impiraturaca, impiraturica, impiraturiici, imperaturiica*; для слова «министр» встречается три написания: *minister, miniister, miniistar*; Санкт-Петербург встречается в четырех написаниях: *sangpitirburge, sangpitirbürge, sangpitirbüürge, sanktapitirbürge*; слово «адъютант» дано в пяти орфографических формах: *adyudanta, adyüdanta, adyutanta, adyütanta, adyutanti*; для имени *Дамбадугар* встречаются два разных написания: *dambaduγar* и *dambadukar*.

Отдельного описания требует написание падежных форм в рассматриваемых летописях бурят. В письменном монгольском языке данные аффиксы пишутся отдельно от основы слова. В транслитерации указанную отдельность выражают дефисы, напр., *morin-u*, форма родительного падежа, «коня». В материале частотного словаря из семи косвенных падежей (Genitive, Dative-Locative, Accusative, Instrumental, Comitative, Ablative, Indefinite) и четырех двойных падежей (Dative.Locative.Ablative — дательно-местно-исходный, Comitative. Instrumental — совместно-инструментальный, Genitive.Dative. Locative — родительно-дательно-местный, Comitative.Accusative — совместно-винительный) старомонгольского языка выявляются слово-

формы в пяти, таких как GEN (родительный п.), DAT-LOC (дательно-местный п.), ABL (исходный п.), INSTR (орудный п.), COM (совместный п.). Во всех выявленных падежных формах наблюдается слитное написание соответствующих аффиксов, напр.: *qadaYIN*[6] «горы», GEN, SG; *yajarA* «на территории», DAT-LOC; *angqanAČA* «от/из начала», ABL; *čaʏaǰaBAR* «законом», INSTR; *nökürTEI* «с другом», COM.

3. Заключение

Диахронический корпус как онлайн-ресурс разрабатывается для использования при исследованиях этнолингвистического, исторического и географического контекста опубликованных старомонгольских памятников бурят в латинской транслитерации. Представляется, что разрабатываемый корпус важен как для изучения истории языка бурят, так и для поиска возможных путей его дальнейшего развития, поскольку его современное состояние общепризнано как сложное в плане сохранности.

Вышеописанные работы должны послужить формированию грамматического словаря старомонгольского языка, разработке морфологического анализатора текстов на данном языке, а также системы корпусного поиска языковых данных. Разработки названных выше словаря, анализатора и системы поиска представляют основные задачи следующего этапа работ составляемого корпуса. Также важными представляются работы по совершенствованию системы помет (тегов) за счет добавления аффиксов словоизменения (склонение, включая диалектное; спряжение), падежных форм старомонгольских местоимений, частиц в вышеописываемую таблицу, по наполнению парадигмы вариативности в лексической системе бурятских летописей, в группах, включающих заимствования, ономастическую, топонимическую и другую специальную лексику, что, в свою очередь, в перспективе послужит семантической разметке корпуса.

При интегрировании размеченного текста в соответствующий корпус появляется возможность быстрого поиска и получения лексического материала при проведении разнообразных лингвистических исследований: анализ контекстов, лексических/статистических/диалектных/архаистичных данных, сочетаемости, синтаксических кон-

[6] В данных примерах аффиксы для наглядности выделены прописными буквами.

струкций и т. д. Посредством разметки совершенствуется грамматический словарь по текстам, базы данных которых включаются в корпус. Создаваемый грамматический словарь по летописным бурятским текстам может послужить формированию исторического словаря бурятского языка.

Работы по диахроническому корпусу бурятского языка требуют своего продолжения, он дополнит такие ресурсы, как Бурятский корпус, Параллельный корпус (бурятский). В перспективе при дальнейшем их общем усовершенствовании они составят основные направления/части Национального корпуса бурятского языка.

Литература

1. *Бадмаева Л. Б.* (ред.) (2009), История бурятской книги: справочно-библиографический CD-ROM. Улан-Удэ.
2. *Бадмаева Л. Б., Очирова Г. Н.* (2018), Летопись Ш.-Н. Хобитуева как памятник письменной культуры бурят. Улан-Удэ.
3. *Гаврилова Т. С., Шалганова Т. А., Ляшевская О. Н.* (2016), К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв. Вестник ПСТГУ. Серия III: Филология. Вып. 2 (47), с. 7–25.
4. *Орехов Б. В.* (2014), Проблемы морфологической разметки башкирских текстов: Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014, с. 135–140.
5. *Поппе Н. Н.* (1937), Грамматика письменно-монгольского языка. М.; Л.

References

1. *Badmaeva L. B.* (ed.) (2009), Istoriya burjatskoy knigi: spravochno-bibliograficheskiy CD-ROM [History of the Buryat book: reference and bibliographic CD-ROM]. Ulan-Ude.
2. *Badmaeva L. B., Ochirova G. N.* (2018), Letopis' Sh.-N. Hobitueva kak pamyatnik pis'mennoy kul'tury buryat [Chronicle of Sh.-N. Khobitueva as a monument of the Buryats' written culture]. Ulan-Ude.
3. *Gavrilova T. S., Shalganova T. A., Lyashevskaya O. N.* (2016), K zadache avtomaticheskoy leksiko-grammaticheskoy razmetki starorusskogo korpusa XV–XVII vv. [On the problem of automatic lexical and grammatical marking of the Old Russian corpus of the 15th–17th centuries]. In: Bulletin of St.TOU. Ser. III: Philology [перевод на английский]. Issue 2(47), pp. 7–25.
4. *Orekhov B. V.* (2014), Problemy morfologicheskoy razmetki bashkirskikh tekstov [Problems of morphological marking of Bashkir texts]. In: Trudy Kazanskoy shkoly po komp'yuternoy i kognitivnoy lingvistike TEL-2014 [Proceedings of the Kazan School on Computer and Cognitive Linguistics TEL-2014]. Kazan': "Fen" RT' Academy of sciences Publ., pp. 135–140.

5. *Poppe N. N.* (1937), Grammatika pis'menno-mongol'skogo yazyka [Written Mongolian grammar]. Moscow — Leningrad.

Бадмаева Любовь Дашинимаевна

ФГБУН Институт монголоведения, буддологии и тибетологии Сибирского отделения РАН (Россия)

Badmaeva Liubov

FSBSI Institute for Mongolian, Buddhist and Tibetan Sciences, Siberian Branch of Russian Academy of Sciences (Russia)

E-mail: ldbadm@gmail.com

В. А. Баранов

V. A. Baranov

ПАРАЛЛЕЛЬНЫЙ КОРПУС СЛАВЯНСКИХ СПИСКОВ ПАРИМЕЙНИКА: МАТЕРИАЛ И ПОСТАНОВКА ЗАДАЧИ[1]

PARALLEL TEXT CORPUS OF SLAVIC MANUSCRIPTS OF PROPHETOLOGIUM: MATERIAL AND STATEMENT OF THE PROBLEM

Аннотация. Дается общая характеристика четырех рукописей славянского паримейника, транскрипции которых размечаются для демонстрации параллельных чтений. Описываются возможности модуля параллельных корпусов системы «Манускрипт» (manuscripts.ru). Показано несоответствие модели базы данных корпуса и состава богослужебных циклов паримий и предложено решение проблемы. Сформулированы требования к дополнительным формам демонстрации материалов корпуса для сопоставления текстологических особенности рукописей.
Ключевые слова. Параллельный корпус, средневековые славянские рукописи, паримейник.

Abstract. The characteristics of the four manuscripts of the Slavic prophetologium, which are marked up to demonstrate parallel readings, are presented. The capabilities of the parallel corpora module of the "Manuscript" system (manuscripts.ru) are described. The discrepancy between the model of the corpus database and the composition of the liturgical cycles of the parimias is shown, and a solution to the problem is proposed. The task of pointing out the repeating parimias and simultaneously preserving their continuous numbering was solved using a special markup. Requirements for additional forms of visualization of parallel readings for the comparison of textological features of manuscripts are formulated.
Keywords. Parallel text corpus, Slavic medieval manuscripts, prophetologium.

1. Параллельный корпус как способ демонстрации списков средневекового текста

Исторический корпус «Манускрипт» (manuscripts.ru), содержащий транскрипции средневековых славянских кодексов и отрывков X–XV вв., предоставляет пользователям различные средства просмотра (близкие к оригинальным и транслитерированные современной кириллицей или латиницей тексты, прямые и обратные формо- и словоуказатели, однотекстовые и многотекстовые конкордансы и др.) и анализа (количественные и статистические сведения о лингвистических единицах и их сочетаемости) данных.

Одним из способов визуализации некоторых транскрипций корпуса является демонстрация соответствующих друг другу фрагментов

[1] Работа выполняется при поддержке Российского фонда фундаментальных исследований (РФФИ), проект № 20-512-18001.

списков одного текста, например, Евангелий[2], майских служебных миней[3], летописей[4]. Работа с параллельными корпусами (поиск и выбор рукописей, режимы демонстрации, фильтры) обеспечивается специализированным модулем, текстовыми данными для которого являются: а) размеченные на фрагменты транскрипции, б) словари фрагментов, единицы которых выступают инвариантами соответствующих текстовых единиц, в) параметры и значения словарных единиц [Баранов и др., 2008; Баранов и др., 2010; Аникина и др., 2012a; Аникина и др., 2012b].

Интерфейс модуля позволяет выбрать списки текста, указать рукопись, по которой будут выравниваться фрагменты, указать диапазон листов этой рукописи или фрагменты для демонстрации, ввести маску искомой лингвистической единицы, определить порядок следования списков, а также форму визуализации фрагментов — *полный текст*, *инципит* <…> *эксплицит* или только *заголовки*. Порядок следования фрагментов зависит от выбранной формы: при визуализации текстов соответствующие друг другу части располагаются одна рядом с другой, при визуализации заголовков следование фрагментов соответствует их следованию в каждом из списков, а соответствующие друг другу части соединяются линиями, что позволяет увидеть различия в расположении фрагментов в рукописях.

Несмотря на неоднократную апробацию модели базы данных, процедур поиска и демонстрации соответствующих друг другу фрагментов, подготовка нового корпуса требует доработки и настройки менеджера в соответствии с особенностями текстов.

2. Параллельный корпус славянского паримейника

2.1. Рукописи

План работ по проекту «Средневековые тексты в современном контексте (новые методы и принципы представления средневековых текстов сегодняшним пользователям)» (рук. — проф. О. Ф. Жолобов) предусматривает, в частности, создание параллельного корпуса славянских списков паримейника, богослужебной книги, содержащей

[2] http://manuscripts.ru/mns/cred.cred?koll=61092969&f_type=12014

[3] http://manuscripts.ru/mns/cred.cred

[4] http://manuscripts.ru/mns/cred.cred?koll=62133570&f_type=14001

тексты из Ветхого и Нового Заветов, которые читаются на вечерне в дни праздников, в дни Великого поста и Страстной недели и др.

В качестве источников для корпуса выбраны четыре древнейших славянских кодекса XII, XIII и XIV вв. — Лазаревский (РГАДА, ф. 381, № 50), Захариинский (РНБ, Q.п.I.13), Федоровский (РГАДА, ф. 381, № 60) списки, а также рукопись РГБ, Тр. 4.

Рукописи содержат паримии Рождественско-Богоявленского (Рождество, Богоявление, паримии Водосвятия), Триодного (Великий пост, Сырная неделя, Лазарева суббота, Страстная неделя, Пентекостарий), месяцесловного циклов [Алексеев 2008: 234–245]. Списки имеют утраты, набор паримий и входящих в них чтений в ряде случаев не совпадают.

2.2. Модель. Разметка. Словарь

Для создания параллельного корпуса необходимы разметка транскрипций на паримии и соотнесение соответствующих друг другу текстов различных списков друг с другом.

Модель базы данных позволяет создать словарь фрагментов необходимого типа, а в каждом тексте — иерархию *текст — фрагмент — символ* и установить связь между словарными и текстовыми единицами. Каждый фрагмент может иметь характеристики (параметры) и значения характеристик, позволяющие организовать поиск и визуализацию единиц с помощью веб-форм корпуса.

Использование словарных и текстовых единиц для разметки транскрипций паримейников на паримии предполагает создание такого словаря паримий, в котором каждая единица уникальна и имеет идентифицирующие ее характеристики.

2.3. Проблема: традиция vs модель

Традиционно в печатных изданиях греческих профитологиев (Prophetologion) паримии имеют сквозную нумерацию, при которой каждая паримия получает номер, соотносящий ее с циклом или его частью и местом в цикле (или в части цикла). Так, например, восемь паримий, читающихся на Рождество, получили, в издании [Höeg et al. 1939, 1980–1981: 600] номер с L1a по L1h, а паримия, читающаяся на вечерне Страстного понедельника, — L36b, и т.д. При предложенной системе нумерации паримии, содержащие один и тот же текст, например, стихи с 1-го по 13-й первой главы Бытия (Быт. 1.1–13), получа-

ют различные номера (в указанном издании это номера L1a, L2a, L5b, L41b).

Использующийся в печатных изданиях принцип сквозной нумерации паримий противоречит модели базы данных корпуса, так как предполагает соотносить одну и ту же паримию, читаемую в различных циклах, с различными словарными единицами.

Кроме того, использование некоторого условного номера, пусть и соотносимого с циклом или его частью, не является интуитивно понятной «меткой» паримии, которая могла бы использоваться как при разметке корпуса (ср., например, «Быт. 1.1» Библии), так и при поиске фрагментов и их визуализации.

Индивидуальным значением паримии могла бы стать «метка», содержащая сведения о чтениях Библии, составляющих ее текст. Но в связи с тем, что в большом количестве случаев набор стихов, глав (а иногда и книг) паримии достаточно велик (например, вторая паримия на Богоявление содержит тексты Исх. 14:15–18, 21–23, 27–29, паримия на Сретение, 2 февраля, — тексты Исх. 12:51–13:2, 10–12, 14–16, 22:28; Лев. 12:2–4, 6, 8; Числ. 8:16–17 и др.), то использование таких значений в качестве основной индивидуализирующей характеристики не представляется возможным.

2.4. Решение: традиция & модель

В настоящее время в качестве рабочего варианта выбрана комбинированная характеристика паримий словаря, которая включает: а) тип фрагмента, б) индекс, указывающий на цикл (его часть) и литерный порядковый номер в цикле (его части), в) комментарий, содержащий сведения о книге(ах), главе(ах), стихах текстов паримии и г) тему паримии[5]:

Тип функционально-структурного раздела: паримия
Номер: P1b[6]
Комментарий: Чис. 24:2–3, 5–9, 17–18 «Пророчество Валаама».

При такой идентификации отдельной единицей словаря становится паримия, текст которой читается на определенной службе и содержит текст, соотносимый с соответствующими стихами Библии.

[5] См. [Алексеев 2008: 234–245].

[6] Где P — паримия, 1 — паримия на Рождество, b — 2-я паримия службы на Рождество.

Списки также содержат паримии, которые читаются в разное время и состав стихов в которых идентичен: Быт. 1:1–13 (3х — на Рождество, на Богоявление, на вечерне Понедельника 1-й недели Великого поста), Иез. 43:27–44.4 (4х — на 8 сентября, 21 ноября, 25 марта, 15 августа), Ис. 61:10–62.5 (2х — на вечерне Страстной субботы и на 11 мая) и др. (всего 28 паримий). В полутора десятках случаев пересечение частичное: Быт. 1:1–13 (паримия на Рождество и др.) и Быт. 1:1–5 (паримия на вечерне Страстной субботы), Ис. 9:5–6 (паримия на Рождество) и Ис. 8:13–9.6 (паримия на 6-м часе Понедельника 3-й недели Великого поста) и др.[7]. Указание на повторяющуюся паримию осуществляется с помощью значения, включающего несколько номеров, например: паримия Быт. 1:1–13 — на Рождество P1a = P2a = P5b, на Богоявление P2a = P1a = P5b, на вечерне Понедельника 1-й недели Великого поста P5b = P2a = P1a.

Наличие инвариантного (принятого в качестве нормирующего) набора чтений для каждой словарной паримии позволяет описать текстологические особенности паримий в текстах. Отличия могут касаться количества входящих в паримию стихов и порядка их следования. Так, одна из паримий на Вознесение, содержащая, согласно [Алексеев 2008: 240], чтения Зах. 14:1–11, в паримейнике Тр. 4 включает чтения Зах. 14:1–4, 8–11 (110г–111б), а в Захариинском — Зах. 14:1–3, 5, 4, 8–11 (л. 219в). Расхождения между ожидаемым и текстовым составом паримий связаны также с утратами части листов рукописей.

2.5. Автоматизация сопоставления. Разметка на стихи

Автоматизированный анализ текстологических особенностей списков предполагает предоставление пользователю возможности поиска и сопоставительной визуализации не только самих паримий, но и их состава, что возможно при наличии разметки как границ паримий, так и границ фрагментов, соответствующих стихам Библии, внутри них.

Помимо разметки списков на паримии, в рамках работ по проекту осуществляется разметка паримий на фрагменты, соответствующие стихам Ветхого и Нового Заветов. Работа выполняется с помощью словаря стихов Библии, связь контекстов рукописей с которыми обеспечивает поиск и демонстрацию параллельных стихов.

---

[7] Сопоставление паримий по степени совпадения входящих в них чтений имеет особое значение при анализе содержательной стороны паримий, так как «сходные по содержанию праздники имеют тождественный набор паримий» [Алексеев 2008: 167].

Особенностью процедуры разметки текстов на стихи является частое лексическое, грамматическое и синтаксическое своеобразие текстов славянских паримий по сравнению с соответствующими стихами славянской Библии. Следствием этого является сложность соотнесения в ряде случаев границ стихов Библии и границ соответствующих фрагментов паримий.

2.6. Визуализация корпуса

Наличие в корпусе процедур поиска и демонстрации соответствующих друг другу фрагментов разных списков, содержащих одни и те же тексты, позволяет уже на этапе разметки получить первый вариант параллельного корпуса четырех списков славянского паримейника[8]. Основным отличием визуализации транскрипций будет выравнивание списков на основе одновременно обеих разметок — разметки на паримии и вложенной разметки на стихи. Другим необходимым компонентом вывода параллельных контекстов должно стать упорядочение паримий не только по их следованию в одном из списков, но и по их следованию внутри богослужебных циклов. Третьим необходимым условием визуализации является демонстрация отсутствия паримий и/или стихов в том или ином списке.

3. Заключение

Любая прикладная разработка, связанная с представлением средневековых кодексов с помощью компьютерных технологий, безусловно, должна продолжать традиции печатных публикаций и одновременно обладать такими инструментами поиска и визуализации данных, которые обеспечивают многовходовость, настраиваемость, группировку, фильтрацию и другие возможности работы с материалом. При решении задач разметки данных может возникнуть ситуация противоречия между составом и структурой текста, традицией их представления и моделью, положенной в основу компьютерной системы. Предложенная система идентификации паримий позволяет, с одной стороны, сохранить связь с принятой в православной литургике их нумерацией, а с другой — однозначно идентифицировать их при работе с параллельным корпусом.

[8] http://manuscripts.ru/mns/cred.cred?koll=94052725&f_type=12014

Источники

Паримейник (Лазаревский или Сковородкинский), нач. XII в., РГАДА, ф. 381, оп. 1, ед. хр. 50, 126 л.

Паримейник (Захариинский), 1271 г., РНБ, Q. п.I. 13, 264 л.

Паримейник (Федоровский II), втор. пол. XIII в., РГАДА, ф. 381, № 60, 109 л.

Паримейник, 2-я пол. XIV в., РГБ, Тр. 4, 142 л.

Литература

1. *Алексеев А. А.* (2008), Библия в богослужении. Византийско-славянский лекционарий. СПб.
2. *Аникина Р. А., Баранов В. А.* (2012a), Параллельный корпус русских летописей XIII–XV вв. в интернете: инструментарий и методика лингвотекстологического анализа средневекового текста. Интеллектуальные системы в производстве. № 2(20), с. 157–162.
3. *Аникина Р. А., Баранов В. А.* (2012b), Параллельный корпус русских летописей в интернете: цели, задачи, технологическая основа, использование. Информационные технологии и письменное наследие: Материалы IV международной научной конференции, с. 12–18.
4. *Баранов В. А., Гнутиков Р. М.* (2008), Электронное критическое издание средневекового текста: постановка задачи, основные требования и инструментальная подготовка. Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: Материалы международной научной конференции, с. 36–44.
5. *Баранов В. А., Дубовцев С. В.* (2010), Электронное критическое издание средневекового славянского текста: модель данных и визуализация лингвистических единиц. Интеллектуальные системы в производстве. № 1, с. 280–287.
6. *Höeg C., Lake S., Zuntz G., Engberg G.* (1939, 1980–1981), Monumenta Musicae Byzantinae. Lectionaria. Vol. I. Prophetologium. Pars. I. Lectiones Nativitatis et Epiphaniae; Vol. I. Prophetologium. Pars. II. Lectiones anni Immobilis. Copenhagen.

References

1. *Alekseev A. A.* (2008), Biblija v bogosluzhenii. Vizantijsko-slavjanskij lekcionarij [Bible in Worship. Byzantine-Slavic Lectionary]. Saint Petersburg.
2. *Anikina R. A., Baranov V. A.* (2012a), Parallel'nyj korpus russkih letopisej XIII–XV vv. v Internete: instrumentarij i metodika lingvotekstologicheskogo analiza srednevekovogo teksta [Parallel Corpus of Russian Chronicles of the 13th–15th centuries on the Internet: Tools and Methods of Linguistic and Textological Analysis of a Medieval Text]. In: Intellektual'nye sistemy v proizvodstve [Intelligent Systems in Manufacturing]. No. 2(20), pp. 157–162.
3. *Anikina R. A., Baranov V. A.* (2012b), Parallel'nyj korpus russkih letopisej v Internete: celi, zadachi, tehnologicheskaja osnova, ispol'zovanie [Parallel Corpus of Russian Chronicles on the Internet: Goals, Objectives, Technological Basis, Use]. In: Infor-

macionnye tehnologii i pis'mennoe nasledie: Materialy IV mezhdunarodnoj nauchnoj konferencii [Information Technology and Textual Heritage: Proceedings of the 4th International Scientific Conference], pp. 12–18.
4. *Baranov V.A., Gnutikov R.M.* (2008), Jelektronnoe kriticheskoe izdanie srednevekovogo teksta: postanovka zadachi, osnovnye trebovanija i instrumental'naja podgotovka [Electronic Critical Edition of a Medieval Text: Problem Statement, Basic Requirements and Instrumental Preparation]. In: Sovremennye informacionnye tehnologii i pis'mennoe nasledie: ot drevnih tekstov k jelektronnym bibliotekam: Materialy mezhdunarodnoj nauchnoj konferencii [Modern Information Technologies and Textual Heritage: from Ancient Texts to Electronic Libraries: Proceedings of the International Scientific Conference], pp. 36–44.
5. *Baranov V.A., Dubovcev S.V.* (2010), Jelektronnoe kriticheskoe izdanie srednevekovogo slavjanskogo teksta: model' dannyh i vizualizacija lingvisticheskih edinic [Electronic Critical Edition of Medieval Slavonic Text: Data Model and Visualization of Linguistic Units]. In: Intellektual'nye sistemy v proizvodstve [Intelligent Systems in Manufacturing]. No. 1, pp. 280–287.
6. *Höeg C., Lake S., Zuntz G., Engberg G.* (1939, 1980–1981), Monumenta Musicae Byzantinae. Lectionaria. Vol. I. Prophetologium. Pars. I. Lectiones Nativitatis et Epiphaniae; Vol. I. Prophetologium. Pars. II. Lectiones anni Immobilis. Copenhagen.

Баранов Виктор Аркадьевич
Ижевский государственный технический университет
им. М. Т. Калашникова (Россия)
Baranov Victor
Kalashnikov Izhevsk State Technical University (Russia)
E-mail: victor.a.baranov@gmail.com

О. В. Блинова, Н. А. Тарасов

O. V. Blinova, N. A. Tarasov

СЛОЖНОСТЬ РУССКИХ ПРАВОВЫХ ТЕКСТОВ: МЕТОДЫ ОЦЕНКИ И ЯЗЫКОВЫЕ ДАННЫЕ[1]

COMPLEXITY OF RUSSIAN LEGAL TEXTS: ASSESSMENT METHODS AND LANGUAGE DATA

Аннотация. Для создания модели автоматического определения сложности русских правовых текстов было необходимо собрать коллекцию таких текстов, разметить их, выделить параметры оценки сложности в применении к выбранному формату разметки. Эти шаги описываются в настоящей работе. Обозначается состав корпусов современных русских юридических текстов CorRIDA, CorDec, CorCodex общим объемом 8,5 млн токенов. Описываются основания выбора инструментов лингвистической разметки (UDPipe, pymorphy2). Кратко характеризуются языковые признаки оценки сложности, среди которых: простейшие базовые метрики; пять формул читабельности; параметры оценки лексической сложности (значения TTR, Yule's K, количество гапаксов, аббревиатур, абстрактных слов и мн. др.); параметры оценки морфосинтаксической и дискурсивной сложности (значения Noun-Verb Ratio; количество граммем генитива, среднего рода, пассива; относительных предложений, аппозитивных модификаторов, лексических средств дискурсивной связности и пр.).
Ключевые слова. Языковая сложность, правовые документы, читабельность, лексическая сложность, морфосинтаксическая сложность, дискурсивная сложность, русские синхронные юридические корпусы.

Abstract. Our goal is to create a model for the automatic assessment of Russian legal texts complexity. To achieve this goal, it is necessary to create a text collection; perform linguistic markup; highlight the parameters for measuring the complexity, oriented on the selected markup format. These steps are described in this paper. We briefly describe three corpora of modern Russian legal texts "CorRIDA", "CorDes", "CorCodex" with a total size of 8.5 million tokens. We justify the choice of linguistic markup tools (UDPipe, pymorphy2). Then we characterize the linguistic features of the complexity assessment, including: the simplest basic metrics; five readability formulas; parameters for assessing lexical complexity (TTR values, Yule's K, the number of hapaxes, abbreviations, abstract words, etc.); parameters for assessing morphosyntactic and discursive complexity (Noun-Verb Ratio values; the number of grammemes of genitive, neuter, passive; relative sentences, appositive modifiers, lexical devices of discursive connectivity, etc.).
Keywords. Linguistic complexity, legal documents, lexical complexity, morphosyntactic complexity, discursive complexity, synchronous corpora of legal Russian.

[1] При поддержке гранта РНФ № 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика».

1. Введение

Языковой сложности посвящена обширная научная литература, как теоретическая, напр., [Dahl 2004; Бердичевский 2014], так и охватывающая конкретные методы оценки текстов (см., напр., обзоры в [Collins-Thompson 2014; Солнышкина, Кисельников 2015]). Современные методы подразумевают далеко не только использование классических формул читабельности. Как указано в [Crossley et al. 2019], «классические формулы читабельности в меньшей степени предсказывают понимание текста, чем формулы, разработанные с использованием лингвистических параметров, основанные на словесных признаках <…> и на признаках, которые оценивают вхождения лексических и синтаксических конструкций, связность текста, тональность, тематику и семантику». Таким образом, при оценке сложности используется все большее количество сложных для извлечения текстовых признаков, не только морфосинтаксических и лексических, но и дискурсивных (отражающих содержательную и структурную связность).

Юридический язык издавна критикуется за многословие, избыточность, длинноты, синтаксическую переусложненность, архаичную лексику и пр. (см., напр., [Tiersma 1999]).

Неудивительно, что и русские правовые тексты привлекли внимание исследователей, которые, во-первых, сконцентрировались в основном на оценке сложности текстов законов, во-вторых, использовали для оценки сложности либо только формулы читабельности (в [Дмитриева 2017] используется известная формула И. В. Оборневой), либо другие достаточно простые и немногочисленные метрики. Так, в [Кучаков, Савельев 2018] используется одна лексическая метрика (TTR, значение которой зависит от длины текста) и одна синтаксическая (расстояние между главным и зависимым по синтаксическому дереву зависимости, вычисляемое так: «для каждого конкретного текста взято одно значение, которое является максимальным для всех предложений текста» [Там же]).

В новейшей работе [Кнутов и др. 2020] использовано большее количество метрик (девять), среди которых: «доля глаголов в страдательном залоге», «доля глаголов от общего количества слов в тексте», «среднее количество слов в субстантивных именных словосочетаниях», «среднее количество причастных оборотов, расположенных в предложениях после определяемого слова, на одно предложение»,

«среднее количество деепричастных оборотов на одно предложение», «среднее количество слов в предложениях», «среднее расстояние между зависимыми словами в предложении», «среднее количество грамматических основ (предикативных основ, предикативных ядер) предложения (подлежащее, сказуемое или одно из них) в одном предложении», «среднее количество слов в абзаце».

2. Состав и объем юридических корпусов

Нашей целью является создание модели автоматического определения сложности русских правовых текстов, учитывающей значительное количество разнообразных параметров. Для достижения этой цели на начальном этапе было необходимо собрать коллекцию таких текстов, разметить их, выделить параметры оценки сложности в применении к выбранному формату разметки. Эти шаги описываются в разделах настоящей работы.

Мы собрали, предобработали и разметили три коллекции современных русских юридических текстов общим объемом около 8,5 млн токенов.

Во-первых, это коллекция русских локальных документов и актов CorRIDA, содержащая документы, с которыми периодически сталкиваются носители языка — неюристы (формы информированных согласий, договоров и пр., скачанные с сайтов государственных учреждений). Корпус CorRIDA состоит из 1546 документов и содержит 1 млн 784 тыс. токенов.

Во-вторых, это коллекция решений Конституционного Суда РФ CorDec, включающая 584 документа, 3427 тыс. токенов. Решения пишутся высокопрофессиональными юристами и адресованы широкому кругу граждан, описание см. в [Blinova et al. 2020a].

В-третьих, это коллекция нормативных документов CorCodex, содержащая 279 текстов кодексов, федеральных законов и постановлений Правительства РФ (в общей сложности 3 млн 227 тыс. токенов). Такие тексты вынуждены читать прежде всего профессиональные юристы.

Размеченные файлы корпусов в формате *.json будут полностью опубликованы на сайте plaindocument.org в конце 2021 г.

3. Разметка корпусов

Известно, что синтаксические признаки хорошо предсказывают языковую сложность, см., например, [Ivanov et al. 2018]. Корпусы, размеченные в формате UD (Universal Dependencies), в последнее время все более активно используются при оценке морфосинтаксической сложности как при межъязыковом сопоставлении, так и при сравнении текстов (коллекций текстов) на одном языке, см., например, [Berdicevsks et al. 2018; Çöltekin, Rama 2018; Yan, Kahane 2018; Dyer 2018].

Поэтому в качестве базового инструмента разметки наших трех коллекций выбран UDPipe. Как инструмент подробного морфологического анализа взят pymorphy2 [Korobov 2015].

Отдельной задачей стал выбор между доступными моделями UDPipe (существуют модели ru-syntagrus, ru-gsd, ru-taiga). Основанием для принятия решения стала статистика метрик, показывающая аккуратность работы моделей, представленная М. Стракой [Universal Dependencies 2.5 Models for UDPipe; Straka 2017]. Согласно этой статистике, где даны значения метрик аккуратности по параметрам UAS, LAS, MLAS и BLEX [CoNLL 2018 Shared Task], модель russian-syntagrus-ud-2.5 работает лучше, поэтому нами выбрана именно она.

После предобработки выполнена автоматическая лемматизация, морфологическая и синтаксическая разметка корпусов. Каждой словоформе присвоена двойная частеречная помета — в терминах UDPipe и в терминах pymorphy2. Частеречная разметка pymorphy2 позволяет различать ADJF — полные прилагательные, ADJS — краткие прилагательные, VERB — глаголы в личной форме, INFN — инфинитивы, PRTF — полные причастия, PRTS — краткие причастия и GRND — деепричастия. Это удобно для оценки морфосинтаксической сложности, в частности, потому, что наблюдается положительная корреляция между количеством полных прилагательных (а также причастий и деепричастий) и сложностью и отрицательная корреляция между количеством глаголов в личной форме и сложностью текстов [Дружкин 2016].

4. Выбранные параметры оценки языковой сложности

С целью оценки языковой сложности юридических текстов в составе собранных корпусов отобрано более 50 параметров. Значения каждого из параметров будут записаны в состав метаданных к текстам корпусов.

Среди параметров — значения простых (базовых) метрик текстов и индексы сложности, вычисленные при помощи формул читабельности. Кроме того, более содержательным образом будут оцениваться в значительной степени условно различаемые лексическая, морфосинтаксическая и дискурсивная сложность.

В качестве базовых метрик решено использовать (среди прочих): ASL — среднюю длину предложения в словах, ASW — среднюю длину слова в слогах. Из формул читабельности, адаптированных для русского языка, выбраны FRE (GL), SMOG, ARI, DCI, CLI, см. [Бегтин 2016; Solovyev et al. 2018].

В качестве параметров оценки лексической сложности выбраны: значения простого TTR; значения метрик лексического разнообразия Yule's K и Yule's I, не зависящие от длины текстов, см. [Blinova et al. 2020a]; количество гапаксов; количество лемм со значениями индексов частотности Zipf Value, в том числе лемм с низкой общеязыковой частотностью [Blinova et al. 2020b]; количество аббревиатур и сокращений; количество слов с абстрактным значением; количество юридических терминов. При оценке сложности будут использованы пользовательские словари, в том числе словарь абстрактных слов [Solovyev et al. 2020].

Для оценки морфосинтаксической сложности выбраны, кроме прочего: количество неслужебных слов, в частности, существительных; значения Noun-Verb Ratio; количество подчинительных и сочинительных союзов; количество тегов граммем родительного падежа, среднего рода, пассива; количество относительных предложений, аппозитивных модификаторов и предложений, содержащих отношения parataxis.

Среди учитываемых дискурсивных признаков, например, количество лексических средств дискурсивной связности, в частности, дискурсивных маркеров (в UD тег discourse); повтор именных групп в соседних предложениях (эта информация полезна для оценки референциальной связности); повтор значений граммем времени и вида у глаголов в личной форме.

5. Перспективы

Корпусы собраны и размечены, параметры оценки сложности текстов отобраны. Ближайшим шагом станет автоматическое вычисление значений этих параметров. Затем с помощью методов машинного

обучения мы создадим модель автоматического определения сложности русских текстов.

После определения численных значений (итоговых индексов) сложности будут использованы средства машинного обучения. Применение моделей трансферного обучения позволит не только с большой точностью определять сложность текстов, но и установить, какие параметры текстов в наибольшей степени коррелируют с целевой сложностью. В качестве примера нейросети с подобной архитектурой можно привести модель Universal Sentence Encoder [Cer et al. 2018]. Эта модель кодирует текст в многомерные векторы, которые можно использовать для классификации, определения семантического сходства, кластеризации и других задач обработки естественного языка. Задача определения сложности текстов в данном контексте является задачей регрессии, и метод USE в ней можно использовать для первоначального кодирования текстов и дальнейшего обучения модели методами глубокого обучения или градиентного бустинга.

Литература

1. *Бегтин И. В.* (2016), Оценка читабельности текста. URL: https://github.com/ivbeg/readability.io/wiki/API (дата обращения: 01.07.2021).
2. *Бердичевский А.* (2012), Языковая сложность (Language Complexity). Вопросы языкознания. № 5, с. 101–124.
3. *Дмитриева А. В.* (2017), «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации. Сравнительное конституционное обозрение. Т. 118, № 3, с. 125–133.
4. *Дружкин К. Ю.* (2016), Метрики удобочитаемости для русского языка: выпускная квалификационная работа магистра. М.: НИУ ВШЭ.
5. *Кнутов А. В., Плаксин С. М. и др.* (2020), Сложность российских законов. Опыт синтаксического анализа. М.: ИД Высшей школы экономики.
6. *Кучаков Р., Савельев Д.* (2018), Сложность правовых актов в России: лексическое и синтаксическое качество текстов. СПб.: Институт проблем правоприменения Европейского университета в Санкт-Петербурге.
7. *Солнышкина М. И., Кисельников А. С.* (2015), Сложность текста: этапы изучения в отечественном прикладном языкознании. Вестник Томского государственного университета. Филология. № 6(38), с. 86–99.

References

1. *Begtin I. V.* (2016), Ocenka chitabel'nosti teksta [Text Readability Assessment], URL: https://github.com/ivbeg/readability.io/wiki/API (date of access: 01.07.2021).
2. *Berdicevskij A.* (2012), Jazykovaja slozhnost' [Language Complexity]. In: Voprosy Jazykoznanija [Topics in the study of language]. No. 5, pp. 101–124.

3. *Berdicevskis A., Çöltekin Ç., Ehret K., Prince K., Ross D., Thompson B., Yan C., Demberg V., Lupya G., Rama T., Bentz C.* (2018), Using Universal Dependencies in cross-linguistic complexity research. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018).
4. *Blinova O. V., Belov S. A., Revazov M. A.* (2020a), Decisions of Russian Constitutional Court: Lexical Complexity Analysis in Shallow Diachrony. R. V. Bolgov, A. V. Chugunov, A. E. Voiskounsky (eds.). In: CEUR Workshop Proceedings. Vol. 2813. Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, Russia, 17–20 June 2020, pp. 61–74.
5. *Blinova O. V., Tarasov N. A., Modina V. V., Blekanov I. S.* (2020b), Modeling Lemma Frequency Bands for Lexical Complexity Assessment of Russian Texts. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii [Computational Linguistics and Intellectual Technologies]. Vol. 19(26), pp. 76–92.
6. *Cer D., Yang Y., Kong S., Hua N., Limtiaco N. et al.* (2018), Universal Sentence Encoder. ArXiv: 1803.11175.
7. *Collins-Thompson K.* (2014), Computational assessment of text readability: a survey of current and future research. In: Th. François, D. Bernhard (eds.). Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics. Vol. 165(2), pp. 97–135.
8. *Çöltekin Ç., Rama T.* (2018), Exploiting universal dependencies treebanks for measuring morphosyntactic complexity. In: Proceedings of the First Shared Task on Measuring Language Complexity, pp. 1–8.
9. CoNLL 2018 Shared Task. URL: http://universaldependencies.org/conll18/evaluation.html (date of access: 01.07.2021).
10. *Crossley S. A., Skalicky S., Dascalu M.* (2019), Moving beyond classic readability formulas: new methods and new models. In: Journal of Research in Reading. No. 42(3-4), pp. 541–561.
11. *Dahl Ö.* (2004), The growth and maintenance of linguistic complexity. Amsterdam.
12. *Dmitrieva A. V.* (2017), «Iskusstvo juridicheskogo pis'ma»: kolichestvennyj analiz reshenij Konstitucionnogo Suda Rossijskoj Federacii. Sravnitel'noe konstitucionnoe obozrenie ["The art of legal writing": A quantitative analysis of Russian Constitutional Court rulings]. In: Sravnitel'noe konstitucionnoe obozrenie [Comparative Constitutional Review]. No. 118(3), pp. 125–133.
13. *Druzhkin K. Ju.* (2016), Metriki udobochitaemosti dlja russkogo jazyka: vypusknaja kvalifikacionnaja rabota magistra [Readability metrics for Russian: master's theses]. Moscow: Higher School of Economics.
14. *Dyer W.* (2018), Integration complexity and the order of cosisters. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pp. 55–65.
15. *Ivanov V. V., Solnyshkina M. I., Solovyev V. D.* (2018), Efficiency of text readability features in Russian academic texts. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii 2018 [Computational Linguistics and Intellectual Technologies]. Vol. 17, pp. 277–287.
16. *Knutov A. V., Plaksin S. M. et al.* (2020), Slozhnost' rossijskih zakonov. Opyt sintaksicheskogo analiza [The complexity of Russian laws. Syntactic analysis experience]. Moscow: Publishing house of the Higher School of Economics.

17. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Analysis of Images, Social Networks and Texts, pp. 320–332.
18. *Kuchakov R., Saveliev D.* (2018), Slozhnost' pravovyh aktov v Rossii: leksicheskoe i sintaksicheskoe kachestvo tekstov [The complexity of legal acts in Russia: Lexical and syntactic quality of texts: analytic note]. European University at Saint Petersburg, St. Petersburg.
19. *Solovyev V., Solnyshkina M., Andreeva M., Danilov A., Zamaletdinov R.* (2020), Text Complexity and Abstractness: Tools for the Russian Language. In: R. V. Bolgov, A. V. Chugunov, A. E. Voiskounsky (eds.). CEUR Workshop Proceedings. Vol. 2813. Proceedings of the International Conference "Internet and Modern Society" (IMS-2020), St. Petersburg, Russia 17–20 June 2020, pp. 75–87.
20. *Solnyshkina M. I., Kisel'nikov A. S.* (2015), Slozhnost' teksta: jetapy izuchenija v otechestvennom prikladnom jazykoznanii [Text complexity: study phases in Russian linguistics]. In: Vestnik Tomskogo Gosudarstvennogo Universiteta, Filologiya [Tomsk State University Journal of Philology]. No. 6(38), pp. 86–99.
21. *Tiersma P.* (1999), Legal Language. Chicago: University of Chicago Press.
22. Universal Dependencies 2.5 Models for UDPipe. URL: https://github.com/jwijf-fels/udpipe.models.ud.2.5/blob/master/inst/udpipe-ud-2.5-191206 (date of access: 01.07.2021).
23. Universal Dependencies. URL: https://universaldependencies.org.

Блинова Ольга Владимировна
Санкт-Петербургский государственный университет (Россия)
Национальный исследовательский университет
«Высшая школа экономики» (НИУ ВШЭ СПб) (Россия)
Blinova Olga
Saint Petersburg State University (Russia)
National Research University Higher School of Economics (Russia)
E-mail: o.blinova@spbu.ru

Тарасов Никита Андреевич
Санкт-Петербургский государственный университет (Россия)
Tarasov Nikita
Saint Petersburg State University (Russia)
E-mail: tarasovn2468@yandex.ru

Н. В. Богданова-Бегларян

N. V. Bogdanova-Beglarian

МАРКЕРЫ-КСЕНОПОКАЗАТЕЛИ В РУССКОЙ ПОВСЕДНЕВНОЙ РЕЧИ: АННОТИРОВАНИЕ РЕЧЕВОГО КОРПУСА, ТИПОЛОГИЯ И КОЛИЧЕСТВЕННЫЕ ДАННЫЕ[1]

'XENO'-MARKERS IN RUSSIAN EVERYDAY SPEECH: ANNOTATION OF THE SPEECH CORPUS, TYPOLOGY AND QUANTITATIVE DATA

Аннотация. В статье излагаются результаты анализа аннотированных выборок из двух речевых корпусов. Разметка проводилась для выявления в материале прагматических маркеров; описанию, систематизации и количественной характеристике подверглись маркеры одного из десяти выделенных классов – ксенопоказатели. В статье приведена их типология и данные об их реализации в речи разных говорящих, в разных коммуникативных ситуациях.
Ключевые слова. Повседневная речь, звуковой корпус, аннотирование корпуса, прагматикализация, прагматический маркер, ксенопоказатель.

Abstract. The article proposes the results of the analysis of annotated samples from two speech corpora created at St. Petersburg State University: the corpus of everyday Russian speech "One Speaker's Day" (ORD) and the corpus of Russian monologic speech "Balanced Annotated Text Library" (SAT). The corpus annotation was carried out to identify pragmatic markers (PM) in the material. The markers of one of the ten identified PM classes were described – 'xeno'-markers (PMK): *grit/gr'u, takoi/aya/ie, tipa (togo chto), vrode (togo chto), znaesh', slushaj, tak*, etc. The grammatical specificity of these units allows us to consider them not as particles but as pragmatic markers. The article provides a typology of PMK and quantitative data of their use in different speakers' discourse and diverse communicative situations.
Keywords. Everyday speech, speech corpus, corpus annotation, pragmatikalization. pragmatic marker, 'xeno'-marker.

1. Введение

Ксенопоказатели рассматриваются в статье как класс прагматических маркеров (ПМ). Последние возникают в устной речи в результате последовательно действующих процессов *грамматикализации* и *прагматикализации*, в ходе которых единицы сначала меняют свой грам-

[1] Исследование проведено при финансовой поддержке гранта Санкт-Петербургского государственного университета (проект № 75254082 «Моделирование коммуникативного поведения жителей российского мегаполиса в социально-речевом и прагматическом аспектах с привлечением методов искусственного интеллекта»).

матический статус, а затем и вовсе практически утрачивают и лексическое, и грамматическое значение и приобретают прагматическое, т. е. определенную функцию в дискурсе. Для маркеров-ксенопоказателей (ПМК) такой функцией становится указание на «присутствие Другого» [Арутюнова 2000: 448]: это и чужая речь, и своя собственная, сказанная ранее или только планируемая на будущее, собственные или чужие мысли и даже речевое, «говорящее», поведение другого человека, его реакции и т. д. «Классические» ксенопоказатели, возникшие из глаголов, — *мол, де, дескать* — остановились в своем развитии на этапе грамматикализации и сегодня описываются как частицы. «Новые» ксенопоказатели чаще проходят другой путь грамматикализации: через вводные единицы — к ПМК. Их перечень в современной речи очень обширен и имеет столь же обширный набор «протоединиц»: вводные единицы, глаголы, наречия, местоимения, союзы. В результате многие ПМК сохранили ограниченную грамматическую изменяемость (*грю/гришь/грит/грим; такой/ая/ие*), что и позволяет числить их в классе не частиц, но — прагматических маркеров (ср.: [Fraser 1996]).

2. Источники материала и методика исследования

Основными источниками материала для настоящего исследования стали два речевых корпуса, созданных в СПбГУ: корпус повседневной русской речи «Один речевой день» (ОРД) и корпус русской монологической речи «Сбалансированная аннотированная текстотека» (САТ) (см. о них: [Звуковой корпус… 2013; Русский язык… 2016]). В целях получения максимально полного инвентаря тех функциональных единиц, которые использует человек в ходе речепорождения, а также в целях описания их функционирования в речи разных говорящих и в разных коммуникативных ситуациях корпусный материал был проаннотирован с выявлением в нем прагматических маркеров. Такая разметка оказалась весьма затруднена тем обстоятельством, что ПМ внешне ничем не отличаются от своих «протоединиц» и лишь в контексте реализуют свой новый статус. В результате ПМ, как и многие дискурсивные слова, «можно изучать только через их употребление» [Дискурсивные слова… 1998: 8–10]. Кроме того, ПМ часто полифункциональны и склонны выстраиваться в цепочки однофункциональных единиц (о проблемах корпусной разметки ПМ и способах их решения см.: [Bogdanova-Beglarian et al. 2018, 2020; Богданова-Бегларян

и др. 2019а, 2019б]). Дополнительным источником материала для наблюдений и выводов стал устный подкорпус (УП) НКРЯ.

На этом материале был получен, в числе прочего, относительно полный набор ПМК и статистика их встречаемости в речи разных говорящих и в разных коммуникативных ситуациях.

3. Ксенопоказатели устной речи в зеркале статистики

Исследование произведено на аннотированных выборках ОРД (195 эпизодов «речевых дней» 104 информантов, более 300 тыс. словоупотреблений) и САТ (монологи разного типа от 34 информантов, более 50 тыс. словоупотреблений). Общий словник ПМ, полученных на этом материале, составил 60 единиц. Их типология включает 10 разновидностей: хезитативы (Х), рефлексивы (Ф), метакоммуникативы (М), разграничители (Г), ксенопоказатели (К), аппроксиматоры (А), дейктические (Д) и ритмообразующие (Р) маркеры, маркеры самокоррекции (С) и заместители (З) [Прагматические маркеры… 2021].

В общем частотном списке функций ПМ ксенопоказатели имеют ранг 5 и уступают по встречаемости в нашей речи только хезитативам, разграничителям, метакоммуникативам, а также смешанной группе ГХ. Такая ситуация свойственна и большинству социолектов, выделенных по признакам пола, возраста, образования и статуса говорящих, уровня их речевой компетенции (УРК) и рода деятельности. Обнаружились лишь некоторые исключения. Больше всего ксенопоказателей (ранг 2) оказалось в речи работников службы сервиса, а также учащихся и руководителей, гуманитариев и людей с незаконченным высшим образованием (ранг 3). Совсем их не встретилось в речи детей, что наводит на мысль, что способность употреблять такие маркеры приобретается только с речевым опытом.

Тот же ранг 5 имеют ксенопоказатели и в речи экстравертов, тогда как в топ-5 частотного списка функций в речи интровертов ПМК вообще не попали. Наиболее частотен в речи экстравертов оказался маркер ГОВОРИТ (в его редуцированной форме *грит*) — третье место после ВОТ и ТАМ, с частотой 6,05 %.

Учет типа и места коммуникации показал, что реже всего ПМК используются в публичной и учебной речи (ранги 8 и 11 соответственно), в казарме и в кафе/ресторанах (ранги 11 и 10). Чаще всего — в разговорах в поликлинике (ранг 2).

Из конкретных ксенопоказателей наиболее частотен маркер *говорит* — в разных его редуцированных формах: *грит* (1116 ipm), *грю* (669), *грят* (47), *гришь* (10). Частотность остальных ПМ, как инвариантов, так и структурных разновидностей, установить затруднительно в силу их высокой полифункциональности. Конкретную функцию маркера можно установить лишь в контексте.

4. Опыт систематизации маркеров-ксенопоказателей

Систематизировать ксенопоказатели в корпусном материале можно, например, с опорой на их прототип (условная чужая речь в контекстах подчеркнута, ПМК выделен).

1. Вводная единица → ПМК:
   - *Вы пришли в булочную и Вас там у кассы облаяли / потому что **видите ли** у Вас крупная бумага / а у них сдачи нет* (УП);
   - *ну или надо посмотреть где арен… в аренду парики сдают / ну там же не будет **так сказать** ой нам надо всего на десять минут дайте нам его / за сто рублей и они скажут идите в ж-ж*пу может быть* (ОРД);
   - *он такой / понял **короче** / по… посмеялся / поулыбался // *Н без темы // *П ну вот / я помогу **короче*** (ОРД)[2].

2. Союз → ПМК:
   - *какой будет / прогноз ? *П # да **вроде у** нас теплеет* (ОРД);
   - *Ну / она* [нрзб] *когда отвечаю / то в окно смотрит там / то лицо такое делает / **типа** я тупая* (УП);
   - *Если взять так / тот же Билл Гейтс / **якобы** образовавший компанию «Майкрософт» / который самый богатый человек…* (УП).

3. Глагол → ПМК:
   - *ну вот // *В я говорю **говорю** / я **говорю** / барышня ! **говорю** / какая вы молодца **говорю** // правильно / () а то / я некрасивая ! сделайте меня покрасивше **говорю*** (ОРД);
   - *и(:) Наталья_Георгиевна % мне (э) всё говорила / **знаешь** / их тогда никто не обойдёт на(:) этапе поставки* (ОРД);
   - *и он блин мне написал **слушай** сорри не могу* (ОРД).

[2] Об особенностях дискурсивной транскрипции материала ОРД см.: [Русский язык… 2016: 242–243].

4. Местоимения → ПМК:
   - *я вчера их встречаю / на улице / ну в «Пятёрочку»$ / я шла как раз в магазин / а был вечер / полдесятого // я* ***такая*** */* *о-о / вы приехали // а у вас завтра занятия будут? / они* ***такие*** */* *будут / я говорю / да-а / не повезло мне // а что такое ? // а у меня завтра у вас три … два семинара* (ОРД);
   - *Её / на сутки позже / принесли на кормление… Его оживляли тоже. Бук… буквально оживляли. В общем я там взволновалась / но они / очень боялись что у меня там молоко пропадёт /* ***то-сё*** */ «не волнуйтесь / не волнуйтесь»* (УП);
   - *Я ж и говорю, вам бы все читать, а если вам живого человека дают, так вы,* ***это самое****, зарежет меня живой человек* (ОП) (единственный пример из основного подкорпуса НКРЯ).

5. Наречие → ПМК:
   - *а-а *П э судья на Колю% **так** / тщ(:) ! *П вы не на базаре* (ОРД).

6. Комбинации разных частей речи → ПМК:
- *я подошёл к тренеру и говорю ему* ***так и так вот*** *я занимался водным поло / энное количество лет / я хочу попасть в команду* (САТ);
- *вот один из них спросил таким / *К *П так / это просто со странным немножко (…) отношением и взглядом / *П а зачем ? *П ну* ***вроде того что*** *их интересует только живая / (…) музыка // живой фольклор* (ОРД);
- *и тут звонок в дверь // стоит этот мужик // *П* ***типа того что*** *блин / давайте общаться !* (ОРД);
- *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* (ОРД);
- *Вот / пишет мне* ***типа*** *ну вот* ***так и так*** */* ***типа*** *перевод / я говорю / «Ну деньги сначала / да / там все дела»* (УП).

Возможны, вероятно, и иные способы систематизации ксенопоказателей: проаннотированный корпусный материал предоставляет такие возможности.

5. Некоторые выводы

Анализ материала показал, что круг маркеров-ксенопоказателей в повседневной речи очень широк и значительно превышает известные из литературы перечни (см., например: [Левонтина 2020]). Как правило, все ПМК полифункциональны (исключение — *якобы*: только К — см.: [Прагматические маркеры… 2021]) и склонны выстраиваться в цепочки. Анализ демонстрирует «отрыв» ПМК от глаголов говорения (ср.: *де, дескать, мол*) и явное преобладание ПМК в диалоге по сравнению с монологом: ОРД — 2336 ipm, ранг 5; САТ — 420 ipm, ранг 8.

Полный перечень ПМК может быть полезен в самых разных прикладных аспектах лингвистики: преподавание РКИ, практика перевода, лингвистическая экспертиза, автоматическая обработка речи, лингвокриминалистика.

Литература

1. *Арутюнова Н. Д.* (2000), Показатели чужой речи *де, дескать, мол.* Язык о языке: сб. статей. М.: Языки русской культуры, с. 437–452.
2. *Богданова-Бегларян Н. В., Блинова О. В., Шерстинова Т. Ю., Зайдес К. Д., Попова Т. И.* (2019а), Корпус естественной речи: проблемы ручного аннотирования прагматических маркеров и пути их решения. Анализ разговорной русской речи (АР^3-2019): Труды восьмого междисциплинарного семинара. СПб.: Политехника-принт, с. 5–10.
3. *Богданова-Бегларян Н. В., Блинова О. В., Мартыненко Г. Я., Шерстинова Т. Ю., Зайдес К. Д., Попова Т. И.* (2019б), Аннотирование прагматических маркеров в русском речевом корпусе: проблемы, поиски, решения, результаты. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 18(25), с. 72–85.
4. Дискурсивные слова русского языка: Опыт контекстно-семантического описания (1998). М.: Метатекст, 448 с.
5. Звуковой корпус как материал для анализа русской речи: Коллективная монография. Ч. 1. Чтение. Пересказ. Описание / Авторы: *Н. В. Богданова-Бегларян (Богданова), И. С. Бродт, В. В. Куканова, О. В. Павлова (Ильичева), Е. М. Сапунова, И. А. Суббота* и др. (2013). СПб: Филологический ф-т СПбГУ, 532 с.
6. *Левонтина И. Б.* (2020), Об арсенале ксенопоказателей в русском языке. Вопросы языкознания. № 3, с. 52–77.
7. Прагматические маркеры русской повседневной речи: Словарь-монография (2021). СПб.: Нестор-История, 520 с.
8. Русский язык повседневного общения: особенности функционирования в разных социальных группах. Коллективная монография / Авторы: *Н. В. Богданова-Бегларян, Т. Ю. Шерстинова, Е. М. Баева, О. В. Блинова, Г. Я. Мартыненко* и др. (2016). СПб: ЛАЙКА, 244 с.

9. *Bogdanova-Beglarian N., Blinova O., Martynenko G., Sherstinova T., Zaides K.* (2018), Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks. In: Proceedings of the FRUCT'23. Bologna, Italy. FRUCT Oy, Finland, pp. 69–77.
10. *Bogdanova-Beglarian N., Blinova O., Sherstinova T., Gorbunova D, Zaides K., Popova T.* (2020), Pragmatic Markers in Dialogue and Monologue: Difficulties of Identification and Typical Formation Models. In: SPECOM 2020. Lecture Notes in Artificial Intelligence, LNAI. Vol. 12335. Springer, Switzerland, pp. 68–78.
11. *Fraser B.* (1996), Pragmatic Markers. In: Pragmatics. Vol. 6(2), pp. 167–190.

References

1. *Arutyunova N. D.* (2000), Pokazateli chuzhoj rechi *de, deskat', mol* [Indicators of Someone Else's Speech *de, deskat', mol*]. In: Jazyk o jazyke: sb. statej [Language about Language: a Collection of Articles]. Moscow: Languages of Russian Culture, pp. 437–452.
2. *Bogdanova-Beglarian N., Blinova O., Sherstinova T., Zaides K., Popova T.* (2019a), Korpus jestestvennoj rechi: problemy ruchnogo annotirovania pragmaticheskikh markerov i puti ikh reshenia [Natural Speech Corpus: Problems of Manual Annotation of Pragmatic Markers and Ways to Solve Them]. In: Analiz razgovornoj russkoj rechi (AR^3-2019): Trudy vos'mogo mezhdisciplinarnogo seminara [Analysis of Colloquial Russian Speech (AR^3-2019): Proceedings of the Eighth Interdisciplinary Seminar]. Saint Petersburg: Polytechnic-print, pp. 5–10.
3. *Bogdanova-Beglarian N., Blinova O., Martynenko G., Sherstinova T., Zaides K., Popova T.* (2019b), Annotirovanie pragmati-cheskikh markerov v russkom rechevom korpuse: problemy, poiski, reshenia, rezul'taty [Annotation of Pragmatic Markers in the Russian Speech Corpus: Problems, Searches, Solutions, Results]. In: Kompjuternaya lingvistika i intellektual'nye tekhnologii: Po materialam jezhegodnoj mezhdunarodnoj konferencii "Dialog" [Computational Linguistics and Intelligent Technologies: Based on the Materials of the Annual International Conference "Dialogue"]. Iss. 18(25). Moscow, pp. 72–85.
4. *Bogdanova-Beglarian N., Blinova O., Martynenko G., Sherstinova T., Zaides K.* (2018), Pragmatic Markers in Russian Spoken Speech: an Experience of Systematization and Annotation for the Improvement of NLP Tasks. In: Proceedings of the FRUCT'23. Bologna, Italy. FRUCT Oy, Finland, pp. 69–77.
5. *Bogdanova-Beglarian N., Blinova O., Sherstinova T., Gorbuno-va D., Zaides K., Popova T.* (2020), Pragmatic Markers in Dialogue and Monologue: Difficulties of Identification and Typical Formation Models. In: SPECOM 2020. Lecture Notes in Artificial Intelligence, LNAI. Vol. 12335, Springer, Switzerland, pp. 68–78.
6. Diskursivnye slova russkogo jazyka: Opyt kontekstno-semanticheskogo opisania *(1998)* [Discursive Words of the Russian Language: Experience of Contextual-Semantic Description]. Moscow: Metatext, 280 p.
7. *Fraser B.* (1996), Pragmatic Markers. In: Pragmatics. Vol. 6(2), pp. 167–190.
8. *Levontina I. B.* (2020), Ob arsenale ksenopokazatelej v russkom jazyke [On the Arsenal of xeno-markers in the Russian Language]. In: Voprosy jazykoznania [Linguistic Issues]. No. 3, pp. 52–77.

9. Pragmaticheskie markery russkoj povsednevnoj rechi: slovar'-monografia (2021) [Pragmatic Markers of Russian Everyday Speech: Dictionary-Monograph]. Saint Petersburg: Nestor-History, 520 p.
10. Russkij jazyk povsednevnogo obshchenia: osobennosti funktsi-onirovania v raznykh sotsial'nykh gruppakh. Kollektivnaya monografija (2016). Authors: N. V. Bogdanova-Beglaryan, T. Yu. Sherstinova, E. M. Baeva, O. V. Blinov, G. Ya. Martynenko et al. [Russian Language of Everyday Communication: Features of Functioning in Different Social Groups. Collective monograph] Saint Petersburg: Laika, 244 p.
11. Zvukovoj korpus kak material dl'a analiza russkoj rechi. Kollektivnaya monografija. CH. 1. Chtenie. Pereskaz. Opisaniye (2013) Authors: N. V. Bogdanova-Beglaryan (Bogdanova), I. S. Brodt, V. V. Kukanova, O. V. Pavlova (Ilyicheva), E. M. Sapunova, I. A. Subbota et al. [Sound Corpus as a Material for the Analysis of Russian Speech: Collective Monograph. Part 1. Reading. Retelling. Description]. Saint Petersburg: Faculty of Philology, St. Petersburg State University, 532 p.

Богданова-Бегларян Наталья Викторовна
Санкт-Петербургский государственный университет (Россия)
Bogdanova-Beglarian Natalia
Saint Petersburg State University (Russia)
E-mail: n.bogdanova@spbu.ru

Е. А. Вольф, Ю. О. Короткова, К. И. Семенов

E. A. Volf, Y. O. Korotkova, K. I. Semenov

АВТОМАТИЧЕСКАЯ РАЗМЕТКА ЗАИМСТВОВАНИЙ ИЗ РУССКОГО ЯЗЫКА В КИТАЙСКИХ ТЕКСТАХ: ПРОБЛЕМЫ СЛОВОДЕЛЕНИЯ И МОРФОПАРСИНГА[1]

AUTOMATIC ANNOTATION OF THE RUSSIAN LOANWORDS IN CHINESE TEXTS: ISSUES IN WORD SEGMENTATION AND POS-TAGGING

Аннотация. Статья посвящена вопросам автоматической аннотации китайских текстов в Русско-китайском параллельном корпусе НКРЯ в двух аспектах: разделения текста на слова и морфологического парсинга. Особое внимание уделяется предложениям, содержащим фонетические заимствования из русского языка. Результаты исследований планируется применить при улучшении лингвистической разметки в Русско-китайском параллельном корпусе НКРЯ.
Ключевые слова. стандартный китайский язык (путунхуа), автоматическая сегментация, морфологическая аннотация, проблема слов вне словаря.

Abstract. The article addresses the issues in the automatic annotation of the Chinese texts in the Russian-Chinese Parallel Corpus of RNC in two aspects: word segmentation, and PoS-tagging. We paid particular attention to the processing of the Russian loanwords that are abundant in the Corpus data. We plan to take the results of the research into consideration while creating the new preprocessing pipeline of the Chinese texts in the corpus.
Keywords. Standard Mandarin (Putonghua), Chinese word segmentation (CWS), PoS-tagging, out-of-vocabulary problem (OOV).

1. Введение

Задачи токенизации и морфологической разметки китайских текстов связаны с рядом трудностей в области орфографии, фонетики и морфосинтаксиса. Все вышеперечисленные проблемы усугубляются в случае, если в тексте присутствуют заимствованные слова.

С набором этих проблем столкнулась команда разработчиков Русско-китайского параллельного корпуса НКРЯ (далее — ruzhcorp). Несмотря на то, что на текущий момент ruzhcorp обладает несколькими уровнями разметки для китайских текстов, в том числе разделением на слова (далее — CWS, от Chinese word segmentation), все они пред-

[1] Проект «Лингвоспецифическая разметка китайских текстов в Русско-китайском параллельном корпусе НКРЯ» выполнен при поддержке Комиссии по поддержке образовательных инициатив Факультета гуманитарных наук (ФГН) Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ) в рамках Конкурса проектных групп для обучающихся ФГН НИУ ВШЭ.

ставляют собой достаточно примитивные алгоритмы, выдающие много ошибок при обработке заимствований.

В настоящей статье мы представляем два исследования по улучшению текущего положения дел в разметке китайских текстов в ruzhcorp: экспериментальную работу в области словоделения (Раздел 2) и PoS-тэггинга (Раздел 3).

2. Исследования в области словоделения

2.1. Обзор стандартов и алгоритмов CWS и экспериментального набора данных

В китайском языке отсутствует кодифицированный набор иероглифов, регулярно используемых при транслитерации заимствований. Более того, в китайской орфографии практически не существует маркеров, указывающих на начало или конец последовательности символов, которые необходимо читать «фонетически», игнорируя семантику иероглифов. Таким образом, одна и та же комбинация полнозначных китайских иероглифов может быть распознана и как набор несущих смысл китайских морфем, и как набор знаков, передающих только звучание.

Учитывая, что в ruzhcorp присутствуют как тексты, содержащие заимствования из русского языка в больших количествах, так и тексты большого размера, состоящие исключительно из «стандартных» китайских слов (например, китайская литература XX в.), задача CWS для корпуса распадается на два блока: во-первых, качественное разделение «стандартного» (т. е. не содержащего заимствований) китайского текста на слова; во-вторых, максимально корректное выделение фонетических заимствований. Для решения обеих задач мы сравнили качество ведущих алгоритмов CWS в области выделения заимствований.

В настоящий момент существует несколько популярных стандартов CWS, при этом ни один из них не является общепринятым ни в науке, ни в NLP. Стандарты словоделения могут принципиально различаться, так как они берут за основу различные аспекты и уровни языковой структуры — семантику, морфосинтаксис или лексику. Более того, для большинства этих стандартов существует более одной их реализации. В нашем исследовании было использовано 17 алгоритмов и их вариантов. Краткое описание каждого стандарта словоделения и реализующих его алгоритмов представлено в табл. 1.

Таблица 1. **Сравнительное описание стандартов CWS**

Стандарт (по [Emerson 2005])	**Основные критерии**	**Алгоритмы, соответствующие стандарту**
Пекинский университет (PKU)	Лексическая семантика и сочетаемость	PKUSeg [Luo et al. 2019], fastHan // {pku, sxu} [Geng et al. 2020], LTP [Che et al. 2021]
Academia Sinica (CNS)	Морфосинтаксис	Ckiptagger [Li, Ma 2019], fastHan // {as, cnc}
Microsoft Research Asia	Морфосинтаксис	fastHan // msr
Penn Chinese Treebank (CTB); вариация — UD	Синтаксис	Stanza [Qi et al. 2018], fastHan // {ctb, udc, wtb, zx}, Spacy [Honnibal, Montani 2017], UDPipe [Straka 2018]
Словарные стандарты	Наличие слова в словаре	NLPIR [张[Zhang] & 商[Shang, 2019]

Цель нашего исследования — определить наилучший алгоритм CWS из вышеперечисленных, который будет оптимально справляться с выделением заимствований из русского языка. Для проверки качества алгоритмов мы создали датасеты, состоящие из 408 предложений из художественной литературы и 87 предложений из текстов СМИ, в которых содержались заимствования.

Исследование качества алгоритмов CWS на наших данных включает как количественный, так и качественный их анализ. В силу ограничений по формату мы вынуждены представить здесь лишь наши соображения относительно качественного анализа; подробное описание количественного анализа представлено в [Семенов и др. 2021]. Отметим лишь, что наиболее эффективными оказались нейросетевые алгоритмы fastHan (в особенности на стандарте CTB), LTP и PKUSeg, дающие качество 80–95 % на каждом датасете.

2.2. Сравнение алгоритмов CWS на данных корпуса: качественный анализ

Ошибки, сделанные сегментаторами на наших данных, можно разделить на две большие категории. Первая представляет разделение на слова, которое, несомненно, нарушает целостность заимствований и некорректно с точки зрения любого алгоритма словоделения. Эти

вхождения мы называем «однозначными ошибками». Вторая группа представляет два класса заимствований, к выделению которых различные стандарты CWS подходят по-разному; поэтому данная группа ошибок (по сути, являющаяся не в полном смысле слова «ошибками») названа «неоднозначными вхождениями».

Однозначные ошибки можно разделить на две категории — случаи чрезмерной токенизации (заимствование разделяется на большее, чем следует, количество слов) и недостаточной токенизации (в заимствование включаются соседние иероглифы из «стандартных» китайских слов).

Случаи чрезмерной токенизации можно объединить в несколько групп. Во-первых, часто выделяются иероглифы, выполняющие роль грамматических маркеров, расположенные после заимствований (например, локативный послелог 里, *lǐ* или результативный маркер глагола 来, *lái*). К более примечательным случаям относятся два типа вхождений. Так, заимствования из трех иероглифов, занимающие именные позиции в китайских предложениях, нередко делятся на два слова — односложное (первое) и двусложное (второе). Мы предлагаем объяснять этот тип ошибок структурой китайского личного имени: большинство китайских антропонимов состоит из односложной фамилии и следующего за ним двусложного личного имени. Так, транслитерация фамилии «Рогожин» разделяется сегментаторами PKUSeg и Ckiptagger на два слова: 罗 *luō* (китайская фамилия Ло) и 戈任 *gērèn* (может служить личным именем Гэжень).

Другой интересный тип ошибок связан с сегментацией четырехсложных заимствований: нередко они делятся на два двусложных слова. Мы предполагаем, что это следствие частотности двусложных китайских слов, которые, согласно [Wong, Xu 2010: 45], составляют 75 % китайского лексикона. Согласно ряду работ, например, [Duanmu 1999], тяготение китайского лексикона к двусложным словам объясняется фонотактическими причинами. Итак, сталкиваясь с последовательностью из четырех слогов (графически соответствующих четырем иероглифам), алгоритм может посчитать, что более правдоподобным будет разделение на два двусложных слова, а не на одно четырехсложное.

Явления недостаточной токенизации обычно происходят в случае, если не отделяется односложная лексема, идущая после длинного заимствования. На первый взгляд этот процесс кажется обратным чрезмерной токенизации одиночных иероглифов в конце заимствований. Однако подробный анализ выявляет обратное: случаи недостаточной

токенизации односложных лексем происходят с полнозначными глаголами (такими как 拿 *ná* «брать, поднимать» и 说, *shuō* «говорить»), в то время как чрезмерная токенизация происходит со служебными лексемами (например, с копулой 是 *shì*).

Обратимся теперь к неоднозначным вхождениям. Первый их тип связан с, пожалуй, единственным указателем на иностранные имена, принятым в китайской орфографии, — т. н. срединной точкой. Этот знак (·) употребляется для разделения имени собственного, которое в языке-доноре состояло из нескольких слов (в случае с русскими заимствованиями он будет ставиться, например, между именем и отчеством). Стандарт CNS предписывает разделять слова по этому знаку и считать их разными токенами, в то время PKU и CTB предлагают относиться ко всему транслитерированному кластеру, содержащему срединную точку, как к одному слову. При этом алгоритмы, реализующие эти стандарты, имеют тенденцию работать противоположным образом: UDPipe, Stanza и fastHan (версия udc), основанные на стандарте CTB, последовательно разделяют слова по срединной точке, а основанный на CNS алгоритм Ckiptagger этого никогда не делает.

Следующий тип неоднозначных вхождений связан с родовыми словами. Так называются односложные (в большинстве случаев) морфемы, которые нередко употребляются в постпозиции имени собственного с целью его более точного определения. Они занимают неоднозначную позицию с точки зрения как своих синтаксических, так и семантических характеристик. Эту нестабильность отражают и стандарты CWS: одни из них (государственный стандарт КНР, предшественник PKU) считают любые сочетания с родовым словом единым токеном, в то время как другие (например, PKU и CNS) дифференцируют разные родовые слова по разным признакам.

Алгоритмы в большинстве случаев не подчиняются стандартам, однако в их действиях можно найти две тенденции. Первая: чем длиннее родовое слово, тем вероятнее оно будет отделено от имени собственного. Так, двусложный элемент 森林 *sēnlín*, «лес», отделяется от предшествующего имени значительно чаще, чем односложный 村 *cūn*, «деревня». Вторая тенденция заключается в более частом выделении званий людей (которые занимают позицию родовых слов), нежели в выделении родовых слов, характеризующих топонимы. Так, 将军 *jiāngjūn*, «генерал», всегда отделяется от предшествующей фамилии, в то время как 城 *chéng*, «город», обычно образует единое целое с предшествующим названием.

Необходимо отметить, что две вышеописанные тенденции более всего совместимы с синтаксическим стандартом Penn Chinese Treebank, который предписывает отделять титулы людей и многосложные постпозитивные элементы компаундов и при этом не отделять односложные постпозитивные морфемы. Это любопытно, так как в данном случае за основу в выделении слов берутся не морфосинтаксические критерии, а фонотактические соображения, ведь стремление выделять двусложные родовые слова и не выделять односложные вновь напоминает нам тенденцию к употреблению двусложных токенов в китайском языке, рассмотренную выше.

3. Исследования в области морфосинтаксической аннотации

Как уже было отмечено, в отличие от других языков при разметке китайских текстов PoS-тэггинг не самая тривиальная задача. Это связано с тем, что существует несколько мнений о том, что такое части речи в китайском языке и как они выделяются (дискуссия по этому вопросу проиллюстрирована, например, в сборнике [Софронов 1989: 37–126]). Исходя из этого, существует большое количество стандартов морфопарсинга для китайского языка. Более того, в ряде систем PoS-тэггинга различают не только морфологические классы слов, но и семантические. Так, имена людей и названия географических объектов могут размечаться различными PoS-тэгами.

Мы проверили алгоритмы китайского PoS-тэггинга на корректность разметки заимствованных слов, к которым относятся имена людей и названия географических объектов. В качестве данных были взяты те же предложения, что и для оценки качества алгоритмов CWS (Раздел 2). Все предложения были разделены на две группы: содержащие топонимы и содержащие антропонимы (так как в ряде стандартов этим группам присваиваются различные PoS-тэги). Мы рассмотрели ряд алгоритмов, основанных на различных стандартах PoS-тэггинга китайских текстов; все они представлены в табл. 2.

Как и в случае со словоделением, о количественном исследовании подробно рассказано в [Семенов и др. 2021]. Отметим, что в этой задаче, аналогично CWS, явным лидером является fastHan.

Обратимся к качественному анализу ошибок PoS-тэггеров. Самые частотные из них — это присвоение заимствованным существительным PoS-тэгов глаголов, прилагательных и наречий. При этом в некоторых случаях можно попробовать интерпретировать подобные

Таблица 2. **Перечисление использованных стандартов и алгоритмов PoS-тэггинга**

Стандарт PoS-тэггинга	Инструмент
Chinese national standard (CNS)	Ckiptagger
Peking university (PKU)	PKUSeg
PKU (модифицированный)	NLPIR, LTP
Penn Chinese Treebank (CTB)	fastHan
Universal Dependencies + Penn Chinese Treebank (UPOS)	stanza, spacy

ошибки: например, Ckiptagger разметил транслитерацию топонима «Китеж» (基捷日 *jījiérì*) как «слово со значением времени» (отдельная морфологическая категория в китайском, тэг Nb), вероятно, потому, что в составе этого токена присутствует 日 *rì* («день, солнце»), составляющая частотная для временных слов.

Наконец, частотна ошибка, когда инструмент путает антропонимы и топонимы. При транслитерации фамилии «Гарин» (加林 *jiālín*) Ckiptagger отнес имя человека к топонимам (тэг Nc), вероятнее всего, из-за родового слова 林 *lín, «лес»*. NLPIR, наоборот, при транслитерации города «Псков» (普斯科夫 *pǔsīkēfū*) приписывает тэг антропонима nrf. Здесь можно предположить, что алгоритм введен в заблуждение морфемой 夫 *fū, «муж»,* которое нередко используется с личными именами или для образования профессий.

4. Заключение

В результате исследования мы сделали следующие выводы: с точки зрения логичности в области словоделения и морфосинтаксической аннотации наиболее подходящим для нашего корпуса является стандарт Penn Chinese Treebank, реализованный в CWS- и PoS-алгоритмах fastHan.

Учитывая, что финальной целью исследований является улучшение алгоритма лингвистической аннотации ruzhcorp, мы планируем встроить вышеописанные алгоритмы в систему препроцессинга китайских текстов. Тем не менее, нам предстоит доработать ряд алгоритмов (в частности, в области срединных точек), а также провести фундаментальные исследования в области родовых слов и их морфосинтаксической и фонотактической трактовки.

Кроме того, наша рабочая группа продолжает практические исследования и эксперименты в области разметки китайских текстов, а именно: дообучение наилучших моделей fastHan на наборах предложений из ruzhcorp; создание альтернативного модуля CWS, включающего нейросетевой модуль с определением смены кодов; эксперименты в области параллельной морфологической разметки (привлекая PoS-тэги соответствующих русских предложений).

Литература

1. *Семенов К. И., Короткова Ю. О., Коновалова А. С., Вольф Е. А.* (2021) [в печати], Автоматическая лингвистическая разметка китайских текстов, содержащих заимствования: Словоделение, транскрипция, PoS-тэггинг. DIALOG-2021: 27th International Conference on Computational Linguistics and Intellectual Technologies. M.
2. *Софронов М. В.* (ред.). (1989), Новое в зарубежной лингвистике. Вып. 22. Языкознание в Китае. М.: Прогресс.
3. *Che W., Feng Y., Qin L., Liu T.* (2021), N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models. arXiv:2009.11616 [cs]. URL: http://arxiv.org/abs/2009.11616 (дата обращения: 01.07.2021).
4. *Duanmu S.* (1999), Stress and the Development of Disyllabic Words in Chinese. Diachronica. Vol. 16(1), pp. 1–35. doi.org/10.1075/dia.16.1.03dua
5. *Emerson T.* (2005), The Second International Chinese Word Segmentation Bakeoff. URL: http://sighan.cs.uchicago.edu/bakeoff2005/ (дата обращения: 01.07.2021).
6. *Geng Z., Yan H., Qiu X., Huang X.* (2020), fastHan: A BERT-based Joint Many-Task Toolkit for Chinese NLP.arXiv:2009.08633 [cs]. URL: http://arxiv.org/abs/2009.08633 (дата обращения: 01.07.2021).
7. *Honnibal M., Montani I.* (2017), spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. URL: https://spacy.io/
8. *Li P.-H., Ma W.-Y.* (2019), CkipTagger. URL: https://github.com/ckiplab/ckiptagger (дата обращения: 01.07.2021).
9. *Luo R., Xu J., Zhang Y., Ren X., Sun X.* (2019), PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. arXiv:1906.11455 [cs]. URL: http://arxiv.org/abs/1906.11455 (дата обращения: 01.07.2021).
10. *Qi P., Dozat T., Zhang Y., Manning C. D.* (2018), Universal Dependency Parsing from Scratch. doi.org/10.18653/v1/K18-2016
11. *Straka M.* (2018), UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. 197–207. doi.org/10.18653/v1/K18-2020
12. *Wong K.-F., Xu R.* (2010), Introduction to Chinese natural language processing. Morgan & Claypool Publ.
13. 13.张[*Zhang*], 华平[*Huaping*], 商[*Shang*], 建云[*Jianyun*]. (2019), NLPIR-Parser：大数据语义智能分析平台[NLPIR-Parser: An intelligent semantic analysis toolkit for big data]. 语料库语言学[Corpus Linguistics]. Vol. 6(1), pp. 87–104.

References

1. *Che W., Feng Y., Qin L., Liu T.* (2021), N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models. arXiv:2009.11616 [cs]. URL: http://arxiv.org/abs/2009.11616 (date of access: 01.07.2021).
2. *Duanmu S.* (1999), Stress and the Development of Disyllabic Words in Chinese. Diachronica. Vol. 16(1), pp. 1–35. doi.org/10.1075/dia.16.1.03dua
3. *Emerson T.* (2005), The Second International Chinese Word Segmentation Bakeoff. URL: http://sighan.cs.uchicago.edu/bakeoff2005/ (date of access: 01.07.2021).
4. *Geng Z., Yan H., Qiu X., Huang X.* (2020), fastHan: A BERT-based Joint Many-Task Toolkit for Chinese NLP.arXiv:2009.08633 [cs]. URL: http://arxiv.org/abs/2009.08633 (date of access: 01.07.2021).
5. *Honnibal M., Montani I.* (2017), spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. URL: https://spacy.io/
6. *Li P.-H., Ma W.-Y.* (2019), CkipTagger. URL: https://github.com/ckiplab/ckiptagger (date of access: 01.07.2021).
7. *Luo R., Xu J., Zhang Y., Ren X., Sun X.* (2019), PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. arXiv:1906.11455 [cs]. URL: http://arxiv.org/abs/1906.11455 (date of access: 01.07.2021).
8. *Qi P., Dozat T., Zhang Y., Manning C. D.* (2018), Universal Dependency Parsing from Scratch. doi.org/10.18653/v1/K18-2016
9. *Semenov K. I., Korotkova Y. O., Konovalova A. S., Volf E. A.* (2021) [in print], Avtomaticheskaya lingvisticheskaya razmetka kitajskix tekstov, soderzhashhix zaimstvovaniya: Slovodelenie, transkripcziya, PoS-te`gging [Automatic Annotation of the Chinese Texts that Contain Loanwords: Word Segmentation, Transcription, PoS-tagging]. DIALOG-2021: 27th International Conference on Computational Linguistics and Intellectual Technologies, Moscow.
10. *Sofronov M. V.* (ed.) (1989), Novoe v zarubezhnoj lingvistike. Vy'pusk 22. Yazy'koznanie v Kitae [New Issues in Foreign Linguistics. Volume XXII: Chinese Language Science]. Moscow: Progress. 472 p.
11. *Straka M.* (2018), UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. 197–207. doi.org/10.18653/v1/K18-2020
12. *Wong K.-F., Xu R.* (2010), Introduction to Chinese natural language processing. Morgan & Claypool Publ.
13. *Zhang H., Shang J.* (2019), NLPIR-Parser: An intelligent semantic analysis toolkit for big data. In: Corpus Linguistics. Vol. 6(1), pp. 87–104.

Вольф Елена Александровна
Национальный исследовательский университет
«Высшая школа экономики» (Россия)
Volf Elena
HSE University (Russia)
E-mail: eavolf@edu.hse.ru

Короткова Юлия Олеговна
Национальный исследовательский университет
«Высшая школа экономики» (Россия)
Korotkova Yulia
HSE University (Russia)
E-mail: yuokorotkova@edu.hse.ru

Семенов Кирилл Игоревич
Институт проблем передачи информации РАН (Россия)
Semenov Kirill
Insitute for Information Transmission Problems of
the Russian Academy of Sciences (Russia)
E-mail: kir.semenow@yandex.ru

А. О. Гребенников, Т. Г. Скребцова

A. O. Grebennikov, T. G. Skrebtsova

КОРПУС РУССКИХ РАССКАЗОВ (1900–1930). УСТОЙЧИВОСТЬ ЛИНГВОСТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК[1]

CORPUS OF THE RUSSIAN SHORT STORIES (1900–1930). VALIDITY OF LINGVO-STASTISTICAL PARAMETERS

Аннотация. На материале выборки из корпуса русских рассказов оценивается устойчивость рангового среднего и энтропии, а также использование функций Вейбулла и Хауштайна для аппроксимации нарастания значений указанных параметров. Показано, что для подобного материала ранговое среднее является лишь относительно устойчивой характеристикой, а устойчивость энтропии зависит от характера текстов. Выбор функции Вейбулла является предпочтительным.
Ключевые слова. Частотные словари, словари языка писателя, статистическое моделирование, корпус, стилеметрия.

Abstract. Further to the previous experiments, validity of the rank mean and entropy for describing frequency dictionary of fiction as well as the use of Weibull and Haustein functions for the approximation of the dependence between sample size and resulting values of the parameters are analyzed. A representative sample from the Corpus of the Russian Short Stories (1900–1930) was chosen as the material for the experiment (total volume is more than 1 000 000 tokens). The rank mean is shown to be only a relative valid parameter for describing a large-scale corpus of fiction, while the relative validity of entropy is greatly affected by the nature of the texts analyzed. Weibull function is proved to be the preferable one for the approximation of the parameters growth.
Keywords. Frequency dictionary, authors' lexicography, statistical modeling, corpus, stylometry.

В последние годы в квантитативной лингвистике ведется поиск переменных и статистических метрик, способных описывать ранговые распределения лексем (в форме частотного словаря) или предсказывать изменение их характеристик с увеличением размера словника [Sherstinova et al. 2020]. Наши предыдущие исследования позволили установить состоятельность таких величин как ранговое среднее и энтропия для описания частотного словаря языка писателя, а также обосновать выбор аппроксимирующих функций [Гребенников 1998; Гребенников 2017; Гребенников и др. 2018].

[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)».

В рамках настоящего исследования будет проанализировано поведение указных характеристик на материале гораздо более значительного объема текстов, написанных при этом разными авторами.

С этой целью мы обратились к материалам корпуса русских рассказов, насчитывающего несколько тысяч русских рассказов, написанных в первые три десятилетия прошлого века [Мартыненко и др. 2018]. Материалы корпуса делятся на три временных среза: 1) довоенный период: начало XX века до Первой мировой войны (1900–1913), 2) военно-революционные годы: Первая мировая война, Февральская и Октябрьская революции и Гражданская война (1914–1922) и 3) советский период (1923–1930) [Гребенников и др. 2019].

Для настоящего исследования была использована выборка из корпуса объемом в 310 рассказов (приблизительно по 100 за каждый период, без каких-либо ограничений по длине, тематике и т. п. и не более чем по одному рассказу одного и того же автора). Рассказы в подкорпусе для каждого периода последовательно объединялись порциями по 10 и для этих объединений составлялись ранговые частотные словари лексем: 24 316 лексем, 376 513 словоформ для первого периода; 24 617 лексем, 303 588 словоформ для второго периода; 30 560 лексем; 383 430 словоформ для третьего периода и 124 081 лексема, 1 077 970 словоформ для выборки в целом. На каждом этапе с увеличением объема словаря фиксировались изменения значения рангового среднего (меры концентрации):

$$r = \sum r f_r / N, \quad (1)$$

где r — ранг, f_r — соответствующая этому рангу частота, N — объем выборки;
и энтропии (меры неопределенности):

$$H = -\sum p_i \log_2 p_i, \quad (2)$$

где p_i — вероятность, представляющая собой в случае частотного словаря отношение частоты слова к объему выборки.

Полученные результаты зависимости рангового среднего и энтропии от объема выборки (обнаруживающие постепенное стремление к неким верхним пределам) затем были аппроксимированы по авторизированной методике исследования, разработанной Г. Я. Мартыненко с использованием метода наименьших квадратов [Гребенников 1998].

Для аппроксимации были выбраны:

функция Вейбулла:

$$y = N_{\max} - N_{\max} e^{-cx^d}, \quad (3)$$

и функция Хаустайна:

$$y = \frac{N_{\max} x^{\gamma}}{x^{\gamma} + q}, \quad (4)$$

где N — объем словаря, x — объем выборки, N_{max} — асимптотический объем словаря, c, d, γ, q — параметры распределения [Гребенников 2017; Гребенников и др. 2019]. Полученные результаты и найденные путем аппроксимации теоретические максимумы исследуемых величин фрагментарно (ввиду ограниченного объема статьи) приводятся в табл. 1–4.

Таблица 1. **Результаты аппроксимации для довоенного периода**

Кол-во расск.	Ранг. среднее	Апп. по Вейбуллу (N_{max} = 1506,92)	Апп. по Хаустайну (N_{max} = 1607)	Энтропия	Апп. по Вейбуллу (N_{max} = 11,59)	Апп. по Хаустайну (N_{max} = 11)
10	1001,43	1153,24	1151,91	10,30	10,29	10,28
40	1422,45	1353,18	1349,00	10,43	10,56	10,56
70	1396,59	1422,97	1420,40	10,69	10,68	10,68
90	1389,53	1440,01	1439,42	10,71	10,72	10,72
100	1406,92	1448,39	1449,25	10,76	10,74	10,74

Таблица 2. **Результаты аппроксимации для военно-революционного периода**

Кол-во расск.	Ранг. среднее	Апп. по Вейбуллу (N_{max} = 1624,71)	Апп. по Хаустайну (N_{max} = 1690)	Энтропия	Апп. по Вейбуллу (N_{max} = 11,36)	Апп. по Хаустайну (N_{max} = 12)
10	775,73	860,14	846,13	10,07	10,05	10,05
40	1589,55	1490,31	1484,60	10,71	10,70	10,70
70	1550,61	1578,29	1569,20	10,85	10,85	10,85
90	1519,71	1603,86	1601,76	10,92	10,92	10,92

Таблица 3. **Результаты аппроксимации для советского периода**

Кол-во расск.	Ранг. среднее	Апп. по Вейбуллу (N_{max} = 1967,4)	Апп. по Хауштайну (N_{max} = 2193)	Энтропия	Апп. по Вейбуллу (N_{max} = 11,31)	Апп. по Хауштайну (N_{max} = 11)
10	970,50	1329,62	1333,11	10,41	10,42	10,39
40	1838,33	1633,34	1631,92	11,12	11,12	11,13
70	1682,17	1729,53	1727,29	11,23	11,23	11,23
90	1694,84	1761,74	1760,28	11,23	11,25	11,25
100	1692,77	1777,14	1776,37	11,27	11,26	11,26
110	1693,40	1779,45	1778,81	11,28	11,27	11,26

Таблица 4. **Результаты аппроксимации для всей выборки**

Кол-во расск.	Ранг. среднее	Апп. по Вейбуллу (N_{max} = 1542,5)	Апп. по Хауштайну (N_{max} = 1562)	Энтропия	Апп. по Вейбуллу (N_{max} = 11,37)	Апп. по Хауштайну (N_{max} = 14)
10	1001,43	1110,22	1072,49	10,30	10,06	10,20
40	1422,45	1326,61	1335,77	10,43	10,54	10,54
70	1396,59	1411,21	1418,31	10,69	10,74	10,71
100	1406,92	1444,71	1448,56	10,76	10,82	10,80
130	1437,83	1461,65	1463,57	10,45	10,87	10,85
160	1439,33	1475,62	1475,91	10,62	10,91	10,89
190	1461,84	1488,72	1487,57	10,71	10,95	10,94
220	1530,37	1496,76	1494,84	10,76	10,98	10,97
250	1500,79	1504,41	1501,92	10,81	11,01	11,01
280	1521,34	1510,92	1508,12	10,88	11,04	11,04
310	1530,37	1515,67	1512,81	10,91	11,06	11,07

Из таблиц видно, что прогностические кривые весьма близки друг к другу, иногда с буквальным совпадением значений анализируемых параметров. Для установления факта состоятельности и определения наиболее адекватной для аппроксимации функции нам необходимо

выяснить, возможно ли достижение найденных предельных значений на выборке реально осуществимого объема.

Установив асимптотические значения исследуемых параметров, мы, переходя к обратным функциям, можем вычислить, выборку какого объема требуется произвести, чтобы ее статистические характеристики были максимально близки к данным. Рассчитаем выборку для достижения 99 % от максимально возможного значения исследуемых параметров и для достижения их значения, отличного на единицу. Полученные результаты приводятся в табл. 5–6.

Таблица 5. **Объем выборки, необходимый для достижения максимальных значений рангового среднего при аппроксимации с помощью функций Вейбулла и Хаустайна**

Период	Выборка для достижения 99 % по Вейбуллу	Выборка для достижения 99 % по Хаустайну	Выборка для достижения отличия на единицу по Вейбуллу	Выборка для достижения отличия на единицу Хаустайну
Довоенный	935 629	18 225 456	11 183 249	1,72E+09
Военно-революционный	302 840	1 274 115	1 253 228	16 247 196
Советский	4 434 696	1,53E+09	158 457 657	5,61E+09
Вся выборка	1 518 023	4 379 998	20 282 353	118 588 851

Таблица 6. **Объем выборки, необходимый для достижения максимальных значений энтропии при аппроксимации с помощью функций Вейбулла и Хаустайна**

Период	Выборка для достижения 99 % по Вейбуллу	Выборка для достижения 99 % по Хаустайну	Выборка для достижения отличия на единицу по Вейбуллу	Выборка для достижения отличия на единицу Хаустайну
Довоенный	326 466 720	710 889 595	53 270 861 675	12 032 766
Военно-революционный	2 584 109	3,82212E+11	38 554 494	150 003 255
Советский	4 434 696	305 065	158 457 657	4 740 401
Вся выборка	4 978 764	6,24805E+18	68 388 69	1,56934E+29

При этом выбранные нами в качестве материала 320 рассказов составляют приблизительно 30 % нашего корпуса. Таким образом, весь словарь рассказов в корпусе составил бы свыше трех миллионов словоупотреблений. Следовательно, ни одна из анализируемых характеристик не является абсолютно состоятельной, т. е. достигающей своих предельных значений в рамках реально осуществимого объема выборки.

Ранговое среднее является лишь относительно состоятельной (т. е. допускающей 99-процентное «вычерпывание») характеристикой, причем со значительным разбросом объемов необходимого для этого выборки по периодам. Это подтверждает работоспособность идеи состоятельности при рассмотрении совокупностей, описываемых ранговыми распределениями, в частности, выборок значительного объема, охватывающих ряд отдельных произведений. Для энтропии даже относительная состоятельность подтверждается только для одного, военно-революционного, периода.

Данные выводы относятся только к использованию функций Вейбулла в качестве аппроксимирующей. Полученные с использованием функции Хаустайна объемы выборок, при которых ранговое среднее и энтропия достигают своих предельных значений, в большинстве случаев аномально велики.

Полученные результаты представляются вполне обоснованными, так как мы имеем дело с выборкой, хоть и однородной с точки зрения жанра, но крайне разнородной с точки зрения разброса тем, языковых средств и уровня таланта авторов входящих в нее текстов, тогда как наши предыдущие исследования всегда имели дело с текстами одного автора.

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, с. 151–156.
3. *Гребенников А. О., Ассель А. Н.* (2019), База русского рассказа XIX–XX веков. Модели аппроксимации. Корпусная лингвистика-2019: Труды международной конференции, с. 379–386.
4. *Гребенников А. О., Скребцова Т. Г.* (2019), Языковая картина мира в русском рассказе начала XX века. Философия и гуманитарные науки в информационном обществе. № 3(25), с. 82–92.

5. *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», с. 180–197.
6. *Sherstinova T., Martynenko G.* (2020), Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019). CEUR Workshop Proceedings. Vol. 2552, pp. 105–120.

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]. In: Korpusnaya lingvistika-2017. Trudy mezhdunarodnoj konferentsii [Corpus Linguititcs–2017. Proceedings of the International Conference].]. Saint Petersburg, pp. 151−156.
3. *Grebennikov A. O., Assel A. N.* (2019), Baza russkogo rasskaza XIX–XX vekov. Modeli approksimacii [Russian Short Stories Corpus of the XIX–XX Centuies. Approximation Models.]. In: Korpusnaya lingvistika-2019. Trudy mezhdunarodnoj konferentsii [Corpus Linguititcs–2017. Proceedings of the International Conference]. Saint Petersburg, pp. 379−386.
4. *Grebennikov A. O., Skrebtsova T. G.* (2019), Jazykovaja kartina mira v russkom rasskaze nachala XX veka [Language Picture of the World in Russian Short Stories of the Early XX Century]. In: Filosofija i gumanitarnye nauki v informacionnom obshhestve [Philosophy and Humanities in Information Society]. No. 3(25), pp. 82−92.
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 20th Century]. In: Trudy XV Mezhdunarodnoj konferentsii po komp'yuternoj i kognitivnoj lingvistike "TEL 2018" [Proceedings of the XV International Conference on Computer and Cognitive Linguistics 'TEL 2018']. Kazan, pp. 180−197.
6. *Sherstinova T., Martynenko G.* (2020), Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019). CEUR Workshop Proceedings. Vol. 2552, pp. 105–120.

Гребенников Александр Олегович
Санкт-Петербургский государственный университет (Россия)
Grebennikov Alexander
Saint Petersburg State University (Russia)
E-mail: a.grebennikov@spbu.ru

Скребцова Татьяна Георгиевна
Санкт-Петербургский государственный университет (Россия)
Skrebtsova Tatyana
Saint Petersburg State University (Russia)
E-mail: t.skrebtsova@spbu.ru

Д. О. Добровольский, Анна А. Зализняк

D. O. Dobrovol'skij, Anna A. Zalizniak

ПАРАЛЛЕЛЬНЫЙ КОРПУС КАК ИНСТРУМЕНТ СЕМАНТИЧЕСКОГО АНАЛИЗА: НЕМЕЦКИЙ МОДАЛЬНЫЙ ГЛАГОЛ *SOLLEN*[1]

THE PARALLEL CORPUS AS AN INSTRUMENT OF SEMANTIC ANALYSIS: THE GERMAN MODAL VERB *SOLLEN*

Аннотация. На материале немецкого глагола *sollen* демонстрируется применение «монофокусного» метода контрастивного корпусного анализа, суть которого состоит в том, что перевод на другой язык рассматривается как свидетельство значения анализируемой единицы. Показано, что обращение к множеству эквивалентов, представленных в переводных текстах, позволяет выявить варианты употребления этого глагола, не упоминаемые в его традиционных описаниях, и уточнить природу модального значения, выражаемого глаголом *sollen*.
Ключевые слова. Модальность, модальные глаголы, немецкий язык, русский язык, перевод, параллельный корпус.

Abstract. The given study is based on parallel corpus data concerning the use of the German verb *sollen* and its Russian equivalents. We demonstrate the advantages of the "monofocus" method of contrastive corpus analysis. The essence of this method is that a translation of a linguistic unit into another language is considered to be evidence of the meaning of the analyzed unit of the original language. It is shown that referring to the set of equivalents presented in translated texts makes it possible to identify variants of the use of the verb *sollen* that are not mentioned in its traditional descriptions.
Keywords. Modality, modal verbs, German, Russian, translation, parallel corpus.

1. Постановка задачи

Данная работа является частью нашего проекта по корпусному исследованию семантики конструкций с немецкими модальными глаголами; эмпирическая база проекта — немецко-русский параллельный корпус НКРЯ и созданная на его основе надкорпусная база данных, см. [Dobrovol'skij, Zalizniak 2018]. В рамках данного проекта мы обращаемся к параллельному корпусу как к инструменту для собственно семантического исследования, используя «монофокусный» метод семантического анализа, суть которого состоит в том, что перевод на другой язык рассматривается как свидетельство значения анализируемой единицы языка оригинала. В данном исследовании объектом

[1] Статья написана при частичной поддержке РФФИ, грант № 20-012-00166.

анализа является глагол *sollen*, обозначающий модальность особого типа, для которой в русском языке нет конвенционального способа выражения.

Как отмечалось многими исследователями, глагол *sollen* выражает «ослабленное долженствование»[2]. Мы предлагаем уточнение этого понятия, позволяющее определить место модального значения, выражаемого немецким глаголом *sollen*, среди других модальных значений. Если модальность *долженствования* означает, что положение дел P соответствует **единственно возможной** логике развития событий (устройства мира), модальность *возможности* — что среди возможных путей его развития **существует** такой, где имеет место P, то модальность, выражаемая немецким *sollen*, представляет собой соотнесение положения дел P с **некоторой логикой устройства мира** (развития событий), не единственно возможной, но в каком-то отношении выделенной. Глагол *sollen* во всех типах употребления выражает отсылку к этой логике. Наиболее явно это значение обнаруживает себя в специфическом значении «ослабленного онтологического долженствования», которому будет посвящено отдельное исследование. В данной статье мы покажем, каким образом «ослабленность» долженствования проявляется в другом его значении — деонтическом.

2. Деонтическое значение глагола *sollen*

Деонтический тип значения в общем виде представляет собой **установку** (обозначим ее ‘должно’) некоторого субъекта X относительно ситуации P(Y), где P — ситуация, обозначенная подчиненным глаголу *sollen* инфинитивом, Y — ее субъект. При этом X-ом может быть: 1) говорящий; 2) некоторое третье лицо (конкретное); 3) слушающий. Y-ом тоже может быть любой из трех, и при этом может совпадать или не совпадать с X. Все 9 комбинаций возможны и встречаются в корпусе. Во всех случаях возможен перевод *sollen* на русский язык с помощью модальных слов *должен, нужно, надо, следует*, при этом некоторые комбинации допускают дополнительно свои особые варианты перевода. Именно они будут интересовать нас в первую очередь.

[2] Ср. [Öhlschläger 1989: 172; Duden-Grammatik 2005: 565; Hentschel, Weydt 2003: 80]. Ср. также понятие weak obligation, применяемое к англ. *should* в [Leech 1971: 95; Wierzbicka 1987: 35; Bybee et al. 1994: 186]. Обычная, не «ослабленная», модальность долженствования выражается в немецком языке с помощью глагола *müssen*.

Во всех случаях, кроме двух типов контекстов, указанных ниже, глагол *sollen* выражает «ослабленное» долженствование: речь идет даже не о долженствовании в собственном смысле слова, а о рекомендации, намерении или предпочтении, основанном на соотнесении выбора поведения с представлением о выделенной среди других линией развития событий. Заметим, что из перечисленных выше «стандартных» русских эквивалентов только *следует* передает идею «ослабленного» долженствования, в остальных случаях этот признак в переводе утрачивается.

1. Субъектом установки ‘должно’ является **говорящий**, а подчиненный инфинитив обозначает действие или состояние самого говорящего или другого лица (в том числе слушающего), ср.: *Ich*/*er*/*du sollte*(*st*) *lieber schweigen — Мне*/*ему*/*тебе лучше бы промолчать*. В русском языке это значение, помимо вышеупомянутого «стандартного» набора, передается также словами *лучше* (*бы*), *хорошо бы*, *не мешает*, *стоит, придется*; при субъекте действия, отличном от говорящего, также глаголами *советовать*, *рекомендовать*, *просить*, формой сослагательного наклонения, инфинитивной конструкцией с дательным падежом субъекта. При субъекте действия 3-го лица возможна также конструкция со словом *пусть*; если субъектом действия является слушающий — императив и форма 2-го лица глагола *мочь*.

1.1. Субъектом зависимого инфинитива является говорящий:

(1) Ach, wir ***sollen*** *uns hinsetzen*, zum Teufel, und etwas leisten, wie unsere Vorfahren etwas geleistet haben… [Thomas Mann. Buddenbrooks (1896–1900)] — Черт возьми, ***лучше*** нам *пораскинуть мозгами* да добиться чего-нибудь в жизни, как того добивались наши предки… [Томас Манн. Будденброки (Н. Ман, 1953)][3]

1.2. Субъектом зависимого инфинитива является третье лицо:

(2) <…> ich konnte genug von ihrem Gesicht sehen, um zu wissen, dass sie die Tränen zurückhielt. Sie ***hätte*** *weinen* ***sollen***, heftig und lange. [Heinrich Böll. Ansichten eines Clowns (1963)] — <…> и все же понял, что она с трудом удерживается от слез. ***Лучше бы*** она *заплакала* — бурно, навзрыд. [Генрих Бёлль. Глазами клоуна (Л. Б. Черная, 1964)]

[3] Здесь и далее в примерах из параллельного корпуса глагол *sollen* и его переводной эквивалент мы выделяем п/ж курсивом, подчиненный глаголу *sollen* инфинитив выделен светлым курсивом. Примеры со ссылкой в квадратных скобках взяты из НКРЯ (www.ruscorpora.ru).

(3) <…> das ist ein herrliches Deutsch, und die Zeitungsleute ***sollten*** davon lernen. [Dieter Noll. Die Abenteuer des Werner Holt. Roman einer Heimkehr (1963)] — <…> какой это к тому же стилист, нашим газетчикам ***не мешает*** у него поучиться превосходному немецкому языку. [Дитер Нолль. Приключения Вернера Хольта (В. Курелла, Р. Гальперина, 1962)]

(4) „Oder Zigeuner", sagte ich, „Mutter ***sollte*** einmal welche zum Tee einladen". [Heinrich Böll. Ansichten eines Clowns (1963)] — Есть еще и цыгане, — сказал я. — ***Хорошо бы*** мама пригласила их к себе на файф-о-клок. [Генрих Бёлль. Глазами клоуна (Л. Б. Черная, 1964)]

1.3. Субъектом зависимого инфинитива является слушающий:

(5) „Du ***solltest*** wenigstens etwas von diesem Tomatenzeug *drauftun*", sagte er. [Heinrich Böll. Ansichten eines Clowns (1963)] — ***Советую*** тебе хотя бы *полить* томатным соком, — сказал он. [Генрих Бёлль. Глазами клоуна (Л. Б. Черная, 1964)]

(6) Du ***solltest*** *dich* wirklich *schämen*, Grace! [Kerstin Gier. Rubinrot (2009)] — *Постыдилась* ***бы***, Грейс! [Керстин Гир. Рубиновая книга (С. Вольштейн, 2012)]

(7) Aber einen Versuch ***sollst*** du machen. [Hermann Hesse. Narziß und Goldmund (1930)] — Но ты ***можешь*** сделать попытку. [Герман Гессе. Нарцисс и Гольдмунд (Г. Барышникова, 1993)]

2. Говорящий передает установку 'должно' **третьего лица.** В этом случае для всех трех вариантов субъекта действия эквивалентами для *sollen*, помимо модальных предикатов *должен*, *нужно*, *надо*, *следует*, могут быть глаголы интерперсональной каузации *велеть, требовать, просить,* а также (реже) *заставлять*, *предлагать*, *посылать* и некоторые другие.

2.1. Субъектом зависимого инфинитива является говорящий:

(8) Natürlich, er ist wirklich vom Fach — aber wenn er meint, ich ***sollte*** nach sechs Bühnenjahren noch *ein Studium anfangen* — Unsinn! [Heinrich Böll. Ansichten eines Clowns (1963)] — Ну конечно, это его профессия, но если он думает, что после шести лет работы на сцене мне ***следует*** снова *сесть за парту*… Какая чушь! [Генрих Бёлль. Глазами клоуна (Л. Б. Черная, 1964)]

(9) <…> wir ***sollen*** dich fragen, ob du Interesse hast. [Cornelia Funke. Herr der Diebe (2002)] — Нам ***велели*** спросить, интересует ли тебя такой заказ. [Корнелия Функе. Король воров (М. Л. Рудницкий, 2004)]

(10) Wie ich noch so auf sie hinsehe, fällts auf einmal der andern lustigen Dicken von meinen zwei Damen ein, ich ***sollte*** ihr während der Fahrt eins *singen*. [Josef von Eichendorff. Aus dem Leben eines Taugenichts (1826)] — Пока я на нее глядел, другой даме — веселой и дородной — пришло на ум ***попросить*** меня что-нибудь *пропеть*. [Йозеф фон Эйхендорф. Из жизни одного бездельника (Д. Усов, 1933–1935)]

2.2. Субъектом зависимого инфинитива является третье лицо:

(11) Als Leslie genug gelacht hatte, meinte sie, James ***solle*** mir lieber etwas anderes *beibringen*, und James war ausnahmsweise ihrer Meinung gewesen. [Kerstin Gier. Saphirblau (2009)] — Когда Лесли вдоволь насмеялась, она решила, что Джеймс ***должен*** *научить* меня чему-то другому, и, как ни странно, Джеймс был с ней на этот раз согласен. [Керстин Гир. Сапфировая книга (С. Вольштейн, 2013)]

2.3. Субъектом зависимого инфинитива является слушающий:

(12) Hier, Großmutter, das ist das Geld. Und Mutter lässt herzlich grüßen. Und du ***sollst*** *nicht böse sein*, dass sie in den letzten Monaten nichts geschickt hat. [Erich Kästner. Emil und die Detektive (1929)] — Вот, бабушка, возьми. И мама передает сердечный привет. И ***просит*** тебя *не сердиться*, что ничего не послала в прошлый месяц. [Эрих Кестнер. Эмиль и сыщики (Л. Лунгина, 1971)]

(13) Die Duksel hat Ihnen natürlich gesagt, Sie ***sollen*** *sich nix wissen machen*, wenn's heraus kommt. [Gustav Meyrink. Der Golem (1914)] — Эта стерва, конечно, ***велела*** вам *притвориться*, что вы ничего не знаете? [Густав Майринк. Голем (Д. Выгодский, 1922)]

Отношение говорящего к передаваемой им установке 'должно' другого лица может быть различным: он может солидаризироваться с этой установкой, например, принимать совет, но может относиться к ней нейтрально или даже резко негативно. Так, в примере (14) говорящий не согласен с передаваемой установкой, высмеивает ее (ср. также комментарий «Какая чушь!» в примере (8)):

(14) Du ***sollst*** ihm *den Stuhl vor die Tür setzen* — es ist zum Lachen. [Theodor Fontane. Effi Briest (1894–1895)] — Ты ***должен*** *выставить* его *за дверь* — вот смешно! [Теодор Фонтане. Эффи Брист (Г. Эгерман, Ю. Светланов, 1960)]

3. Субъектом установки 'должно' является **слушающий**, что возможно только в контексте диалога, когда говорящий интерпретирует

какое-то предшествующее высказывание своего собеседника или запрашивает его мнение. Субъект подчиненного инфинитива при *sollen,* как правило, второй участник этого диалога, то есть говорящий. Высказывания с субъектом действия — слушающим или третьим лицом — возможны только при наличии подчиняющего предиката мнения в форме 2-го лица. То, что субъект установки — слушающий, может быть выражено тремя способами:

1) эксплицитно (с помощью выражений типа *meinst du* — ср. пример (15));
2) подразумеваться или следовать из контекста (ср. пример (16));
3) в форме вопроса-предложения к слушающему *Soll ich* <inf.>*?* (примеры (17)–(18)).

(15) Leise fragte er: „Was, meinst du, ***soll*** ich *tun?*" [Hermann Hesse. Siddhartha (1922)] — Потом тихо проговорил: «Что же, по-твоему, я ***должен*** *делать?*» [Герман Гессе. Сиддхартха (Б. Д. Прозоровская, 1990)]

В примере (16) говорящий (шофер) обращает внимание собеседницы (Момо) на то, что именно она высказала ранее просьбу отвезти ее к себе домой; сам глагол *sollen* при этом на русский язык не переводится.

(16) „Zu Gigis Haus, bitte", antwortete Momo. Der Fahrer blickte etwas überrascht drein. „Ich denke, ich ***soll*** *dich zu dir nach Hause bringen.*" [пользуясь принятой нотацией: «X [=*ты*] считаешь, что 'должно' P [=*отвезти тебя домой*] (Y[=*я*]). — К дому Джиги, пожалуйста, — ответила Момо. Шофер озадаченно посмотрел на нее. — Я думал, что мы *поедем к тебе домой*. [Михаэль Энде. Момо (Ю. И. Коринец, 1982)]

(17) Конструкция *Soll ich* <inf.>*?* (без вопросительного местоимения) имеет значение 'говорящий предлагает произвести действие, которое, по его мнению, слушающий хотел бы, чтобы он произвел'. То есть говорящий как бы пытается угадать, что нужно слушающему. Специально отметим, что при наличии вопросительного местоимения реализуется другая конструкция — **вопрос**, который задает говорящий самому себе или адресату, ср. *Was soll ich tun — Что мне делать?*, *Wann soll ich da sein? — Когда мне прийти?* и т. п. (ср. пример (15)). А конструкция *Soll ich…*<inf.>*?* реализует речевой акт **предложения**, ср.: *Soll ich dir Tee eingießen? — Налить тебе*

чаю? (≅***Soll*** ich diese dicke Flasche bis zum Rand *vollfüllen*? [Patrick Süskind. Das Parfum: Die Geschichte eines Mörders (1985] — ***Хотите***, *я заполню до краев* вон ту толстую флягу? [Патрик Зюскинд. Парфюмер: История одного убийцы (Э. Венгерова, 1992)]

(18) ***Soll*** ich dir ein Taxi *bestellen*? [Heinrich Böll. Ansichten eines Clowns (1963)] — Может, *заказать* тебе такси? [Генрих Бёлль. Глазами клоуна (Л. Б. Черная, 1964)]

«Ослабленность» долженствования, выражаемого *sollen*, может утрачиваться в двух типах высказываний: содержащих отрицание с глаголом *sollen* в индикативе и в таких, где идея побуждения усиливается какими-то дополнительными средствами.

Отрицательное предложение с *sollen* в индикативе — это однозначный запрет. Это вытекает из семантики *sollen*: если утвердительное предложение с этим глаголом указывает на существование «некоторой линии развития событий», включающей ситуацию P, то отрицание этого утверждения означает, что такой линии развития событий не существует. Поэтому, например, в библейских заповедях используется именно глагол *sollen* (ср. *Du sollst nicht töten!*), а в русском, где аналогичного модального глагола нет, —– просто форма императива (*Не убий!*); ср. также:

(19) „Nein", sagte der Prügler und strich ihm mit der Rute derartig über den Hals, dass er zusammenzuckte, „du ***sollst*** *nicht zuhören*, sondern *dich ausziehen*". [Franz Kafka. Der Prozess (1914)] — Нет, — отрезал экзекутор и провел розгой по его шее так, что тот вздрогнул, — и вообще, ***не вмешивайся***, *а* ***раздевайся*** поскорее. [Франц Кафка. Процесс (Р. Райт-Ковалева, 1965)]

Снятие «ослабленности» долженствования происходит также при использовании средств, усиливающих категоричность рекомендации или запрета: это, прежде всего, частицы *ja* и *doch*; возможны также эксплицитные отсылки к тому, что данное высказывание нужно интерпретировать как приказ, типа *sage ich dir!*

3. Заключение

Проведенное исследование показывает, что обращение к материалу параллельного корпуса, предоставляющего в распоряжение исследователя реально использованные профессиональными пере-

водчиками переводные эквиваленты, позволяет не только выявить значительное количество потенциальных переводных эквивалентов, не представленных в словарях, но и существенно уточнить семантику рассматриваемой языковой единицы.

Помимо стандартных русских эквивалентов глагола *sollen* — модальных слов *должен, надо, нужно, следует*, которые возможны для всех типов соотношения субъекта установки и субъекта подчиненного инфинитива, в исследованном материале встречаются также глаголы и предикативы, модальный потенциал которых не всегда очевиден. Причем выбор того или иного эквивалента частично коррелирует с лицом субъекта установки и субъекта действия.

Глаголы «жесткой» интерперсональной каузации *велеть*, *требовать* и *заставлять* встречаются преимущественно в том случае, когда субъектом установки ‘должен’ является 3-е лицо, независимо от лица субъекта действия. Глаголы «мягкой» интерперсональной каузации *советовать*, *рекомендовать* и *просить* встречаются преимущественно в случае, когда субъектом установки является говорящий, но возможны также и при субъекте установки 3-го лица.

Только в случаях, когда субъектом установки ‘должен’ является говорящий, в переводе может появляться предикатив *лучше* (*бы*). Это не случайно, поскольку *лучше* (*бы*) выражает субъективную оценку говорящего; этот переводной эквивалент для *sollen* в наиболее явной форме передает семантику выделенности определенной линии развития событий среди других возможных. Исключительно установку говорящего передает также ряд других единиц, выражающих «ослабленное долженствование». Возможность использования в качестве переводного эквивалента для *sollen* глагола *мочь* подтверждает промежуточный между долженствованием и возможностью статус того типа модальности, который выражается глаголом *sollen*.

Наибольшими отличиями в наборе возможных способов передачи модального значения, выражаемого немецким *sollen*, характеризуется ситуация, когда установка ‘должен’ принадлежит слушающему, а субъектом действия является говорящий. В этом случае *sollen* может входить в конструкцию со значением речевого акта предложения, которая передается по-русски вопросительным предложением с модализованным инфинитивом или с глаголом *хотеть* во 2-м лице.

Литература

1. *Добровольский Д. О., Зализняк Анна А.* 2018. Немецкие конструкции с модальными глаголами и их русские соответствия: проект надкорпусной базы данных // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог» (2018). Выпуск 17 (24). М., 2018. С. 172–184.
2. *Bybee J., Perkins R., Pagliuca W.* (1994), The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World. Chicago.
3. Duden-Grammatik (2005), 7., völlig neu erarb. und erweiterte Aufl. Mannheim.
4. *Hentschel E., Weydt H.* (2003), Handbuch der deutschen Grammatik. 2., völlig neu bearb. Aufl. Berlin.
5. *Leech G.* (1971), Meaning and the English Verb. London.
6. *Öhlschläger G.* (1989), Zur Syntax und Semantik der Modalverben des Deutschen. Tübingen.
7. *Wierzbicka A.* (1987), The Semantics of Modality. Folia linguistica. Vol. 21 (1), pp. 25–43.

References

1. *Bybee J., Perkins R., Pagliuca W.* (1994), The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World. Chicago.
2. *Dobrovol'skij D., Zalizniak Anna.* (2018), German Constructions with Modal Verbs and their Russian Correlates: A Supracorpora Database Project, Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2018". Issue 17(24), pp. 172–184.
3. Duden-Grammatik (2005), 7., völlig neu erarb. und erweiterte Aufl. Mannheim.
4. *Hentschel E., Weydt H.* (2003), Handbuch der deutschen Grammatik. 2., völlig neu bearb. Aufl. Berlin.
5. *Leech G.* (1971), Meaning and the English Verb. London.
6. *Öhlschläger G.* (1989), Zur Syntax und Semantik der Modalverben des Deutschen. Tübingen.
7. *Wierzbicka A.* (1987), The Semantics of Modality. Folia linguistica. Vol. 21 (1), pp. 25–43.

Добровольский Дмитрий Олегович
Институт русского языка РАН (Россия)
Институт языкознания РАН (Россия)
Стокгольмский университет (Швеция)
Dobrovol'skij Dmitrij
Russian Language Institute of the Russian Academy of Sciences (Russia)
Institute of Linguistics of the Russian Academy of Sciences (Russia)
Stockholm University (Sweden)
E-mail: dobrovolskij@gmail.com

Зализняк Анна Андреевна
Институт языкознания РАН (Россия)
Институт проблем информатики РАН (Россия)
Zalizniak Anna
Institute of Linguistics of the Russian Academy of Sciences (Russia)
Institute of Informatics Problems of the Russian Academy of Sciences (Russia)
E-mail: anna.zalizniak@gmail.com

О. В. Донина

O. V. Donina

АВТОМАТИЗАЦИЯ ВЫГРУЗКИ РЕЗУЛЬТАТОВ ПОИСКА В КОРПУСАХ ЧЕТВЕРТОГО ПОКОЛЕНИЯ

AUTOMATION OF DOWNLOADING SEARCH RESULTS FROM FOURTH-GENERATION CONCORDANCERS

Аннотация. В рамках доклада предлагается рассмотреть постраничную выгрузку как наиболее оптимальный способ автоматизации выгрузки результатов поискового запроса в корпусах четвертого поколения (на примере корпуса М. Дэвиса NOW). Наш подход позволяет, с одной стороны, сэкономить время лингвистов, избавив их от временных затрат, необходимых на формулирование поискового запроса и перенос выдачи результатов поиска в удобный для последующего лингвистического анализа формат, а с другой – дает возможность изучать лингвистический материал во всей его полноте, не ограничивая количество выгружаемых корпусных контекстов.
Ключевые слова. Корпусы четвертого поколения, NOW Corpus, web scraping, Python.

Abstract. We propose to consider web scraping as the most efficient way to automatically download search query results from fourth-generation concordancers (using M. Davis's NOW corpus as an example). Our approach allows, on the one hand, to save linguists' time, and on the other hand, it makes it possible to study all contexts presented in the corpora without any limitation.
Keywords. Fourth-generation concordancers, NOW Corpus, web scraping, Python.

1. Корпусы четвертого поколения: новые возможности и новые трудности

Т. Макэнери и Э. Харди были выделены четыре поколения корпусных программных средств [McEnery et al. 2012], и хотя большинство современных инструментов корпусной лингвистики относится ими к третьему поколению, корпусная система М. Дэвиса (https://www.english-corpora.org/) рассматривается как инструмент четвертого поколения, позволяющий работать с большими объемами данных, в связи с размещением базы данных на веб-сервере и предварительной индексации материала. На момент написания статьи объемы корпусов М. Дэвиса варьируются от 100 млн до 14 млрд слов, что, с одной стороны, предоставляет широкие возможности для лингвистических исследований, но, с другой, обуславливает временные затраты на выгрузку результатов поиска в корпусах. Так, например, формулировка поискового запроса и перенос 1000 примеров выдачи результатов поиска из корпуса в таблицу формата .xls(x) для последующего лингвистического анализа занимает в среднем 25 минут. Учитывая объемы корпу-

сов, результаты выдачи могут исчисляться сотнями тысяч. Наиболее простым решением данной проблемы, связанной с трудоемкостью, может являться ограничение материала исследования первой тысячей контекстов поисковой выдачи, что, на наш взгляд, может привести к ограничению и нашего лингвистического знания. В связи с этим целью данной работы является предложить способ автоматизации выгрузки результатов поиска, который, во-первых, позволит сэкономить время исследователей, а во-вторых, даст возможность изучать лингвистический материал во всей его полноте.

Разработка загрузчика, необходимого для сбора данных из лингвистических корпусов, происходила на базе корпуса NOW (Newspapers on the Web) [Davies 2016], содержащего 12,2 млрд словоупотреблений в двадцати государственных вариантах английского языка и пополняющегося на 4 млн слов ежедневно.

2. Способы автоматизации выгрузки поисковых запросов

Наиболее распространенный вариант выгрузки данных с различных сайтов — через API. Как было выяснено при общении с командой М. Дэвиса, публичный API у них отсутствует в связи с необходимостью минимизации нагрузки на сервер. При этом они поощряют создание другими исследователями полуавтоматической выгрузки запросов (но технической поддержки для этого не предоставляют). В качестве альтернативных вариантов были предложены:

1) приобретение групповой подписки (стоимость: 300$ за 1 год для доступа 30 человек); нам этот вариант не подходит, так как групповая подписка только увеличивает количество ежедневных запросов, но функционал остается прежним, доступна только ручная выгрузка;
2) приобретение ограниченных архивных версий корпусов (стоимость одного корпуса варьируется от 245$ до 795$); и хотя в этом случае появляется возможность автоматизировать необходимые корпусные запросы, стоимость оказывается довольно высокой, и, кроме того, эта опция не предполагает получения обновлений, что неприемлемо при работе с корпусом NOW, который обновляется каждый день.

В качестве альтернативного варианта была рассмотрена возможность использования Macro Recorders (таких, как Auto Mouse Clicker,

EasyClicks Macros, TinyTasks, Mouse Record Premium, Macro Toolworks, Mini Mouse Macro, Ghost Mouse, AutoHotkey), которые запоминают движения мышью и нажатия клавиш клавиатуры (аналогично записи макросов) и повторяют их. В этом случае можно было бы вводить исследуемые лингвистические единицы в поля поиска вручную, а процесс копирования контекстов из корпуса в файл .xls(x) формата записать при помощи этой программы в виде последовательности действий и запускать этот процесс необходимое количество раз. Но в связи с полуавтоматическим характером этого процесса и только относительным его упрощением было принято решение отказаться от этого варианта.

3. Постраничная выгрузка результатов поискового запроса

В итоге наиболее соответствующим нашим задачам и материалу способом автоматизации сбора информации была признана постраничная выгрузка из корпуса (по сути, Web Scraping, т. е. сбор данных в сети, их последующая очистка и извлечение требующейся информации). Рассмотрим процесс реализации постраничной выгрузки из корпуса NOW.

Подход состоял в том, чтобы запрограммировать действия, производимые пользователем самостоятельно при ручной выгрузке. Для этого в «Инструментах разработчика» браузера были изучены запросы к серверу и получаемые ответы. Программируемые этапы имитации действий пользователя при ручной выгрузке и используемые инструменты отражены на рис. 1. Стоит отметить, что при реализации проекта было решено использовать универсальный мультипарадигменный скриптовый язык программирования Python, что, в первую

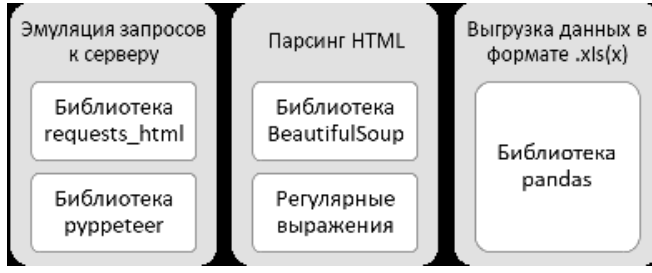


Рис. 1. **Этапы имитации действий пользователя при ручной выгрузке и используемые инструменты**

очередь, обусловлено наличием большого количества библиотек для решения разнообразных задач (в том числе связанных с автоматической обработкой естественного языка). Далее рассмотрим имитируемые этапы подробнее.

Первым этапом выступает эмуляция запросов к серверу, состоящая в свою очередь из трех запросов:

- Запрос ключа сессии (эмулирует вход пользователя на сайт),
- Ввод необходимых для выгрузки данных,
- Получение соответствующих запросу примеров.

Изначально для всех трех запросов использовалась библиотека requests_html для работы с HTTP, которая по сути является усовершенствованной версией разработанной ранее для аналогичных целей библиотеки urllib2. Библиотека requests_html позволяет выполнять запросы к серверу, обрабатывать его ответы и выгружать содержимое веб-страниц для последующего парсинга в виде HTML. Но при использовании данной библиотеки возникла проблема, заключающаяся в том, что сервер возвращал данные в формате, отличном от того, что есть на сайте. Было принято решение использовать беззаголовочный браузер pyppeteer (headless chrome browser automation library), способный эмулировать работу браузера. До недавнего времени для решения подобных задач требовалось прибегать к использованию таких проектов, как PhantomJS. С появлением беззаголовочных браузеров появилась возможность визуализировать и анализировать веб-страницы без использования пользовательского интерфейса (UI — user interface), получая тот же результат, что и в традиционном режиме с UI. Пионером в области автоматизации действий веб-браузера и удаленного управления браузером считается инструмент Selenium, разработанный на Java. Pyppeteer же, позволяющий управлять браузером из кода Python с помощью относительно простого и высокоуровневого API, можно рассматривать как современную альтернативу использованию традиционного Selenium. Pyppeteer позволяет получить практически полный контроль над браузером Chrome, в том числе открывать вкладки, в реальном времени анализировать объектную модель документа (DOM), выполнять Javascript и многое другое. В связи с тем, что в рамках решения нашей задачи не было очевидно, на каком именно запросе возникают проблемы, мы пробовали использовать данную библиотеку для всех трех запросов. Опытным путем было показано, что возникающая проблема связана с первым запросом. Таким обра-

зом, в итоговой версии кода для получения ключа сессии (т. е. при первом запросе) теперь используется беззаголовочный браузер pyppeteer, а для второго и третьего запросов продолжает применяться библиотека requests_html (из-за простоты использования).

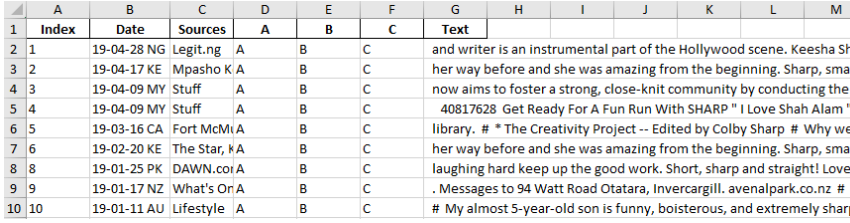
На втором этапе для извлечения требующихся данных с веб-страниц использовалась библиотека для XML/HTML-парсинга BeautifulSoup, которая, согласно информации разработчиков, способна преобразовывать даже неправильную разметку (tag soup), откуда и получила свое название (с этой точки зрения BeautifulSoup считается надежнее, чем ее более оперативно работающий аналог lxml). Несмотря на то, что, согласно документации, библиотека BeautifulSoup способна определять местоположение закрывающихся тегов даже при их отсутствии, в процессе работы у нас возникла именно такая проблема: тег Font открывался, но не закрывался. В связи с этим, чтобы убрать невалидные элементы HTML, которые мешали парсингу BeautifulSoup, мы использовали регулярные выражения, удалив все теги Font.

На последнем этапе для выгрузки данных в формате .xls(x) мы воспользовались библиотекой для обработки и моделирования данных — pandas. Несмотря на то, что данная библиотека предоставляет множество способов анализа данных (в том числе: группировка, создание сводных таблиц, визуализация при помощи графиков (при наличии matplotlib) и многое другое), для решения нашей задачи была актуальна возможность чтения и записи всех самых распространенных форматов данных (например, файлов .xls(x), HTML, SQL, .txt и пр.), а также наличие объекта табличной структуры данных DataFrame, т. е. проиндексированного многомерного массива значений.

4. Результаты

В результате выполнения данного кода мы получаем файл формата .xls(x) (рис. 2), содержащий данные, аналогичные результатам, отражающимся во вкладке Context корпуса NOW, но в табличном формате, позволяющем проводить дальнейший лингвистический анализ. Возможности выдачи корпуса в выгружаемом нами формате ограничены контекстом из 30 слов во избежание нарушения авторских прав.

Стоит отметить масштабируемость данного подхода, который мы в дальнейшем планируем применять для остальных корпусов Марка Дэвиса, с которыми мы работаем (COCA, GloWbE, iWeb).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Index**	**Date**	**Sources**	**A**	**B**	**C**	**Text**						
2	1	19-04-28 NG	Legit.ng	A	B	C	and writer is an instrumental part of the Hollywood scene. Keesha Shar						
3	2	19-04-17 KE	Mpasho K	A	B	C	her way before and she was amazing from the beginning. Sharp, smart,						
4	3	19-04-09 MY	Stuff	A	B	C	now aims to foster a strong, close-knit community by conducting the Sh						
5	4	19-04-09 MY	Stuff	A	B	C	40817628 Get Ready For A Fun Run With SHARP " I Love Shah Alam " Fu						
6	5	19-03-16 CA	Fort McM	A	B	C	library. # * The Creativity Project -- Edited by Colby Sharp # Why we lo						
7	6	19-02-20 KE	The Star, K	A	B	C	her way before and she was amazing from the beginning. Sharp, smart,						
8	8	19-01-25 PK	DAWN.cor	A	B	C	laughing hard keep up the good work. Short, sharp and straight! Love fr						
9	9	19-01-17 NZ	What's On	A	B	C	. Messages to 94 Watt Road Otatara, Invercargill. avenalpark.co.nz # SH						
10	10	19-01-11 AU	Lifestyle	A	B	C	# My almost 5-year-old son is funny, boisterous, and extremely sharp. I						

Рис. 2. **Сгенерированный файл формата .xls(x), содержащий результаты корпусного запроса**

Литература

1. *Davies M.* (2016), NOW Corpus. URL: http://corpus.byu.edu/now/.
2. *McEnery T., Hardie A.* (2012), Corpus Linguistics: method, theory and practice, Cambridge University Press.

References

1. *Davies M.* (2016), NOW Corpus. URL: http://corpus.byu.edu/now/.
2. *McEnery T., Hardie A.* (2012), Corpus Linguistics: method, theory and practice, Cambridge University Press.

Донина Ольга Валерьевна
Воронежский государственный университет (Россия)
Donina Olga
Voronezh State University (Russia)
E-mail: olga-donina@mail.ru

А. В. Захарова

A. V. Zakharova

ВОЗМОЖНОСТЬ КОРПУСА: ОТПРАВНЫЕ ТОЧКИ И ОПЫТ СОЗДАНИЯ КОРПУСОВ МУЛЬТИМОДАЛЬНЫХ ТЕКСТОВ

THE POSSIBILITY OF A CORPUS: BACKGROUND AND PRACTICE OF BUILDING MULTIMODAL TEXT-IMAGE CORPORA

Аннотация. В статье описывается один из современных, активно развивающихся подходов к созданию корпусов мультимодальных визуальных текстов. Приведены базовые понятия и концепции, кратко описана методика аннотирования материала, показан исследовательский потенциал корпусного представления мультимодальных текстов. Предполагается, что изучение опыта создания корпусов в рамках этого подхода может представлять особый интерес для изучения эффективности коммуникации.
Ключевые слова. Мультимодальность, мультимодальные тексты, корпусная лингвистика, речевые жанры, коммуникативная эффективность.

Abstract. The article contains a brief introduction to a currently developing approach towards analyzing multimodal visual documents within the Genre and Multimodality framework. This approach includes building corpora of visual documents as a tool to identify patterns in their structure, assess coherence within multimodal texts and afterwards contribute to genre topology. The author lists fundamental concepts and ideas, describes a method for corpus annotation and briefly displays research potential of the approach stated as well as examples of corpora of visual documents created worldwide. It is proposed that the body of work with multimodal corpora may significantly shift the current research of communicative effectiveness.
Keywords. Multimodality, multimodal texts, corpus linguistics, speech genres, communicative efficiency.

Большинство текстов, представленных в привычной визуальной коммуникации, задействуют сразу несколько систем выразительных средств, которые также называют модусами в рамках развивающейся в зарубежной лингвистике концепции мультимодальности [Kress 2010]. Это касается как более традиционных печатных текстов, так и цифровых. Тем не менее, представляется, что взаимоотношения между различными системами выразительных средств в пределах одного текста привлекали сравнительно мало внимания и находились скорее на периферии лингвистики, в области семиотики, на границе с психологией, социологией, культурологией.

Хотя концепция мультимодальности применялась для исследования самых разных знаковых систем и их сочетаний (музыки, изо-

бразительного искусства, кино [Van Leeuwen 1999; Bateman, Schmidt 2011]) и даже изучения коммуникации обезьян [Waller et al. 2013], в данной работе нас интересует, как в рамках мультимодальности разрабатывается процедура анализа сложных (креолизованных) визуальных текстов, одновременно задействующих вербальные, графические и иные средства и системы для передачи информации. В современной лингвистике выделяется кластер англоязычных работ, посвященных этой проблематике, рассмотренной с позиции мультимодальности. Его ядро представлено работами британского лингвиста Джона Бейтмана, ученого и практика Роберта Воллара и их коллег и соавторов, например, [Bateman 2008; Bateman et al. 2018; Waller 2017].

Большая часть этих работ нацелена на практические результаты. В основе лежит положение о том, что взаимоотношение выразительных систем в конкретных текстах напрямую влияет на их эффективность (т. е. успешность выполнения коммуникативной задачи) для намеченного адресата или группы адресатов. Предполагается, что исследования в этой области необходимы как для оценки эффективности самих текстов в коммуникации, так и для оценки конкретных решений, принятых во время создания, оформления и «производства» текстов. Для разработки этого положения Джон Бейтман предлагает, во-первых, использовать понятие **жанра** как синтетическое, объединяющее разные способы выражения значений в текстах, и, во-вторых, проводить **корпусные исследованиия** таких текстов [Bateman 2008: 50–65]. Создание корпусов мультимодальных текстов было бы полезно не только для анализа способов представления информации, но и для изучения их восприятия и понимания, что является актуальной прикладной задачей и привлекает внимание академических исследователей коммуникации и практических специалистов из самых разных областей.

Согласно подходу Бейтмана [Bateman 2015: 221–222], в рамках концепции мультимодальности исследуется совместное использование выразительных средств разной природы (визуальных, аудиальных, вербальных, графических и пр.) и их взаимодействие в формировании сообщений. Первоисточником идеи анализа мультимодальных текстов признаются работы Г. Кресса и Т. ван Левена (в частности, [Kress, Van Leeuwen 2001]), в свою очередь обращавшихся к концепции социальной семиотики М. Хэллидея [Halliday 1978]. В изложении Бейтмана, эти авторы положили начало изучению текстов в широком понимании, не входивших в число канонических объектов исследования

лингвистики и включавших фильмы, живопись, веб-сайты и пр. В основе анализа Г. Кресса и Т. ван Левена лежало представление о грамматике как системе выбора выразительных средств и их сочетания для передачи смысла и достижения нужного эффекта в коммуникации.

Представление о жанре в работах Бейтмана и других рассматриваемых авторов, которое используется в разработке метода анализа мультимодальных текстов, в явном виде заимствует два критерия из теории речевых жанров. Первый — общность структуры текстов — прямо соответствует одному из трех критериев выделения жанра в работах М. М. Бахтина (общность темы, стилистики и композиции), второй — коммуникативная задача — выработан в работах А. Вежбицкой (см. также [Гиндин 2015]). В неявном виде применяется также критерий общности стилистических средств. Кроме названных, важнейшим критерием выделения жанра признается воспроизводимость и узнаваемость типов текстов в некотором сообществе. Этот критерий выходит на первый план при оценке эффективности: привычный жанр формирует у адресата ожидания как относительно содержания текста, так и модели взаимодействия с ним.

Гипотезу о применимости понятия жанра к мультимодальным текстам (и как таковую возможность их жанровой классификации) предполагалось проверить на специально отобранном материале, включающем печатные и цифровые тексты. В качестве ограничительного критерия для технического удобства была выбрана условно страничная организация текстов (page-based texts). Для систематизации и анализа на основе собранного материала предполагалось создание тестового исследовательского корпуса, что подразумевало разработку системы аннотирования сложных, нелинейно организованных мультимодальных текстов.

При разработке схемы представления материала было решено использовать анализ структуры текстов, методы качественного контент-анализа, изучение их риторической организации и связности. Мультимодальный текст представляли состоящим из нескольких слоев, каждый из которых соответствовал уровню анализа со своим методическим аппаратом и формой представления:

1. На первом уровне (base layer) происходило выделение базовых, визуально отделимых друг от друга текстовых (предложения, абзацы, заголовки, подписи и т. д.) и графических элементов (изображения, рисунки, иконки и т. д.) — минимальных составляющих (base units).

2. На следующем уровне — композиционном (layout layer) — составлялось описание выделенных составляющих в соответствии с их расположением относительно других элементов (группировка по близости) и на странице (расположение в пространстве). Результат анализа представляли в виде древовидной структуры. Также на этом уровне описывались особенности оформления составляющих.
3. Далее происходило определение смысловых (риторических) отношений между элементами в терминах теории риторической структуры [Mann, Thompson 1988] и составлялась их схема (rhetorical layer).
4. Наконец, на последнем, «навигационном» уровне (navigation layer) отмечались особые элементы, поддерживающие структуру целого текста и создающие интертекстуальные связи (нумерация, отсылки и т. д.) внутри целого текста или некоторого гипертекста.

Кроме подтверждения гипотезы о жанровой классификации мультимодальных текстов, с помощью такой схемы описания материала предполагалось проверить, насколько композиционная и риторическая структуры соответствуют друг другу и как это влияет на успешное понимание текстов. Иными словами, насколько важна поддержка смысловой структуры пространственным расположением и оформлением текстовых и графических элементов для правильной интерпретации целого текста адресатом. Забегая вперед, скажем, что в процессе анализа материала действительно были выявлены случаи несоответствия, которые потенциально могут привести к ошибкам при восприятии и понимании текстов, а также при усвоении информации, содержащейся в них, что показывает возможность использования такого рода корпусного материала в прикладных исследованиях эффективности обучающих материалов и информирования в широком понимании [Bateman 2017].

Сейчас известно не менее десяти исследований на основе локальных корпусов разноформатных мультимодальных материалов: туристических брошюр, постеров, веб-страниц и пр. (краткий обзор см. [Hiippala 2017]). В большинстве случаев тексты размечались вручную, что представляло известную сложность в масштабировании подхода и создания одного общего корпуса мультимодальных текстов. Для ряда современных исследований разметка корпусов также проводится вручную. Тем не менее, благодаря накопленному опыту рабо-

ты с такими текстами, а также развитию технологий и возможностей коллективных исследований (например, совместная разработка открытого кода программного обеспечения, краудсорсинг и пр.), стали появляться попытки автоматизации системы аннотирования текстов. В частности, они включают использование технологии компьютерного зрения для разметки базовых элементов текстов и их группировки, алгоритмов распознавания текстовых фрагментов и возможностей совместной работы с результатами анализа на каждом уровне (например, [Hiippala 2016]). Тем не менее, даже при такой локальной автоматизации каждый этап работы пока предполагает настройку и верификацию исследователем и определенный объем ручной разметки.

Несмотря на сложность быстрого набора и обработки материала для мультимодальных корпусов, уже известны примеры их применения для позитивного решения исследовательских задач разного рода: разработка упомянутой выше проблемы классификации жанров печатных и цифровых сложных текстов; сравнения применения выразительных средств для решения коммуникативных задач — как синхронно, так и в диахронии; изучения соотношения текста и изображения; проблемы перевода и культурной адаптации мультимодальных текстов; оценки эффективности текстов в коммуникации и др. В перспективе потенциал и качество корпусов такого рода можно значительно усовершенствовать при обогащении современных наработок богатым теоретическим и практическим научным аппаратом, а также, как представляется, опытом отечественной теории текста и корпусной лингвистики.

Литература

1. *Гиндин С. И.* (2015), Болевые точки теории речевых жанров. Русский язык сегодня. Вып. 6: Речевые жанры современного общения. М.: Флинта: Наука, с. 55–61.
2. *Bateman J.* (2008), Multimodality and genre: A foundation for the systematic analysis of multimodal documents. Springer.
3. *Bateman J.* (2017), Multimodality and genre. Issues for information design. In: Information design research and practice. London: Routledge, pp. 221–241.
4. *Bateman J., Wildfeuer J., Hiippala T.* (2017), Multimodality: Foundations, research and analysis. A problem-oriented introduction. De Gruyter Mouton.
5. *Bateman J., Schmidt K.-H.* (2011), Multimodal Film Analysis: How Films Mean. London: Routledge.
6. *Halliday M.* (1978), Language as social semiotic: The social interpretation of language and meaning. Maryland: University Park Press.

7. *Hiippala T.* (2016), Semi-automated annotation of page-based documents within the Genre and Multimodality framework. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. The Association for Computational Linguistics, pp. 84–89.
8. *Hiippala T.* (2017), An overview of research within the Genre and Multimodality framework. In: Discourse Context Media, pp. 276–284.
9. Kress G. (2010), Multimodality: A Social Semiotic Approach to Contemporary Communication. New York: Routledge.
10. *Kress G., Van Leeuwen T.* (2001), Multimodal Discourse: The Modes and Media of Contemporary Communication. Oxford UK: Oxford University Press.
11. *Mann W. C., Thompson S. A.* (1988), Rhetorical structure theory: Toward a functional theory of text organization. In: Text-Interdisciplinary Journal for the Study of Discourse. Vol. 8, No. 3, pp. 243–281.
12. *Van Leeuwen T.* (1999), Speech, Music, Sound. London: Palgrave MacMillan.
13. *Waller B. M., Liebal K., Burrows A. M., Slocombe K. E.* (2013), How can a multimodal approach to primate communication help us understand the evolution of communication? In: Evolutionary Psychology. Vol. 11(3), pp. 538–549.
14. *Waller R.* (2017), Practice-based perspectives on multimodal documents: corpora vs connoisseurship. In: Discourse, Context and Media. Vol. 20, pp. 175–190.

References

15. *Bateman J.* (2008), Multimodality and genre: A foundation for the systematic analysis of multimodal documents. Springer.
16. *Bateman J.* (2017), Multimodality and genre. Issues for information design. Information design research and practice. London: Routledge, pp. 221–241.
17. *Bateman J., Wildfeuer J., Hiippala T.* (2017) Multimodality: Foundations, research and analysis. A problem-oriented introduction. De Gruyter Mouton.
18. *Bateman J., Schmidt K.-H.* (2011), Multimodal Film Analysis: How Films Mean. London: Routledge.
19. *Gindin S. I.* (2015), Bolevye tochki teorii rechevyh zhanrov [The pain points of the theory of speech genres]. In: Russkij jazyk segognja. Vyp. 6: Rechevye zhanry sovremennogo obshhenija [Russian today. Vol. 6: Speech genres of modern communication]. Moscow: Flinta, Nauka Publ., pp. 55–61.
20. *Halliday M.* (1978), Language as social semiotic: The social interpretation of language and meaning. Maryland: University Park Press.
21. *Hiippala T.* (2016), Semi-automated annotation of page-based documents within the Genre and Multimodality framework. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. The Association for Computational Linguistics, pp. 84–89.
22. *Hiippala T.* (2017), An overview of research within the Genre and Multimodality framework. In: Discourse Context Media, pp. 276–284.
23. Kress G. (2010), Multimodality: A Social Semiotic Approach to Contemporary Communication. New York: Routledge.

24. *Kress G., Van Leeuwen T.* (2001), Multimodal Discourse: The Modes and Media of Contemporary Communication. Oxford UK: Oxford University Press.
25. *Mann W. C., Thompson S. A.* (1988), Rhetorical structure theory: Toward a functional theory of text organization. In: Text-Interdisciplinary Journal for the Study of Discourse. Vol. 8, No. 3, pp. 243–281.
26. *Van Leeuwen T.* (1999), Speech, Music, Sound. London: Palgrave MacMillan.
27. *Waller B. M., Liebal K., Burrows A. M., Slocombe K. E.* (2013), How can a multimodal approach to primate communication help us understand the evolution of communication? In: Evolutionary Psychology. Vol. 11(3), pp. 538–549.
28. *Waller R.* (2017), Practice-based perspectives on multimodal documents: corpora vs connoisseurship. In: Discourse, Context and Media. Vol. 20, pp. 175–190.

Захарова Анна Викторовна
Центр теории текста и лингвистического обеспечения коммуникации
Российского государственного гуманитарного университета (Россия)
Zakharova Anna
Center for Text Theory and Linguistic Support of Communication of
Russian State University for Humanities (Russia)
E-mail: zaharova.av@rggu.ru

С. С. Земичева, А. А. Васильченко

S. S. Zemicheva, A. A. Vasilchenko

ДРУГ, ПОДРУГА, ТОВАРИЩ, ЗНАКОМЫЙ: СПЕЦИФИКА СЕМАНТИКИ И ФУНКЦИОНИРОВАНИЯ В ДИАЛЕКТНОЙ РЕЧИ (ПО КОРПУСНЫМ ДАННЫМ)[1]

DRUG, PODRUGA, TOVARISHH, ZNAKOMYJ ('FRIEND'): SPECIFICITY OF SEMANTICS AND FUNCTIONING IN DIALECT SPEECH (BASED ON CORPORA DATA)

Аннотация. Цель статьи – выявить своеобразие использования номинаций дружбы в речи носителей традиционной культуры. Основным инструментом исследования является Томский диалектный корпус объемом более 2 000 000 словоупотреблений, для сопоставления используются данные других корпусных ресурсов. Новизна работы обусловлена обращением к исследованию функционирования общерусских слов в диалектном дискурсе с использованием корпусных методик. Проанализировано более 600 контекстов. На основе анализа частотности впервые выявлена гендерная специфика функционирования обозначений дружбы.
Ключевые слова. Диалектный корпус, Томский диалектный корпус, русские говоры Сибири, лексико-семантическое поле «Дружба».

Abstract. The purpose of the article is to reveal the originality of the use of friendship nominations in the speech of members of traditional culture. The main research tool is the Tomsk dialect corpus with a size of more than 2,000,000 tokens; data from other corpora are used for comparison. The novelty of the work is due to the appeal to the study of the semantics of all-Russian words in dialect discourse using corpus techniques. More than 600 contexts have been analyzed. Based on the analysis of frequency, the gender specificity of the functioning of friendship nominations has been revealed for the first time.
Keywords. Dialect corpus, Tomsk dialect corpus, Siberian dialects of Russian, lexico-semantic field "Friendship".

1. Введение

В 2021 г. исполняется 20 лет с момента публикации на русском языке книги А. Вежбицкой «Понимание культур через посредство ключевых слов», где «дружба» рассматривается как один из национально специфичных концептов, по-разному представленный в английской, русской, польской, немецкой, австралийской культурах [Вежбицкая 2001]. С тех пор лексика дружбы многократно изучалась в разных

[1] Исследование выполнено за счет гранта Российского научного фонда (проект № 19-78-10015 «Разработка электронных ресурсов для исследования народно-речевой культуры Среднего Приобья»).

аспектах. Подробный обзор работ по данной проблематике сделан Т. В. Леонтьевой [Леонтьева 2016]. Особое внимание ученые уделяли описанию различий в семантике синонимических единиц, связанных в русском литературном языке с понятием «дружба» [Вежбицкая 2001: 106–135; Урысон 2003; Шмелев 2005].

В то же время семантика и функционирование этих лексем за пределами литературного языка изучены в меньшей степени. Некоторые наблюдения на материале разговорной речи и жаргона сделаны в работе О. А. Араповой и Р. М. Гайсиной [Арапова, Гайсина 2005]. В диссертации Е. В. Коняевой проанализированы изменения концепта «дружба» в научном, обыденном и обыденно-научном типах сознания носителей русского языка на основе данных ассоциативного эксперимента [Коняева 2015]. Т. В. Леонтьевой проведен мотивационный анализ диалектных единиц, обозначающих понятия «друг», «дружба», «дружить», в русских говорах [Леонтьева 2011]. Словарный состав обозначений дружбы в русских говорах исследовала также Т. И. Вендина [Вендина 2020: 505–512].

Цель данной статьи — выявить своеобразие использования общерусских единиц, входящих в лексико-семантическое поле «дружба», *в* диалектном дискурсе. Под диалектным дискурсом понимается речь сельских жителей, записанная в ходе диалектологических экспедиций. Яркими особенностями этого дискурса являются устный характер его бытования, а также принадлежность говорящих к традиционной народно-речевой культуре.

2. Материал исследования

Материалом для работы послужили контексты со словами *друг, подруга, товарищ, знакомый* в Томском диалектном корпусе [ТДК]. Корпус представляет собой электронный ресурс на основе архива диалектных записей, сделанных с 1947 по 2019 г., насчитывающий в настоящее время более двух миллионов словоупотреблений. В корпусе реализована возможность поиска по лемме, а также по экстралингвистическим параметрам (год, место записи, возраст, пол, уровень образования информанта), по теме и жанру. Запись материалов производилась на территории западносибирского региона (Томской и Кемеровской областей). В зону анализа вовлечены, в основном, записи, сделанные с конца 1970-х по 2019 год. В отдельных случаях привлекались более ранние материалы. Основная часть записей представля-

ет собой не лингвистические опросники, а разговоры «на свободные темы», приближенные к естественной коммуникации. Информантами в большинстве случаев являются женщины в возрасте от 55 лет и старше. По уровню образования информантов в выборку попали следующие категории: неграмотные — около 22 %, имеющие начальное образование — 19 %, неполное среднее — 17 %, полное среднее или среднее специальное — 8 %, высшее — 2 % (в остальных случаях данные об образовании отсутствуют). Всего проанализировано более 600 высказываний. Во второй части работы для сопоставительного анализа использовались количественные данные других электронных ресурсов — Национального корпуса русского языка [НКРЯ], Устьянского диалектного корпуса [Даниэль и др. (2013-2018)].

3. Семантика и функционирование лексем

3.1. Друг и подруга

Т. В. Леонтьева отмечала, что в традиционной культуре дружба «имеет социальный характер, в отличие от современной культуры, которой свойственно осмысление категорий дружбы преимущественно в контексте межличностных отношений» [Леонтьева 2011: 204–205]. Эта мысль подтверждается и при анализе диалектного дискурса, причем на разных уровнях.

Во-первых, сами информанты неоднократно говорят о том, что в деревне все дружат друг с другом: [*А друзей детства помните?*] *Ну, как* ***друзей.*** *Все* ***местные*** *все друзья были* (с. Монастырка, 2018). Такие представления типичны и объясняются тем, что в традиционном деревенском сообществе выживание отдельного человека практически невозможно без опоры на общину.

Еще одним аргументом, подтверждающим социальный характер дружбы, является неупотребительность прилагательных *единственный, близкий, лучший в* сочетании со словами *друг/подруга* (см. табл. 1).

Для оценки статистической значимости различий дополнительно использовалась мера Log-Likelihood, рассчитываемая по формуле:

$$2\left(a\ln\left(\frac{a}{E1}\right)+b(\ln\left(\frac{b}{E2}\right)\right), \quad (1)$$

где a — абсолютное число употреблений конструкции в интересующей нас коллекции (в данном случае ТДК), b — число словоупотреблений в НКРЯ.

Таблица 1. **Абсолютная и относительная частотность отдельных словосочетаний с номинациями дружбы**

	НКРЯ	ТДК	НКРЯ ipm	ТДК ipm
Прилагательное	*друг/подруга*/всего			
единственный(ая)	265/23/288	0/1/1	0,89	0,48
близкий(ая)	2223/194/2465	1/1/2	7,6	0,95
лучший(ая)	1433/242/165	1/0/1	5,2	0,48
Всего слов в корпусе	321 712 061	2 098 152		

Величина E1 рассчитывается по формуле:

$$c\frac{a+b}{c+d}. \quad (2)$$

где c — объем исследуемого корпуса (ТДК), d — объем НКРЯ.

Величина E2 рассчитывается по формуле:

$$d\frac{a+b}{c+d}. \quad (3)$$

Проверка показала, что значимые отличия в частотности словосочетаний имеются только для конструкций с прилагательным *близкий* (LL = 20), в других случаях различия статистически не значимы (*лучший* LL = 0,05, *единственный* LL = 0,75).

3.2. Знакомый

Если люди, связанные с говорящим общностью территории проживания, «автоматически» воспринимаются как друзья, то жители других населенных пунктов обозначаются как *знакомые*: *Я с матерью приехал в церкву, у ее знакомый был, начальник почты.* (с. Нарым, 1972). *Возьметь кадочку масла и… какой-то Александр Александрович там был продавец в Кривошеине. Знакомый, в общем, свой магазин у него был* (с. Новокривошеино, 1982). Из 85 употреблений слова *знакомый* в функции существительного не выявлено ни одного, где речь шла бы о жителях своего села. Таким образом, *знакомство,* как и в литературном языке, предполагает менее близкие отношения, однако в диалектной речи отличия дружбы и знакомства напрямую связаны с пространственными категориями.

3.3. Товарищ

А. Д. Шмелёв отмечал, что «прототипическим для этого слова является обозначение людей, объединенных на основе «мужской солидарности, обусловленной «боевым товариществом» или совместным участием в одном деле» [Шмелёв 2005]. В диалектной речи это прототипическое значение в значительной степени сохраняется: *Мы часа в четыре туда пришли, ишшо' темновато было, и вот, наверно, перед обедом, одного моего товарища ранило, второго товарища, попала мина, перерезала живот, насмерть сразу* (п. Белый Яр, 1988). *Приходилось на медведя [охотиться]. Поташшы'лись с товарищем* (д. Баранаково, 1971). *Вот закупит лошадей купец и по Иркутскому тракту ездил с товарищами на Восток* (с. Старая Шегарка, 1982). Слово *товарищ* в анализируемых материалах практически всегда характеризует именно мужскую дружбу, даже если используется в женской речи: *А тут была то'ко записка, товарищ его прислал, что Николай погиб* (с. Зырянское, 1982).

4. Частотность лексем: гендерные особенности

«Дружба» в диалектном дискурсе оказывается преимущественно женским концептом: 85 % всех примеров зафиксировано в женских текстах. При этом частотность конкретных номинаций довольно сильно различается в речи мужчин и женщин. Для верификации полученных наблюдений были сделаны подсчеты частотности исследуемых слов в других диалектных корпусах. Так как объем выборки в каждом из диалектных корпусов относительно небольшой, материалы были обработаны вручную (учитывались только примеры, относящиеся к семантическому полю «Дружба», отсеивались контексты с омонимией). Результаты, представленные в табл. 2, демонстрируют, что номинации *подруга* и *подружка* действительно используются чаще в речи женщин, но и слова *друг и товарищ* также оказываются «женскими». Учитывая специфику материала диалектных корпусов, можно предположить, что это обусловлено их несбалансированностью (преобладанием женских текстов). Однако определить объем подкорпусов мужской и женской речи по отдельности оказалось невозможным ни в диалектном подкорпусе НКРЯ, ни в Устьянском корпусе.

В ТДК различия между частотностью обозначений в речи мужчин и женщин статистически значимы (см. табл. 3).

Таблица 2. **Абсолютная частотность обозначений в диалектных корпусах**

	ТДК		НКРЯ Диал.		Устьянский	
Пол/слово	М.	Ж.	М.	Ж.	М.	Ж.
подруга	13	107	–	33	6	36
подружка	13	194	–	1	1	14
товарищ	35	57	7	8	3	26
друг	50	81	1	9	24	30
Объем корпуса	2 098 152		395 440		1 245 162	
Объем подкорпусов	489 884	1 462 989	?	?	?	?

Таблица 3. **Коэффициент логарифмического правдоподобия LL score**

подруга	32
подружка	82
товарищ	22,5
друг	29

5. Выводы

Проведенный анализ позволяет уточнить сделанные ранее наблюдения о специфике концепта «Дружба» в диалекте. Подтверждается вывод о социальном, а не межличностном характере дружбы, что отражается в редкости словосочетаний, обозначающих дружбу «малым кругом». Установлено, что отличия дружбы и знакомства в диалекте связаны с пространственными категориями: жители своего села обычно обозначаются как друзья, жители других населенных пунктов — как знакомые. Показано, что слово *товарищ* в диалектной речи характеризует преимущественно мужскую дружбу, сохраняя архаические черты семантики. Выявлена гендерная специфика функционирования обозначений дружбы. Лексемы *подруга, подружка* чаще используются в речи женщин, а *товарищ, друг — в* речи мужчин.

Литература

1. *Арапова О. А., Гайсина Р. М.* (2005), Дружба. Антология концептов. Под ред. В. И. Карасика, И. А. Стернина. Т. 1, с. 58–80.
2. *Вежбицкая А.* (2001), Понимание культур через посредство ключевых слов. М.
3. *Вендина Т. И.* (2020), Антропология диалектного слова. СПб.
4. *Даниэль М., Добрушина Н., фон Вальденфельс Р.* (2013–2018), Говор бассейна Устьи. Корпус севернорусской диалектной речи. URL: www.parasolcorpus.org/Pushkino (дата обращения: 03.03.2021).
5. *Коняева Е. В.* (2015), Модификаты концепта «Дружба» в русском языке конца XX — начала XXI вв.: автореф. дис. … канд. филол. наук. Екатеринбург.
6. *Леонтьева Т. В.* (2016), Лексика дружбы: перспективы изучения. Социокультурное пространство России и зарубежья: общество, образование, язык. № 5, с. 71–78.
7. *Леонтьева Т. В.* (2011), Средства номинации понятий «друг», «дружба», «дружить» в русских народных говорах: мотивационный анализ. Язык. Текст. Дискурс. № 9, с. 196–207.
8. НКРЯ — Национальный корпус русского языка. URL: http://www.ruscorpora.ru (дата обращения: 01.03.2021).
9. ТДК — Томский диалектный корпус. Лаборатория общей и сибирской лексикографии НИ ТГУ. URL: http://losl.tsu.ru/corpus (дата обращения: 28.02.2021. Режим доступа: для зарегистрированных пользователей).
10. *Урысон Е. В.* (2003), Друг, товарищ, приятель. Новый объяснительный словарь синонимов русского языка, с. 297–299.
11. *Шмелев А. Д.* (2005), Дружба в русской языковой картине мира. Ключевые идеи языковой картины мира, с. 289–303.

References

1. *Arapova O. A., Gajsina R. M.* (2005), Druzhba [Friendship]. In: Antologija konceptov [Anthology of concepts]. Vol. 1, pp. 58–80.
2. *Daniel M., Dobrushina N., fon Waldenfels R.* (2013–2018), Govor bassejna Ust'i. Korpus severnorusskoj dialektnoj rechi [The language of the Ustja river basin. A corpus of North Russian dialectal speech]. URL: www.parasolcorpus.org/Pushkino (date of access: 03.03.2021).
3. *Konjaeva E. V.* (2015), Modifikaty koncepta "Druzhba" v russkom jazyke konca XX — nachala XXI vv. [Modifications of the concept "Friendship" in the Russian language of the late XX — early XXI centuries], Abstract of Philology Cand. Diss, Ekaterinburg.
4. *Leont'eva T. V.* (2016), Leksika druzhby: perspektivy izucheniya [Friendship vocabulary: study perspectives]. In: Sotsiokul'turnoye prostranstvo Rossii i zarubezh'ya: obshchestvo, obrazovaniye, yazyk [Sociocultural space of Russia and abroad: society, education, language]. No. 5, pp. 71–78.
5. *Leont'eva T. V.* (2011), Sredstva nominacii ponjatij "drug", "druzhba", "druzhit'" v russkih narodnyh govorah: motivacionnyj analiz [Means of nominating the concepts

"friend", "friendship", "befriend" in Russian folk dialects: motivational analysis]. In: Jazyk. Tekst. Diskurs [Language. Text. Discourse]. No. 9, pp. 196–207.
6. Natsional'nyy korpus russkogo yazyka [Russian National Corpus]. URL: http://ruscorpora.ru/search-main.html (date of access: 01.03.2021).
7. *Shmelev A. D.* (2005), Druzhba v russkoj jazykovoj kartine mira [Friendship in the Russian linguistic world-view]. In: Kljuchevye idei jazykovoj kartiny mira [Key ideas of the linguistic world-view in Russian]. Moscow, pp. 289–303.
8. Tomskij dialektnyj korpus [Tomsk dialect corpus]. Laboratorija obshej i sibirskoj leksikografii NI TGU [Laboratory of General and Siberian Lexicography, TSU]. URL: http://losl.tsu.ru/corpus (date of access 25.02.2021).
9. *Uryson E. V.* (2003), Drug, tovarishh, prijatel'. Novyj objasnitel'nyj slovar' sinonimov russkogo jazyka [Friend, comrade, buddy. New explanation Comprehensive dictionary of synonyms of the Russian language]. Moscow, pp. 297–299.
10. *Vendina T. A.* (2020), Antropologija dialektnogo slova [Anthropology of the dialect word]. Saint Petersburg.
11. *Vezhbickaja A.* (2001), Ponimanie kul'tur cherez posredstvo kljuchevyh slov [Understanding cultures through their key words]. Moscow.

Земичева Светлана Сергеевна
Томский государственный университет (Россия)
Zemicheva Svetlana
Tomsk State University (Russia)
E-mail: optysmith@gmail.com

Васильченко Анна Анатольевна
Томский государственный университет (Россия)
Vasilchenko Anna
Tomsk State University (Russia)
E-mail: annavasilchenko.95@mail.ru

В. И. Зубов, Е. И. Риехакайнен

V. I. Zubov, E. I. Riekhakaynen

ВМЕСТЕ ИЛИ ВРОЗЬ: НЕОДНОСЛОВНЫЕ ЕДИНИЦЫ В КОРПУСАХ И В МЕНТАЛЬНОМ ЛЕКСИКОНЕ НОСИТЕЛЯ РУССКОГО ЯЗЫКА[1]

TOGETHER OR NOT: MULTIWORD UNITS IN CORPORA AND IN THE MENTAL LEXICON OF A RUSSIAN SPEAKER

Аннотация. В статье описывается первый этап исследования, направленного на определение статуса неоднословных единиц в ментальном лексиконе носителя русского языка. На материале справочников по русской орфографии и текстов Национального корпуса русского языка начиная с 1956 года составлен список из наиболее употребительных двусловных сочетаний, имеющих пары среди однословных единиц. Этот список (201 пара) может использоваться в практике преподавания русского языка. В качестве наиболее вероятных кандидатов на вхождение в ментальный лексикон носителя русского языка выделены семь неоднословных сочетаний, которые являются высокочастотными или сопоставимы по частотности со своими однословными вариантами. Следующим этапом исследования станет анализ фонетических реализаций отобранных единиц в естественной русской устной речи.
Ключевые слова. Неоднословные единицы, русский язык, ментальный лексикон, Национальный корпус русского языка, орфография.

Abstract. The paper describes the first stage of a study of multiword units in the mental lexicon of a Russian speaker. Based on the guides on Russian spelling and the texts from the Russian National Corpus since 1956, we compiled a list of 201 most common two-word combinations that have pairs among one-word units. Seven two-word units, which are high-frequency or comparable in frequency with their one-word counterparts, were considered the most likely candidates for entering the mental lexicon of a Russian speaker. The next step will be the analysis of the phonetic realizations of these units in natural Russian speech.
Keywords. Multiword units, Russian, mental lexicon, Russian National Corpus, orthography.

1. Предпосылки исследования

Для решения одной из наиболее актуальных задач современной психолингвистики — описания структуры ментального лексикона (внутреннего словаря) носителя языка — необходимо среди прочего ответить на вопрос о единицах ментального лексикона. Представляется, что методы корпусной лингвистики и существующие на данный момент корпусы устных и письменных текстов могут быть использованы для поиска ответа на этот вопрос. В статье будет описан начальный этап исследования, посвященного определению статуса некоторых неоднословных единиц в ментальном лексиконе носителя русского языка.

[1] Исследование выполнено при поддержке гранта РФФИ № 19-012-00629.

В работах по порождению и восприятию речи активно обсуждается, могут ли самостоятельными единицами ментального лексикона являться сочетания нескольких слов. В качестве кандидатов в такие единицы предлагаются, например, коллокации, т. е. сочетания знаменательных слов, которые характеризуются частичной невыводимостью [Ellis, Simpson-Vlach 2009 и др.], или т.н. составные слова / эквиваленты слова, которые представляют собой сочетания нескольких орфографических слов, но при этом функционально приближаются к словам и часто произносятся как единое целое (например, *потому_что, то_есть* и др.) (см. обзор русскоязычных работ в [Мустайоки, Копотев 2004]).

На наш взгляд, еще одной группой неоднословных единиц, которые хранятся в ментальном лексиконе носителя языка как самостоятельные единицы, могут быть сочетания двух и более орфографических слов, имеющие пары среди однословных единиц (и *так — итак*, *на счет — насчет*), т. е. отличающиеся от последних только наличием пробела. Можно предполагать, что фонетическая близость подобных сочетаний к отдельным словам будет способствовать их объединению в одну единицу с соответствующими однословными единицами на уровне перцептивного словаря, т. е. того уровня ментального лексикона, на котором представлен фонетический облик слов (подробнее см. в [Венцов 2007]). Для проверки этого предположения необходимо проведение комплексного исследования, соединяющего в себе методы корпусной лингвистики и психолингвистики, а именно нужно выяснить, действительно ли пары, различающиеся только наличием/отсутствием пробела, всегда реализуются одинаково в естественной устной речи, и если это так, как происходит выбор необходимой интерпретации в процессе восприятия речи.

Исследования на материале русского языка, в которых сопоставляются именно пары «одно графическое слово — два графических слова», остаются единичными. Насколько мы можем судить, на данный момент в русскоязычной традиции отсутствует даже устоявшийся термин для обозначения подобных пар единиц. В [Ягунова 2008] такие пары выделяются в особые фонетические слова, имеющие омоним. В поэзии для близкого явления используется термин «пантограмма», когда строки полностью совпадают по буквенному составу, но различаются расстановкой пробелов (*ночей — но чей*; *задело — за дело*) [Бубнов 2002]. Если принимать во внимание орфографическую близость таких пар, то их можно отнести к т. н. орфографическим со-

седям — близким по написанию словам, которые отличаются только одним графическим элементом, а именно наличием/отсутствием пробела. Однако обычно к соседям относятся слова с перестановкой, добавлением или удалением букв; обсуждаемые же единицы на данный момент не включены в базу данных орфографических соседей в русском языке [Алексеева, Слюсарь 2017]. Таким образом, первым этапом нашего исследования стало формирование списка слов, которые станут объектом изучения.

2. Принципы формирования списка

В качестве источников для формирования первоначального списка неоднословных единиц, которые могут претендовать на самостоятельность в ментальном лексиконе, мы воспользовались справочниками по русской орфографии [Лопатин 2009; Розенталь 2011], где перечислены слова, слитное/раздельное написание которых вызывает сложности у носителей языка (*вбок — в бок; помногу — по многу*). В первую очередь нас интересовали служебные части речи, а также наречия и сочетания самостоятельных слов с предлогами. Можно предположить, что возникновению затруднений в выборе корректного написания способствует как раз то, что сочетание двух орфографических слов, совпадающее по буквенному составу с одним словом, претендует на некоторую степень устойчивости в ментальном лексиконе носителей языка. На данном этапе мы сосредоточились только на слитном или раздельном написании, поэтому слова с дефисным написанием не учитывались.

Слова из орфографических справочников были дополнительно проверены по Орфографическому академическому ресурсу «АКАДЕМОС» (http://orfo.ruslang.ru/). Так, пары типа в *общем — вобщем, по-суху — по суху в* список включены не были, потому что нормативным является единственный вариант написания. Для пар же типа *вперегиб — в перегиб, внакладку — в накладку* слитное написание является нормативным, а также существуют слова типа *перегиб*, *накладка*, которые могут быть употреблены с предшествующим предлогом. Всего в список была включена 271 пара, имеющая слитный или раздельный варианты написания. Часть из них однозначно по-разному реализуется в устной речи, поскольку единицы, образующие пару, имеют разные ударения (*впервые — в первые; втихую — в тихую*) — таких пар 47, они были исключены из дальнейшего анализа. Все оставшиеся в списке

единицы были проверены на наличие в основном подкорпусе Национального корпуса русского языка (далее — НКРЯ) начиная с 1956 года, то есть с года принятия Правил русской орфографии и пунктуации. Для части низкочастотных слов мы не встретили ни одного словоупотребления, например: *вперегиб — в перегиб*, *вперегонку — в перегонку*, *вприхватку — в прихватку*, *изнизу — из низу*; в некоторых парах одно из слов было представлено в корпусе, другое — нет: *навырез* (0)[2] — *на вырез* (3), *наудалую* (5) — *на удалую* (0), *навыкат* (29) — *на выкат* (0). Всего таких пар было 23, и они также были исключены из анализа.

Оставшаяся 201 пара вошла в основной список единиц, реализация которых в устной речи представляет интерес для дальнейшего исследования. Хотя частеречный анализ единиц не входил в задачи нашей работы на данном этапе, можно отметить, что в группу слов со слитным написанием вошли предлоги (*ввиду*, *насчет*), союзы (*зато*, *тоже*), наречия (*впустую*, *наутро*), вводные слова (*например*), а также слова, которые могут являться несколькими частями речи (*навстречу* — предлог, наречие, *оттого* — союз, наречие). В группе слов с раздельным написанием подавляющее большинство — сочетания самостоятельных слов с предлогами *в*, *до*, *за*, *из*, *на*, *от*, *под*, *по*, *при*, *с*, а также местоимения с частицами или союзами и *так*, *так же*, *что бы*. Слова со слитным написанием встречаются в корпусе значительно чаще слов с раздельным написанием (среднее значение частотности — 64,52 ipm и 6,59 ipm соответственно). Отметим, однако, что подобное соотношение справедливо для 161 пары (более частотными являются, например, *чтобы*, *потом*, *сейчас*, *причем*, *например* и др.), для остальных 40 раздельный вариант представлен чаще (*с ходу, с плеча, в начале*, *от того* и др.).

Одна из проблем, с которой мы столкнулись при подсчете частотности, — это орфографические ошибки в корпусе. Например, для единицы *на долго* из 25 случаев употребления встретилось всего пять с верным написанием (например, *похожим* ***на долго*** *показываемую фигу*), остальные 20 — орфографические ошибки (*там* ***на долго*** *зависать нельзя*), для единицы *на вынос* из 53 случаев употребления лишь 18 не содержат ошибки, остальные 35 — это некорректное написание наречия *навынос*. Подобные случаи, с одной стороны, свидетельствуют о сложности выбора верной орфографической формы для таких единиц, с другой стороны, могут быть косвенным подтверждением

[2] В скобках указано количество вхождений в подкорпус объемом 163 271 973 с/у.

того, что близкие по написанию слова представляют собой одну единицу перцептивного словаря. Отметим также, что для корректного определения частотности по основному подкорпусу НКРЯ необходимо исключить подобные ошибки, что весьма трудоемко в случае высокочастотных слов вроде *зато*, *притом* и т. п., поэтому решено было вычислить частотность и по подкорпусу НКРЯ со снятой омонимией, также начиная с 1956 года (объем корпуса — 4 759 671 с/у), где, как можно ожидать, все ошибки были устранены в ходе разрешения грамматической неоднозначности. Кроме того, мы указали для каждой единицы частотность в текстах устного подкорпуса НКРЯ за тот же период (целиком — 12 538 108 с/у и только со снятой омонимией — 205 994 с/у), что позволит сопоставить частоту употребления той или иной единицы в устной и письменной речи. Получившийся список, на наш взгляд, имеет самостоятельную практическую ценность: он может использоваться в практике преподавания русского языка, так как наглядно показывает, на какие пары сложных для написания единиц из представленных в справочниках по русской орфографии стоит обращать внимание при обучении русской орфографии в первую очередь в силу их высокой частотности. С перечнем единиц можно ознакомиться по ссылке: https://osf.io/bsy2t/.

3. Неоднословные сочетания — кандидаты в единицы ментального лексикона

Поскольку частотность является одним из ключевых факторов, влияющих на структуру ментального лексикона (см. обзор в [Риехакайнен 2016: 51–56]), мы предположили, что прежде всего на статус самостоятельных единиц ментального лексикона должны претендовать: 1) высокочастотные неоднословные единицы; 2) неоднословные единицы, которые сопоставимы по частотности с их однословными «соседями».

1) Высокочастотными мы считали двусловные сочетания из нашего списка, ipm которых хотя бы по одному из корпусов выше 100. Таких сочетаний оказалось четыре: *так же*, *и так*, *то же*, *в начале*, причем последнее имеет необходимую частотность только в основном подкорпусе. При этом соотношения по частотности внутри пар с этими сочетаниями оказались различными и в ряде случаев зависят от типа корпуса. Так, неоднословное сочетание частотнее однословного в парах *итак — и так* и *вначале — в начале*, но во второй паре

единицы в устной речи более близки по частотности, чем в письменной. В паре *тоже — то же* во всех подкорпусах намного частотнее однословный вариант. В паре *также — так же* в письменных текстах преобладает однословный вариант. В устном подкорпусе оба варианта практически не различаются по частотности, но если анализировать только устные тексты со снятой омонимией, неоднословный вариант будет иметь более высокую частотность.

2) Мы считали, что однословная и неоднословная единицы в паре имеют сопоставимую частотность, если их показатели ipm различались не более чем в два раза. Таких пар оказалось 24, но большинство из них являются редкими: только в трех парах (*зато — за то, оттого — от того* и *притом — при том*) вариант с раздельным написанием имеет ipm больше 10 хотя бы одном устном и письменном подкорпусе. Таким образом, мы включили в итоговый список неоднословных единиц, которые могут быть представлены в ментальном лексиконе носителя языка в качестве целостных единиц, семь сочетаний (см. табл. 1).

Таблица 1. **Список единиц для дальнейшего фонетического анализа**

Единица	ОП, ipm	УП, ipm	ОС, ipm	УС, ipm
также	499,52	190,06	451,92	179,62
так же	172,35	189,74	171,65	199,03
итак	64,80	98,26	55,68	58,25
и так	167,57	581,51	211,57	422,34
тоже	819,57	2003,57	959,52	1995,20
то же	159,68	229,38	144,34	237,87
вначале	33,10	35,01	29,83	29,13
в начале	101,35	71,30	72,06	38,84
зато	130,67	74,33	132,36	53,40
за то	66,90	89,25	75,43	77,67
оттого	50,33	21,77	52,10	9,71
от того	53,12	74,25	50,84	101,94
притом	17,59	15,39	19,12	19,42
при том	14,75	14,44	12,82	29,13

ОП, УП — основной и устный подкорпусы; ОС, УС — основной и устный подкорпусы со снятой омонимией.

Следующим этапом исследования станет сопоставление того, как двусловные сочетания и соответствующие им однословные единицы реализуются в устной речи. Если результаты фонетического анализа покажут, что однословные и неоднословные единицы в каждой паре имеют одинаковые варианты произнесения, то в дальнейшем в психолингвистических экспериментах можно будет проверить гипотезу о влиянии соотношения по частотности внутри каждой пары на то, каким образом носитель языка интерпретирует такие единицы при восприятии их на слух.

Литература

1. *Алексеева С. В., Слюсарь Н. А.* (2017), Орфографические соседи в русском языке: база данных и эксперимент, направленный на изучение морфологической декомпозиции. Вопросы психолингвистики. № 32, с. 12–27.
2. *Бубнов А. В.* (2002), Палиндромия: от перевертня до пантограммы. Новое литературное обозрение. № 57, с. 295–312.
3. *Венцов А. В.* (2007), Восприятие устной речи и ментальный лексикон. Русская языковая личность: Материалы шестой выездной школы-семинара. Череповец, с. 63–69.
4. *Лопатин В. В.* (ред.) (2009), Правила русской орфографии и пунктуации. Полный академический справочник. М.
5. *Мустайоки А., Копотев М.* (2004), К вопросу о статусе эквивалентов слова типа *потому что*, в *зависимости от*, к *сожалению.* Вопросы языкознания. № 3, с. 88–107.
6. *Риехакайнен Е. И.* (2016), Восприятие русской устной речи: контекст + частотность. СПб.
7. *Розенталь Д. Э.* (2011), Русский язык. Орфография и пунктуация. М.
8. *Ягунова Е. В.* (2008), Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь.
9. *Ellis N. C., Simpson-Vlach R.* (2009), Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. In: Corpus Linguistics and Linguistic Theory. Vol. 5, pp. 61–78.

References

1. *Alexeeva S. V., Slioussar N. A.* (2017), Orfograficheskie sosedi v russkom yazyke: baza dannyh i eksperiment, napravlennyj na izuchenie morfologicheskoj dekompozicii [Orthographic Neighbors in Russian: a Database and an Experiment Aimed at Studying Morphological Decomposition]. In: Voprosy Psiholingvistiki [Journal of Psycholinguistics]. No. 32, pp. 12–27.
2. *Bubnov A. V.* (2002), Palindromiya: ot Perevertnya do Pantogrammy [Palindromia:

From Shifter to Pantogram]. In: Novoe Literaturnoe Obozrenie [New Literary Observer]. No. 57, pp. 295–312.
3. *Ellis N. C., Simpson-Vlach R.* (2009), Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. In: Corpus Linguistics and Linguistic Theory. No. 5, pp. 61–78.
4. *Lopatin V. V.* (ed.) (2009), Pravila russkoj orfografii i punktuacii. Polnyj akademicheskij spravochnik [Rules of Russian spelling and punctuation. The Complete Academic Guide]. Moscow.
5. *Mustajoki A., Kopotev M.* (2004), K voprosu o statuse ekvivalentov slova tipa *potomu chto*, *v zavisimosti ot*, *k sozhaleniyu* [On the Status of Word-Equivalents of the Type *potomu chto*, *v zavisimosti ot*, *k sozhaleniyu*]. In: Voprosy Yazykoznaniya [Topics in the study of language]. No. 3, pp. 88–107.
6. *Riekhakajnen E. I.* (2016), Vospriyatie russkoj ustnoj rechi: kontekst + chastotnost' [Perception of Russian Oral Speech: Context + Frequency]. Saint Petersburg.
7. *Rozental' D. E.* (2011), Russkij yazyk. Orfografiya i punktuaciya [Russian. Orthography and Punctuation]. Moscow.
8. *Ventsov A. V.* (2007), Vospriyatie Ustnoj Rechi i Mental'nyj Leksikon [Oral Speech Perception and Mental Lexicon]. In: Russkaya Yazykovaya Lichnost': Materialy Shestoj Vyezdnoj Shkoly-Seminara [Russian Language Personality: Materials of the sixth field school-seminar]. Cherepovets, pp. 63–69.
9. *Yagunova E. V.* (2008), Variativnost' Strategij Vospriyatiya Zvuchashchego Teksta (Eksperimental'noe issledovanie na materiale russkoyazychnyh tekstov raznyh funkcional'nyh stilej) [Strategies Variability of the Oral Text Perception (An Experimental Study of Russian-Language Texts of Different Functional Styles)]. Perm'.

Зубов Владислав Иванович
Санкт-Петербургский государственный университет (Россия)
Zubov Vladislav
Saint Petersburg State University (Russia)
E-mail: v.zubov@spbu.ru

Риехакайнен Елена Игоревна
Санкт-Петербургский государственный университет (Россия)
Riekhakaynen Elena
Saint Petersburg State University (Russia)
E-mail: e.riehakajnen@spbu.ru

А. Е. Колесников, Л. А. Малахов

A. Ye. Colesnicov, L. A. Malahov

ДИГИТИЗАЦИЯ ПЕЧАТНЫХ МОЛДАВСКИХ ДИАЛЕКТНЫХ ЗАПИСЕЙ[1]

DIGITIZATION OF PRINTED MOLDAVIAN DIALECT RECORDS

Аннотация. Румынские лингвистические ресурсы, накопленные молдавскими филологами до перехода в 1989 году на латинскую графику, представлены на кириллице. Это препятствует широкому доступу к ним. В работе описывается технология дигитизации диалектных записей, напечатанных в молдавской кириллической фонетической транскрипции. Тексты распознаются и конвертируются в латинскую фонетическую транскрипцию. Описана разработанная программа-конвертор.
Ключевые слова. ИТ в лингвистике, дигитизация, молдавские диалекты, транслитерация фонетической записи с кириллицы на латиницу.

Abstract. The Romanian linguistic resources accumulated by Moldavian philologists before the transition to the Latin script in 1989 are presented in Cyrillic. This prevents widespread access to them. This paper describes the technology of digitization of published dialect records printed in Moldavian Cyrillic phonetic transcription. The texts are recognized and converted to the Latin phonetic transcription. The developed converter program is described.
Keywords. IT in linguistics, digitization, Moldovan dialects, transliteration of phonetic recording from Cyrillic to Latin.

1. Введение

Диалекты и говоры всегда привлекали внимание исследователей разнообразием фонетических, грамматических и лексических форм. В 1957–1965 гг. сектор диалектологии Института языка и литературы АН МССР организовал полевые исследования в местах проживания молдаван в 240 населенных пунктах на территории СССР.

Собранный на местах материал был включен в Молдавский лингвистический атлас (1968–1973) [3]. Были изданы в 6 томах «Тексте диалектале» [4]. Это первая работа по молдавским диалектам румынского языка на территории Советского Союза, которая является важным источником для изучения многих фонетических феноменов и диалектической лексики.

С учетом неоспоримой научной ценности опубликованных диалектных текстов Институт румынской филологии им. Богдана Петри-

[1] Проект (2019–2023): «Научное освоение национального языкового наследия в контексте европейской интеграции», № 20.80009.1606.01 в Госреестре научных и инновационных проектов (Молдова).

чейку-Хашдеу (бывший Институт языка и литературы АН МССР) инициировал проект [5], целью которого является возрождение диалектных записей как части национально-культурного наследия, доступного в электронной форме.

Использование новых информационных технологий дает ряд конкретных преимуществ. Автоматизация процесса дигитизации достигается объединением усилий лингвистов и компьютерных специалистов [2]. В частности, необходимо разработать правила транслитерации фонетических кириллических символов в латинскую транскрипцию и программное обеспечение.

Этот подход обеспечивает лучшее качество печатной версии (первое издание было напечатано на ротапринте). В дальнейшем может быть создан электронный корпус диалектных текстов для включения в более крупные корпусы текстов разного типа на румынском языке, что сделало бы их еще более доступными и более удобными. Тома диалектных текстов, опубликованные в советский период на кириллице, обновленные после оцифровки и транслитерации, войдут в серию томов, издаваемых румынскими диалектологами.

Проект является комплексным, средствами ИТ решая филологическую проблему. Данная работа рассматривает разработку и апробацию технологии и программной поддержки переноса диалектных кириллических фонетических записей из бумажной в электронную форму на базе IPA. Предполагается последующее переиздание записей на латинице. В [5] содержится обзор проекта с точки зрения филолога.

В разделе 2 описана применяемая технология оцифровки текстов, записанных молдавской кириллической фонетической транскрипцией.

В разделе 3 описан разрабатываемый нами на Python конвертор для транслитерации распознанного текста диалектов румынского языка из молдавской кириллической фонетической в латинскую транскрипцию на основе IPA.

2. Оцифровка текстов на молдавской фонетической транскрипции

Фонетическая транскрипция для исходного варианта состоит из кириллицы с некоторыми дополнениями [4: 8–11].

Процесс оцифровки текстов состоит из нескольких этапов: сканирование, постобработка отсканированных ресурсов, подготовка к распознаванию, распознавание текста (OCR), автоматическая и ручная проверка [2].

На первом шаге было проведено сканирование с разрешением до 600 dpi. На этапе постобработки полученных файлов использовалось приложение Scan Tailor. Это интерактивный инструмент постобработки отсканированных страниц. Он выполняет такие операции, как разделение страниц, выравнивание, добавление/удаление страниц, преобразование в двоичную форму, очистку и другие функции [6].

На этапе распознавания были разработаны шаблоны со специальным алфавитом, учитывающим символы кириллической фонетической записи, и расширены существующие словари. Использовалась программа ABBYY FineReader v.14 [1]. Мы использовали словарь из ранее правильно распознанных слов. Следует отметить, что это не всегда помогало, так как произношение любого слова у разных носителей диалектов бывает непохожим.

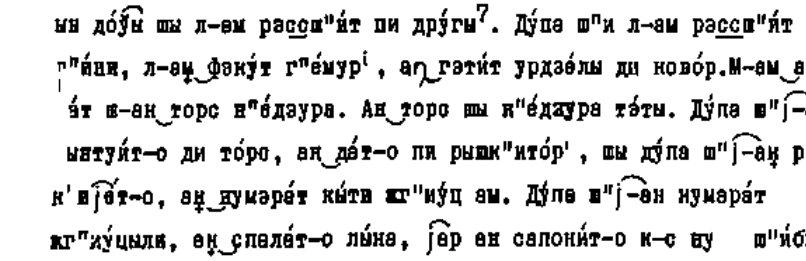


Рис. 1. **Исходное изображение страницы (фрагмент)**

Автоматическая проверка проводилась в режиме редактирования с подсказкой из построенного словаря. Вручную исправлялось лишь около 5–10 % ошибок. На сегодняшний день завершена оцифровка первого тома книги [4].

ын до́ўы шы л-ам рэссш"и́т пи дру́гы[7]. Ду́па ш"и л-ам рэссш"и́т г"и́ни, л-ам̨ фэку́т г"е́мурі, аŋ гэти́т урдза́лы ди ково́р. М-ам апука́т ш-ан то́рс н"е́дзура. Ан торс шы н"е́дзура тэ́ты. Ду́па ш"j-ан мынтуи́т-о ди то́рс, ан да́т-о пи рышк"ито́рі, шы ду́па ш"j-аŋ̨ рышк'ије́т-о, аŋ̨ нумэра́т кы́ти жг"иу́ц ам. Ду́па ш"j-ан нумара́т жг"иу́цыли, аŋ̨ спала́т-о лына, jар ан сапони́т-о к-с ну ш"и́бы

Рис. 2. **Текст после распознавания и правки**

Примеры исходного изображения и распознанного текста приведены на рис. 1, 2.

3. Конвертор молдавской кириллической фонетической транскрипции

После распознавания выполняется транслитерация в заданный фонетический алфавит. Трудность представляет специфика фонетической транскрипции, использующей буквы с диакритическими знаками над и под буквой, а также позиции символа (суперскрипт). Графические связки проставляются вручную.

Кириллический вариант использует около 80 фонетических символов с 16 диакритическими знаками, плюс еще несколько связок и разделителей, в итоге более 100 символов.

Необходимые символы не всегда есть в шрифтах, и каждый символ с диакритикой изображается цепочкой до 4 точек Юникода. Транслитерация в ряде случаев зависит от позиции символа в слове (начало, середина, конец), контекста (соседних символов) и даже от вертикальной позиции символа (нормальный, суперскрипт).

Были разработаны 273 правила транслитерации (табл. 1).

Таблица 1. **Пример правил транслитерации в латиницу**

ý	→	ú	пу́ни → púni
ў	→	u̯	брыў → brîu̯
ᵘoá	→	ᵘo̯á	сᵘoа́рили → sᵘo̯árili
уы	→	u̯â	доуы → dou̯â
ᵘ + гласная	→	ᵘ̯+ гласная	ᵘогра́ды → ᵘográdâ

Используемый формат файлов docx также не является простым и требует специального подхода. Для разработки конвертора был использован язык Python с расширением — библиотекой python-docx. Библиотека предоставляет обертку (wrapper) для внутреннего XML формата, используемого в docx. Это значит, например, что хотя у нас есть последовательности абзацев и их отрезков, но ряд операций над ними не выполняется, так как механизм обертки динамически воспроизводит текстовое содержимое из XML формата.

Например, текст абзаца есть строка, но ее редактирование не отражается в документе; приходится присваивать строку внешней переменной, править последнюю, а потом переприсваивать полученный текст пустому абзацу. Также приходится отслеживать и переносить в конвертированный текст ряд атрибутов исходного документа, при-

чем на всех уровнях, от документа в целом до отрезков абзаца включительно.

В программе транслитерации требовалось учесть ряд особенностей, осложняющих обработку распознанного текста.

Например, программа распознавания, а также человек, правящий текст после распознавания, иногда вставляют вместо кириллических символов одинаковые по начертанию латинские. Программа является таблично управляемой, и это можно учесть в таблицах. Такое решение повлекло бы повтор информации и заметное увеличение объема таблиц, и при отладке таблиц исправления пришлось бы вносить в нескольких местах.

Поэтому до выполнения транслитерации в дополнительном проходе по тексту все символы одинакового начертания заменяются на кириллицу. Латиница считается ошибкой, подлежащей исправлению. Это допустимо, так как текст является фонетической записью с минимальными пояснениями.

Более сложное явление — неоднозначность порождения символа при наличии нескольких диакритических знаков. Например, символ с двумя диакритическими знаками может быть порожден тремя способами (рис. 3).

В стандарте (Юникод) для таких случаев предусмотрена «каноническая декомпозиция», то есть строго определенный порядок добавления диакритик (курсив на рис. 3).

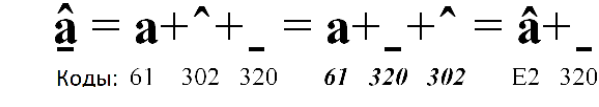


Рис. 3. **Варианты декомпозиции составного символа**

Символы исходного текста подвергаются канонической декомпозиции. В таблицу вносится предписанный стандартом вариант. После подстановки латиницы выполняется обратная операция («каноническая композиция»).

Символ «й» считается составным и превращается в «и» с диакритикой, становясь гласным, что неверно в нашем случае. Поэтому в программе «й» исключен из декомпозиции.

Каноническая декомпозиция также выполняется на отдельном проходе по тексту.

Программа до собственно подстановки латинских эквивалентов взамен кириллицы из таблицы выполняет шесть дополнительных просмотров текста:

- замена символов, которых не может быть в фонетической записи;
- замена символов из эпизодически используемого в кирилличе-ском тексте нестандартного шрифта;
- замена латиницы, каноническая декомпозиция и перевод текста в нижний регистр;
- замена суперскрипта на верхний регистр для простоты обработки;
- контекстная замена сочетаний, содержащих небуквенные символы;
- устранение пробелов перед диакритикой.
- Фрагмент страницы после транслитерации и простановки графических связок показан на рис. 4.

```
în dóu̯â ș̦î l-am răssș̦ît pi drúgâ[7]// dúpa și l-am răssș̦ît
g̑íni/ l-aṃ‿făcút g̑émuri̯/ aŋ‿gătít urd̦álâ di covór// m-am‿apu-
cát ș-an‿tors n̑éd̦ura// an‿tors ș̦î n̑éd̦ura tắtâ// dúpa ŝ-an
mîntuít-o di tórs/ an‿dát-o pi rîș̦k̑itóȓ/ ș̦î dúpa ŝ-an rîș̦-
k̑i̯ét-o/ aṇ‿numărát c̑íti j̑g̑iúț am// dúpa ŝ-an numarát
j̑g̑iúț̂ili/ aṇ‿spalát-o l̂ína/ i̯ar an saponít-o c-s nu șîbâ
```

Рис. 4. **Фрагмент фонетической записи после транслитерации**

4. Заключение

Книга «Тексте диалектале» состоит из 6 томов. Мы распознали первый том, получили набор шаблонов распознавания, пользовательский фонетический алфавит, словарь лексики для автоматизации проверки. Разработали конвертор для представления распознанного текста диалектов румынского языка, записанных кириллической фонетической транскрипцией, на основе выявленных правил транслитерации.

Дальнейшая работа предполагает исследование молдавских диалектов румынского языка посредством дигитизации и транслитерации остальных томов диалектных записей. Оцифрованные и транслитерированные материалы будут размещены на сайте проекта, а также напечатаны по мере возможности.

Литература

1. ABBYY FineReader v.14. URL: https://help.abbyy.com/en-us/finereader/14/ (дата обращения: 01.07.2021).
2. *Cojocaru S., Ciubotaru C., Colesnicov A., Malahov L., Bumbu T.* (2017), Instrumentar pentru digitizarea şi transliterarea textelor tipărite în limba română cu caractere chirilice. În: Revista Bibliotecii Academiei Române. No. 2, pp. 27–38.
3. *Melnic V. et al.* (1968–1973), Atlasul lingvistic moldovenesc. Chişinău.
4. *Melnic V., Stati V., Udler P.* (1969), Texte dialectale, vol. 1. Chişinău.
5. *Popovschi L.* (2020), Despre necesitatea revitalizării textelor dialectale publicate la Chişinău. URL: https://ibn.idsi.md/sites/default/files/imag_file/154-160_7.pdf (дата обращения: 01.07.2021).
6. Scan Tailor official web site. URL: https://scantailor.org/

References

1. ABBYY FineReader v.14. URL: https://help.abbyy.com/en-us/finereader/14/ (date of access: 01.07.2021).
2. *Cojocaru S., Ciubotaru C., Colesnicov A., Malahov L., Bumbu T.* (2017), Instrumentar pentru digitizarea şi transliterarea textelor tipărite în limba română cu caractere chirilice [Instrumentation for digitizing and transliterating texts printed in Romanian with Cyrillic characters]. In: Revista Bibliotecii Academiei Române [Romanian Academy Library Magazine]. No. 2, pp. 27–38.
3. *Melnic V. et al.* (1968–1973), Atlasul lingvistic moldovenesc [Moldavian linguistic atlas]. Chisinau.
4. *Melnic V., Stati V., Udler P.* (1969), Texte dialectale [Dialect texts]. Vol. 1. Chisinau.
5. *Popovschi L.* (2020), Despre necesitatea revitalizării textelor dialectale publicate la Chişinău [On necessity of revitalization of dialect texts published in Chisinau] URL: https://ibn.idsi.md/sites/default/files/imag_file/154-160_7.pdf (date of access: 01.07.2021).
6. Scan Tailor official web site. URL: https://scantailor.org/

Колесников Александр Евгеньевич
Институт математики и информатики им. В. Андрунакиевича (Молдова)
Colesnicov Alexandru
Andrunachievici Institute of Mathematics and Computer Science (Moldova)
E-mail: acolesnicov@gmx.com

Малахов Людмила Андреевна
Институт математики и информатики им. В. Андрунакиевича (Молдова)
Malahov Ludmila
Andrunachievici Institute of Mathematics and Computer Science (Moldova)
E-mail: ludmila.malahov@math.md

М. В. Копотев, О. В. Кисселев, А. А. Климов

M. V. Kopotev, O. V. Kisselev, A. A. Klimov

CAT&KITTENS: КОРПУС РУССКИХ АКАДЕМИЧЕСКИХ ТЕКСТОВ И ОСНОВАННЫЕ НА НЕМ ИНСТРУМЕНТЫ АНАЛИЗА СТУДЕНЧЕСКИХ РАБОТ[1]

CAT&KITTENS: AN ACADEMIC CORPUS AND CORPUS-BASED TOOLS FOR ANALYSIS OF RUSSIAN ACADEMIC WRITING

Аннотация. Проект направлен на создание Корпуса академических текстов на русском языке (CAT) и связанного с ним сервиса исправления ошибок в студенческих текстах (*kittens*). Платформа позволяет искать примеры в большом академическом корпусе и проверять собственный текст на наличие неакадемических вариантов употребления.
Ключевые слова. Корпус, академическое письмо, корпусной анализ.

Abstract. The project aims to create the Corpus of Academic Russian Texts (CAT) a corpus-based service that provides searches for non-standard language usage in novice writers' texts (*kittens*). The platform allows for simple searching in a large academic corpus and provides sophisticated evaluation of a novice writer's text against the large corpus data.
Keywords. Corpus, academic writing, corpus-based analysis.

1. Введение

Изучение академического языка в значительной степени обусловлено необходимостью научить студентов, осваивающих этот жанр, устоявшимся практикам и языковым шаблонам. В течение последних нескольких десятилетий эта область исследований находилась под влиянием двух взаимосвязанных подходов: компьютерного обучения языку (англ. Computer-Assisted Language Learning, CALL) и компьютерной лингвистики, включая подходы и инструменты корпусной лингвистики.

Методы корпусной лингвистики широко используются для создания учебных языковых ресурсов (Crossley et al., 2017). Особенно заметный вклад они внесли в анализ английского академического дискурса (Ackermann & Chen, 2013; Biber et al., 2004; Durrant & Mathews-Aydinli, 2011; Gray & Biber, 2013).

[1] Статья частично основана на нашей публикации (Klimov et al., 2021). Мы благодарим студентов НИУ ВШЭ А. Грилланди, М. Килину, Е. Носову, А. Сидорову и студента Портлендского университета Ивза (Наму) Винеке, без которых эта статья никогда бы не появилась.

Несмотря на то, что компьютерные подходы бурно развивались в последнее десятилетие, практических приложений CALL для изучения русского языка существует немного. Мы можем назвать проекты Revita (Katinskaia et al., 2018), Текстометр (Laposhina et al., 2018), Visualizing Russian (Clancy et al., 2019), Русский конструктикон (Endresen et al., 2020), RuSkell (Apresjan et al., 2016), CoCoCo (Копотев, 2020).

Важно отметить, что существуют корпусные ресурсы, содержащие академические тексты. Например, НКРЯ включает тексты академического жанра (около 27,4 млн токенов). Однако подкорпус не позволяет выделять и сортировать различные поджанры и не сбалансирован с точки зрения представленности академических дисциплин. Наконец, НКРЯ невозможно выгрузить для более сложной языковой обработки. Еще один проект, который дает возможность анализировать академический стиль, — это подкорпус RU-AC в составе проекта IntelliText, возглавляемого С. Шаровым (corpus.leeds.ac.uk/itweb). RU-AC относительно невелик (пять миллионов токенов), состоит из студенческих работ разного качества и не может считаться репрезентативным.

Проект CAT&kittens заполняет эту лакуну в двух направлениях. Мы надеемся, во-первых, внести свой скромный вклад в изучение современного русского академического языка и выявить, хотя бы частично, типичные ошибки, что позволит использовать пропедевтические подходы для их предупреждения у следующих поколений студентов. Во-вторых, мы планируем создать платформу, которая предоставит пользователю возможность анализировать и редактировать собственный текст на основе сравнения с большим корпусом.

2. Структура корпуса

CAT задуман как двухчастный корпус академического русского языка, охватывающий несколько научных областей. Эталонный корпус, собственно CAT, включает 3600 научных статей, взятых из авторитетных рецензируемых российских академических журналов; статьи подкорпуса представляют шесть академических областей: экономика, образование и психология, юридические тексты, лингвистика, история и социология. Все статьи опубликованы в период с 2010 по 2018 год. Размер корпуса CAT — около 14 миллионов токенов, что достаточно для реализации множества задач, однако не всех. Для решения проблем, связанных прежде всего с нейросетевым моделиро-

ванием, этого объема оказалось недостаточно, поэтому был собран дополнительный корпус CyberCAT, основанный на научной онлайн-библиотеке «Киберленинка» (cyberleninka.org). Объем этого корпуса — около 155 миллионов токенов. Принципы создания и аннотирования обоих корпусов совпадают, однако в обучающей платформе будет использоваться только эталонный корпус CAT.

Таблица 1. **Распределение текстов в CAT и CyberCAT**

Корпус / Дисциплина	Эталонный корпус CAT	Корпус для машинного обучения CyberCAT
Экономика	2 494 422	16 548 007
Педагогика и психология	1 880 004	32 773 116
Юриспруденция	2 636 290	17 361 237
Лингвистика	2 691 363	11 985 639
История	2 808 313	55 516 958
Социология	1 500 196	21 127 403
Всего	**14 010 588**	**155 312 360**

3. Обработка текстов

Все тексты в кодировке UTF-8 аннотированы с помощью анализатора UDpipe (Straka & Straková, 2017), который объединяет инструменты токенизации, лемматизации, морфологической разметки и синтаксического анализа в терминах грамматики зависимостей в формате CoNLL-U. Использование этого инструмента позволяет получать не только базовую информацию, например, о частоте токенов или лемм, но и более детальную информацию о грамматических характеристиках токенов и синтаксических зависимостях между ними.

Перед применением UDpipe все тексты были предварительно обработаны: ссылки, таблицы, рисунки, имена авторов, а также названия университетов или журналов и номера страниц в колонтитулах были убраны. Тексты были разделены на отдельные предложения; знаки препинания, обозначающие конец предложения, заменены на точки; остальные знаки препинания (например, запятые, точки с запятой и т. д.) были удалены. Все числа в текстах были заменены токеном NUM. В корпус были включены только основные разделы статей (вве-

дение, обзор литературы, анализ и результаты, обсуждение и заключение и т. п.); все таблицы, рисунки, уравнения и цитаты в тексте, а также библиография были автоматически удалены с помощью скриптов Python, доступных по адресу: github.com/kopotev/CATandkittens.

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# newdoc									
# newpar									
# sent_id = 1									
# text = В работе показан процесс разработки платформы.									
1	В	в	ADP	_	_	2	case	_	_
2	работе	работа	NOUN	_	Animacy=Inan|Case=Loc|Gender=Fem|Number=Sing	3	obl	_	_
3	показан	показать	VERB	VERB	Aspect=Perf|Gender=Masc|Number=Sing|Tense=Past|Variant=Short|VerbForm=Part|Voice=Pass	0	root	_	_
4	процесс	процесс	NOUN	NOUN	Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing	3	nsubj:pass	_	_
5	разработки	разработка	NOUN	_	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	4	nmod	_	_
6	платформы	платформа	NOUN	NOUN	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	5	nmod	_	SpaceAfter=No
7	.	.	PUNCT	_	_	3	punct	_	SpaceAfter=No

Рис. 1. **Пример разметки в формате UDpipe**

4. Инструменты проверки текста

На основе созданных корпусов создается онлайн-сервис CAT&kittens — многофункциональная платформа с различными инструментами поиска и анализа. Помимо стандартных функций поиска по корпусу платформа предоставляет возможности для оценки качества текста.

Первый уровень проверки — это общая оценка текста. С помощью визуализации различных параметров текста система дает автору представление об уровне сложности текста (*readability*), для которой используются тесты *Flesch Reading Ease* и *Flesch-Kincaid Grade*, адаптированные для русских академических текстов (Solovyev et al., 2018) и согласованные с общеевропейской шкалой языковой компетенции CEFR (Little, 2007). Кроме этого, сервис рассчитывает среднюю длину предложений, которые сравниваются с соответствующими значениями в CAT с целью продемонстрировать, насколько анализируемый текст соотносится с общим уровнем сложности, принятым в академическом дискурсе. Слишком длинные или слишком короткие предложения выделяются в тексте для дальнейшей правки. На этом этапе мы проверяем также возможные повторы фрагментов в тексте.

Второй уровень проверки касается лексической вариативности и коллокаций. Эта часть включает прежде всего показатель sTTR (соотношение тип/токен), дающий общее представление о лексическом богатстве текста. На более детальном уровне система анализирует низкочастотные слова, в том числе *hapax legomen*, проверяет чрезмерное или недостаточное использование лексем из выбранной предметной области. Помимо простого определения отклонений от академического стандарта, система предлагает альтернативные варианты, подобранные на основе семантической близости слов, полученных на предобученной на данных CyberCAT модели.

Семантическая близость используется и в алгоритме поиска нестандартных коллокаций, который на первом этапе выявляет в тексте n-граммы, не найденные в большом корпусе, затем подбирает семантически близкие замены и на последнем этапе проверяет полученные коллокации на наличие в CAT, то есть в эталонном академическом дискурсе.

Третий уровень — проверка грамматики. В отличие от доступных средств проверки, например, Орфо, проверка в нашей системе ориентирована на обнаружение отклонений от стандартов академического письма. Эта часть алгоритма сравнивает пару «токен — тегсет» в тексте с аналогичными парами в корпусе и указывает на возможные отклонения от стандартного употребления морфологических форм. Таковым оказываются, например, императивы, нехарактерные для академических текстов, или неграмматичные формы, которым не находится соответствия в корпусе CAT. В этой же части анализируются более сложные морфосинтаксические отклонения от языкового стандарта, например, генитивные цепочки (набор вложенных друг в друга генитивных групп) или пропуск прономинального субъекта (*prodrop*), характерный скорее для разговорных регистров.

5. Заключение

В нашем докладе мы кратко описали создаваемую платформу CAT&kittens, основу которой составляет большой корпус академических текстов. На данном этапе платформа еще не завершена, однако уроки, извлеченные из ее разработки, имеют некоторое значение для методологии использования корпусных инструментов анализа. Один из уроков — сочетание автоматической проверки и человеческой компетенции при создании инструментов анализа. Цель создания подоб-

ной системы состоит не в замене квалифицированного эксперта, не в полной автоматизации процесса, а в том, чтобы облегчить поиски подозрительных фрагментов и предложить автору возможные варианты. Окончательный выбор можно сделать либо обратившись к эксперту, либо, в соответствии с корпусной парадигмой, с помощью поиска в CAT — глубоко аннотированном репрезентативном корпусе русских академических текстов, который предоставляет как учащимся, так и исследователям, изучающим академические жанры, необходимые инструменты анализа.

Литература

1. Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) — A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247.
2. Apresjan, V., Baisa, V., Buivolova, O., Kultepina, O., & Maloletnjaja, A. (2016). RuSkELL: Online Language Learning Tool for Russian. *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity (6–10 September, 2016)*. EURALEX International Congress. Lexicography and Linguistic Diversity, Tbilisi.
3. Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.
4. Clancy, S., Green, D., Egorova, V., & Willis, O. (2019). Foundations of Russian: A cognitive and constructional approach to teaching Russian enriched by frequency data. *SCLC-2019. Book of Abstracts*, 15.
5. Crossley, S., Russell, D., Kyle, K., & Römer, U. (2017). Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics*, *1*.
6. Durrant, P., & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, *30*(1), 58–72.
7. Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, *18*(1), 109–136.
8. Katinskaia, A., Nouri, J., & Yangarber, R. (2018). Revita: A language-learning platform at the intersection of its and call. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
9. Klimov, A., Kisselev, O., & Kopotev, M. (2021). Towards intelligent correction of collocational errors in Russian novice academic texts in the CAT&kittens writing support platform. *Russian Language Journal.*, *71*(1), (в печати).
10. Laposhina, N., Veselovskaya, V., Lebedeva, M.U., & Kupreshchenko, O.F. (2018). Automated text readability assessment for Russian second language learners. *Компьютерная лингвистика и интеллектуальные технологии*, 403–413.
11. Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *91*(4), 645–655.

12. Solovyev, V., Ivanov, V., & Solnyshkina, M. (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent & Fuzzy Systems*, *34*(5), 3049–3058.
13. Straka, M., & Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.
14. Копотев, М. В. (2020). О самом сложном: изучение сочетаемости слов онлайн. *Русский язык за рубежом*, *6*, 36–43.
15. Эндерсен, А., Жукова, В., Мордашова, Д., Рахилина, Е., Ляшевская, О. (2020). Русский конструктикон: Новый лингвистический ресурс, его устройство и специфика. *Компьютерная лингвистика и интеллектуальные технологии*. Вып. 19(26). М.: Изд-во РГГУ, 1–15.

References

1. Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247.
2. Apresjan, V., Baisa, V., Buivolova, O., Kultepina, O., & Maloletnjaja, A. (2016). RuSkELL: Online Language Learning Tool for Russian. *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity (6–10 September, 2016)*. EURALEX International Congress. Lexicography and Linguistic Diversity, Tbilisi.
3. Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.
4. Clancy, S., Green, D., Egorova, V., & Willis, O. (2019). Foundations of Russian: A cognitive and constructional approach to teaching Russian enriched by frequency data. *SCLC-2019. Book of Abstracts*, 15.
5. Crossley, S., Russell, D., Kyle, K., & Römer, U. (2017). Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics*, *1*.
6. Durrant, P., & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, *30*(1), 58–72.
7. Endersen, A., Zhukova, V., Mordashova, D., Rakhilina, E., Lyashevskaya, O. (2020). Russkij konstruktikon: Novyj lingvisticheskij resurs, ego ustrojstvo i specifika. *Komp'yuternaya lingvistika i intellektual'nye texnologii*. 19(26). Moscow: RGGU, 1–15.
8. Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, *18*(1), 109–136.
9. Katinskaia, A., Nouri, J., & Yangarber, R. (2018). Revita: A language-learning platform at the intersection of its and call. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
10. Klimov, A., Kisselev, O., & Kopotev, M. (2021). Towards intelligent correction of collocational errors in Russian novice academic texts in the CAT&kittens writing support platform. *Russian Language Journal.*, *71*(1), (In press).

11. Kopotev, M. B. (2020). O samom slozhnom: izuchenie sochetaemosti slov onlajn. *Russkij yazyk za rubezhom*, *6*, 36–43.
12. Laposhina, N., Veselovskaya, V., Lebedeva, M. U., & Kupreshchenko, O. F. (2018). Automated text readability assessment for Russian second language learners. *Компьютерная Лингвистика и Интеллектуальные Технологии.*, Moscow, 403–413.
13. Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *91*(4), 645–655.
14. 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), 3049–3058.
15. Straka, M., & Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Копотев Михаил Вячеславович
НИУ ВШЭ (Россия)
Хельсинкский университет (Финляндия)
Kopotev Mihail
HSE University (Russia)
University of Helsinki (Finland)
E-mail: mkopotev@hse.ru

Кисселев Олеся Викторовна
Университет Техаса в Сан-Антонио (США)
Kisselev Olesya
The University of Texas at San Antonio (USA)
E-mail: olesya.kisselev@utsa.edu

Климов Александр Антонович
НИУ ВШЭ (Россия)
Хельсинкский университет (Финляндия)
Klimov Aleksandr
HSE University (Россия)
University of Helsinki (Finland)
E-mail: a.klimov@hse.ru

Н. А. Коротаев

N. A. Korotaev

ПОИСК В МУЛЬТИКАНАЛЬНОМ КОРПУСЕ: СОДЕРЖАТЕЛЬНЫЕ ЗАДАЧИ И ТЕХНИЧЕСКАЯ РЕАЛИЗАЦИЯ

SEARCHING IN A MULTICHANNEL CORPUS: RESEARCH QUESTIONS AND TECHNICAL SOLUTIONS

Аннотация. В докладе представлены основные возможности поисковой системы по корпусу «Рассказы и разговоры о грушах», доступной на сайте https://multidiscourse.ru/search/. Корпус состоит из аудио- и видеозаписей однотипных коммуникативных сессий, снабженных разметкой вокального и кинетического поведения. Обсуждаются конкретные исследовательские вопросы, решение которых может быть облегчено благодаря наличию разрабатываемой поисковой системы.
Ключевые слова. Мультиканальный дискурс, корпусы устной речи, разработка поисковых движков, диалог и монолог.

Abstract. The paper presents the search system developed for the "Russian Pear Chats & Stories" (RUPEX) corpus and discusses its main features. RUPEX is a multichannel corpus that includes audio and video recorded communication sessions and their annotations. The search system, which is based on the integrated multichannel annotation, is a one-page browser application that allows users to run queries and view results in a multi-layered format together with relevant video fragments. Several phenomena intrinsic to natural communication can be extensively studied using this research tool.
Keywords. Multichannel discourse, spoken corpora, search engines, dialogue and monologue.

1. Вводные замечания

Среди насущных задач современной корпусной лингвистики — создание и использование ресурсов, которые позволяют изучать язык в ситуациях естественной коммуникации; см., в частности [Kotov, Budyanskaya 2012; Богданова-Бегларян (ред.) 2013; Joo et al. 2017]. Базовой формой такого использования языка является устная речь, при этом только речью естественная коммуникация не ограничивается: общаясь, мы также используем жесты, мимику, движения глаз, расположение тела в пространстве и проч. Подход к анализу языковой деятельности, в котором наряду с речью учитывается и вклад других коммуникативных ресурсов, принято называть мультимодальным или мультиканальным [Müller et al. (eds.) 2013]. Создание мультиканальных корпусов связано с несколькими проблемами. Приведем две из них, наиболее существенные в контексте данного доклада.

Во-первых, крайне трудоемкой является содержательная разметка мультиканального дискурса: здесь (почти) нет успешных автоматизированных решений, поэтому приходится во многом полагаться на ручное аннотирование. Во-вторых, нетривиален вопрос о способах извлечения и представления размеченных данных, особенно в веб-интерфейсе. В этом докладе мы кратко рассмотрим эти вопросы на примере корпуса «Рассказы и разговоры о грушах». Дальнейшее изложение построено следующим образом. В разделе 2 приводятся основные данные о корпусе и обсуждаются принципы разметки. В разделе 3 описывается веб-интерфейс поиска по корпусу. В разделе 4 перечисляются некоторые исследовательские вопросы, решение которых, на наш взгляд, существенно облегчается благодаря наличию основанной на размеченных данных поисковой системы.

2. «Рассказы и разговоры о грушах»: общие принципы разметки

Корпус «Рассказы и разговоры о грушах» (RUPEX; https://multidiscourse.ru/) был записан в 2015 и 2017 годах. Всего было получено 40 записей, имеющих единую структуру: участники всех записей в заранее определенном порядке рассказывают и обсуждают содержание «Фильма о грушах», созданного в 1970-е годы группой под руководством У. Чейфа [Chafe (ed.) 1980]. Поведение участников фиксировалось посредством петличных микрофонов, фронтальных видеокамер и видеокамеры общего плана, у двух участников в каждой записи при помощи портативных айтрекеров также регистрировалась глазодвигательная активность. Общий объем аудио- и видеоматериала составляет около 15 часов; подробнее о корпусе см. [Kibrik, Fedorova 2018].

Разметка записей корпуса выполняется независимо для отдельных коммуникативных каналов: вокального (вербальный и просодический компоненты), канала мануальной жестикуляции, канала глазодвигательной активности и проч. Ниже приведен фрагмент текстовой аннотации (транскрипта) вокального канала одной из записей. Транскрипт выполнен в соответствии с принципами, изложенными в [Kibrik et al. 2020]. Нумерованные строки транскрипта соответствуют элементарным дискурсивным единицам (ЭДЕ) и паузам между ними. Под ЭДЕ понимается минимальный шаг в развитии устного дискурса, соотносимый с единым когнитивным усилием и выделяемый в потоке речи на основании набора просодических характеристик. В скобках указана длительность абсолютных и заполненных пауз; при помощи слэшей и стрелок отображаются движения частоты основного тона

в акцентированных словоформах; пунктуационные знаки в конце строк указывают на иллокутивно-фазовое значение ЭДЕ: дефолтная незавершенность (запятая), незавершенность в контексте неполноты информации (три запятых), завершение сообщения (точка) и др.

(1) Фрагмент вокального транскрипта записи pears04

R-vE409	и /тут мимо проезжает \↑девочка,
pR-305	(0.27)
R-vE410	тоже на /→велосипеде,,,
pR-306	(0.06)
R-vE411	–ну уже на \таком-м || (0.17) \женском /→велосипедике,,,
pR-307	(0.60)
R-vE412	(э 0.48) она с \↑косичками,
R-vE413	с \двумя \↑косичками,
pR-308	(0.27)
R-vN065	{sm 0.05}
R-vE414	(\–во-от,)
R-vE415	и /проезжает \мимо.

Разумеется, текстовый формат аннотации пригоден только для вокального канала. Прочие каналы размечаются в многоуровневом формате, к которому — для получения единой мультиканальной аннотации — в дальнейшем приводятся и речевые данные. Для хранения единой разметки используется программа ELAN [Hellwig et al. 2018], подробнее о принципах мультиканальной аннотации в корпусе см. [Korotaev et al. to appear].

3. Интерфейс онлайн-поиска[1]

При наличии ELAN, разметок и медиафайлов поиск можно выполнять на локальной машине непосредственно по единой мультика-

[1] Серверная часть поисковой системы разработана Г. Б. Добровым, клиентская — А. Н. Хитровым.

нальной аннотации. Однако в этом решении есть ряд очевидных ограничений, обойти которые позволяет браузерная реализация. При ее разработке мы исходили из следующих положений. Во-первых, поиск должен работать в браузере, без необходимости устанавливать дополнительное программное обеспечение и скачивать «тяжелые» медиафайлы. Во-вторых, он должен обладать достаточно понятным интерфейсом — доступным в том числе и для тех пользователей, которые не изучили сложную схему мультиканальной аннотации. В-третьих, при отображении результатов поиска необходимо отображать видеофайлы и ассоциированную с ними содержательную разметку — причем именно те ее аспекты, которые релевантны для конкретного запроса и/или конкретного пользователя.

Разрабатываемая поисковая система доступна по адресу https://multidiscourse.ru/search/. В ней используется SQL-реплика мультиканальных аннотаций формата ELAN; клиентская часть представляет собой одностраничное приложение на языке JavaScript, спроектированное с использованием шаблона MVVM; серверная часть выполнена на языке Java. На апрель 2021 года поиск производится по трем размеченным записям корпуса; пользователям доступны следующие функции (краткое описание основных возможностей также содержится во вкладке «Как искать»):

- Ограничение области поиска отдельными записями и/или этапами записей (вкладка «Область поиска»).
- Составление поисковых запросов: как простых, т. е. содержащих в себе одну поисковую единицу, так и сложных (вкладка «Запрос»). В простых запросах пользователю достаточно выбрать тип искомой единицы и при необходимости указать ее свойства, выбрав требуемые значения в именованных полях: общих для единиц всех типов (длительность, принадлежность участнику с конкретной ролью в записи) или специфических для конкретного типа (наличие и тип акцента для слов; рукость для мануальных жестов; направления взгляда для глазных фиксаций и др.). Каждое поле снабжено краткой справкой. Сложные запросы включают в себя несколько поисковых единиц, связанных между собой ограничениями на расстояние и (опционально) на (не) совпадение участников. При формировании сложного запроса происходит динамическое перестроение «карты запроса».
- Просмотр результатов поиска (вкладка «Результаты»). По умолчанию результаты представлены в виде простого текстового

списка; при нажатии на элемент выдачи отображается фрагмент многоуровневой разметки и плеер с выбором из нескольких синхронизированных видеофайлов.

- Настройка слоев для отображения в расширенном режиме просмотра результатов (вкладка «Настройки выдачи»).

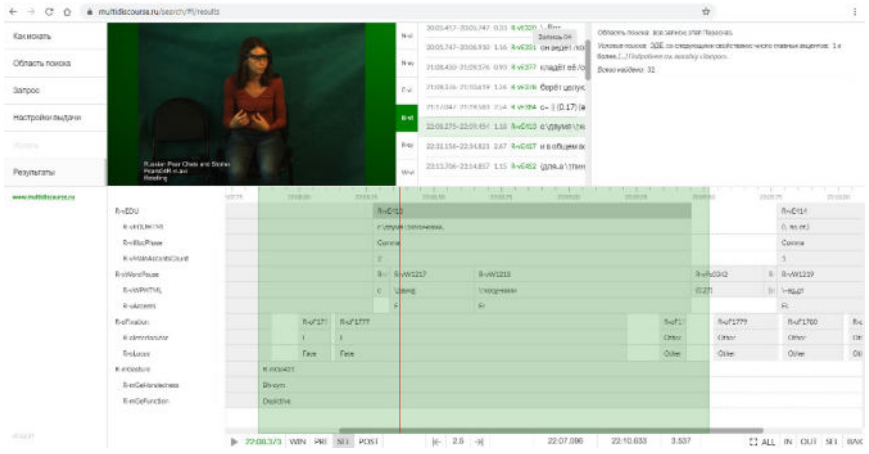


Рис. 1. **Расширенный режим просмотра результатов по поисковому запросу на сайте https://multidiscourse.ru/search/**

На рис. 1 показан расширенный режим просмотра результатов по сложному поисковому запросу, включающему в себя три единицы:

(i) ЭДЕ участника X, содержащие в себе не менее одного акцентированного слова;
(ii) Фиксации взгляда того же участника, направленные на лицо кого-либо из других участников записи и пересекающиеся с ЭДЕ (i);
(iii) Изобразительные жесты того же участника, пересекающиеся с ЭДЕ (i).

Область поиска ограничена этапом пересказа. Для просмотра выбран фрагмент, соответствующей ЭДЕ R-vE413 из транскрипта (1). Видно, что говорящая смотрит на лицо слушателя и сопровождает произнесение ЭДЕ с *двумя косичками* изобразительным жестом, выполняемым двумя руками с симметричной траекторией.

4. Примеры использования и дальнейшие перспективы

Приведем некоторые примеры исследовательских задач, решение которых становится возможным — или существенно облегчается — благодаря наличию мультиканальной аннотации и возможности поиска по аннотированным данным. (Список далеко не полный и носит сугубо иллюстративный характер.)

1) Составление «жестикуляционного портрета» лексических единиц. Например, если задать в поиске слова, которые указывают на регулярно встречающихся в корпусе референтов, и включить в настройках выдачи отображение мануальных жестов, можно получить список изобразительных кинетических техник, которые участники записей используют для сопровождения речевого описания этих объектов. Аналогичный вопрос можно поставить и для единиц, не обладающих предметным значением, — например, дискурсивных маркеров *вот и ну*. Так, согласно данным размеченного подкорпуса, из 372 вхождений самых частотных в корпусе существительных с предметным значением (*груша, мальчик, корзина, шляпа, велосипед*), произносимых с акцентом, 267 (72 %) пересекаются с тем или иным мануальным жестом, реализуемым тем же участником. (В качестве критерия пересечения было задано составное условие вида «начало жеста — от –3000 до 200 мс от начала слова, конец жеста — от 200 до 5000 мс от начала слова».) В то же время для полноударных *вот и ну* аналогичное пересечение с жестами наблюдается лишь в 49 % случаев (100 из 206). При этом из 100 жестов, сопровождающих *вот и ну*, 62 выполняют прагматическую функцию, т. е. преимущественно нацелены на установление контакта между участниками коммуникации; тогда как из 267 жестов, сопровождающих частотные слова с предметным значением, прагматическую функцию реализуют только 100 (37 %). Таким образом, предварительный анализ подобного рода показывает, что различия в лексической семантике слов в определенной степени поддерживаются сопутствующей мануальной жестикуляцией.

2) Интегральное описание мультиканального поведения участников в заданной коммуникативной ситуации. Примером тут могут служить диалогические обмены, инициируемые вопросом или запросом на подтверждение («полуутверждением»). Указав соответствующие значения в поле «Иллокутивно-фазовое значение ЭДЕ», можно получить предварительный список релевантных контекстов. Впоследствии поисковый запрос можно усложнять, вводя условия, относящиеся

к глазодвигательной активности участников (так, инициатор обмена регулярно поддерживает речевой запрос установлением зрительного контакта с тем участником, от которого он ожидает реакции), структуре коммуникативного обмена (инициатор обычно ожидает реакции на свой запрос, а получив ее от назначенного участника, осуществляет завершающую реплику подтверждения, которая при этом может сопровождаться или заменяться кивком головой), жестикуляции и проч. Как указано в разделе 3, все полученные результаты поиска можно просматривать непосредственно в браузере, выбирая нужную камеру и при необходимости редактируя набор отображаемых слоев разметки.

3) Сопоставительный анализ поведения участников на монологических vs. диалогических этапах записей. Поскольку одни и те же участники задействованы как при (пере)рассказах содержания фильма, так и при его обсуждении, материал корпуса располагает к подобного рода контрастивным исследованиям. Так, манипулируя значениями в поле «Этапы записей» на вкладке «Область поиска», можно, в частности, обнаружить, что у одних и тех же участников в диалоге по сравнению с монологом значительно падает частотность заполненных, а также внутренних (т. е. располагающихся внутри ЭДЕ) пауз. Это наблюдение подтверждает сложившееся в литературе представление о большей сложности в порождении монологического дискурса и позволяет дать ему количественную оценку.

Мы полагаем, что при увеличении объема размеченных данных разрабатываемая система поиска позволит получать новые результаты об устройстве устной и — шире — мультиканальной коммуникации. Кроме того, мы надеемся, что предложенные технические решения могут заинтересовать разработчиков других мультиканальных корпусов и поспособствовать созданию единого формата мультиканальной разметки.

Литература

1. *Богданова-Бегларян Н. В.* (ред). (2013), Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 1. Чтение. Пересказ. Описание. СПб, Филологический факультет СПбГУ.
2. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Norwood, NJ, Ablex.
3. *Joo J., Steen F., Turner M.* (2017), Red Hen Lab: Dataset and Tools for Multimodal Human Communication Research. In: KI — Künstliche Intelligenz. No. 31, pp. 357–361.

4. *Hellwig B., Hulsbosch M., Somasundaram A., Tacchetti M., Geerts J.* (2018), ELAN — Linguistic Annotator: version 5.4. Manual updated on 2018-12-05. URL: https://www.mpi.nl/corpus/manuals/manual-elan.pdf (дата обращения: 01.04.2021).
5. *Kibrik A. A., Fedorova O. V.* (2018), Language production and comprehension in face-to-face multichannel communication. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue". Vol. 24(17), pp. 305–316.
6. *Kibrik A. A., Korotaev N. A., Podlesskaya V. I.* (2020), Russian spoken discourse: Local structure and prosody. S. Izre'el et al. (eds.). In search of basic units of spoken language: A corpus-driven approach. Amsterdam, John Benjamins, pp. 37–76.
7. *Korotaev N. A., Evdokimova A. A., Litvinenko A. O., Nikolaeva J. V., Sukhova N. V., Fedorova O. V., Kibrik A. A.* (to appear), Multichannel annotation scheme. O. V. Fedorova, A. A. Kibrik (eds.). The MCD handbook: A practical guide to annotating multichannel discourse. Institute of Linguistics RAS.
8. *Kotov A. A., Budyanskaya E. M.* (2012), The Russian emotional corpus: Communication in natural emotional situations. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue". Vol. 11(18), pp. 296–306.
9. *Müller C., Cienki A., Fricke E., Ladewig S., McNeill D., Tessendorf S.* (eds.) (2013), Body — Language — Communication: An international handbook on multimodality in human interaction. Vol. 1. Berlin, De Gruyter Mouton.

References

1. *Bogdanova-Beglarian N. V.* (ed.) (2013), Zvukovoj korpus kak material dlja analiza russkoj rechi: Kollektivnaja monografija. Chast' 1. Chtenie. Pereskaz. Opisanie [Speech corpus as a base for analysis. Collective monograph. Part 1. Reading. Retelling. Description]. Saint Petersburg, St. Petersburg University, Faculty of Philology.
2. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Norwood, NJ, Ablex.
3. *Joo J., Steen F., Turner M.* (2017), Red Hen Lab: Dataset and Tools for Multimodal Human Communication Research. In: KI — Künstliche Intelligenz. No. 31, pp. 357–361.
4. *Hellwig B., Hulsbosch M., Somasundaram A., Tacchetti M., Geerts J.* (2018), ELAN — Linguistic Annotator: version 5.4. Manual updated on 2018-12-05. URL: https://www.mpi.nl/corpus/manuals/manual-elan.pdf (date of access 01.04.2021).
5. *Kibrik A. A., Fedorova O. V.* (2018), Language production and comprehension in face-to-face multichannel communication. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue". Vol. 24(17), pp. 305–316.
6. *Kibrik A. A., Korotaev N. A., Podlesskaya V. I.* (2020), Russian spoken discourse: Local structure and prosody. S. Izre'el et al. (eds.). In search of basic units of spoken language: A corpus-driven approach. Amsterdam, John Benjamins, pp. 37–76.
7. *Korotaev N. A., Evdokimova A. A., Litvinenko A. O., Nikolaeva J. V., Sukhova N. V., Fedorova O. V., Kibrik A. A.* (to appear), O. V. Fedorova, A. A. Kibrik (eds.). The MCD

handbook: A practical guide to annotating multichannel discourse, Institute of Linguistics RAS.
8. *Kotov A. A., Budyanskaya E. M.* (2012), The Russian emotional corpus: Communication in natural emotional situations. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue". Vol. 11(18), pp. 296–306.
9. *Müller C., Cienki A., Fricke E., Ladewig S., McNeill D., Tessendorf S.* (eds.) (2013), Body — Language — Communication: An international handbook on multimodality in human interaction. Vol. 1. Berlon, De Gruyter Mouton.

Коротаев Николай Алексеевич
Российский государственный гуманитарный университет (Россия)
Korotaev Nikolay
Russian State University for the Humanities (Russia)
E-mail: n_korotaev@hotmail.com

И.Л. Корецкая

I.L. Koretskaya

ПРОБЛЕМЫ ЯПОНСКОЙ КОРПУСНОЙ ЛИНГВИСТИКИ

CHALLENGES IN THE JAPANESE CORPUS LINGUISTICS

Аннотация. В ходе работы над созданием корпуса текстов его разработчики сталкиваются со многими сложностями, некоторые из которых являются универсальными (обеспечение репрезентативности корпуса, сбор материала и т.д.), а некоторые обусловлены спецификой языка. В данной работе представлен краткий обзор современного состояния японской корпусной лингвистики, а также анализ различных вопросов, решением которых занимаются разработчики японских корпусов.
Ключевые слова. Японская корпусная лингвистика, лингвистический корпус, проблема слова в японском языкознании, автоматическая обработка японского языка.

Abstract. When building a text corpus, the developers face various problems, some of which are universal (e.g. ensuring the corpus representativeness, collecting material, etc.) and some are caused by some specific features of the language. In case with the Japanese language, most problems are related to the Japanese writing system. The research on the Japanese corpus linguistics is not very well known outside Japan, as well as quite rare. In this paper, we give a brief overview of the current Japanese corpus linguistics, its development and the main Japanese corpora (both written texts and speech corpora) and analyze different issues the Japanese corpora developers deal with.
Keywords. Japanese corpus linguistics, linguistic corpus, word in the Japanese linguistics, Japanese language processing.

1. Введение

Современные исследования, проводимые в рамках японской корпусной лингвистики, в основном посвящены изучению различных языковых явлений на материале корпусов, а также проблемам создания новых и усовершенствования имеющихся программ по автоматической обработке языка, особенно с использованием нейронных сетей. Примеры таких работ можно найти, например, в сборниках материалов Семинара по языковым ресурсам[1], на сайте Университета Киото[2] и в др. источниках. Работы по японским корпусам устной речи собраны, к примеру, в [Proceedings of LREC 2018 Special Speech Sessions]. История развития корпусной лингвистики в Японии подробно описана в статьях сборника Japanese Linguistics, в том числе [Miyajima 2007].

[1] Сборники за 2016–2020 гг. можно найти в архиве NINJAL: https://repository.ninjal.ac.jp/

[2] https://nlp.ist.i.kyoto-u.ac.jp/EN/?Publications

Многие японские исследователи отмечают отставание развития японской корпусной лингвистики от других стран, указывая на то, что долгое время не было объемных репрезентативных национальных корпусов японского языка, создание которых является одной из ее главных задач (см., например, [Gotoo 2007; National Institute for Japanese Language and Linguistics: Survey and guide 2020/2021]). Критика в адрес японской корпусной лингвистики, видимо, вызвана сравнением ее с европейской корпусной лингвистикой. При разработке корпусов японцы ориентировались на идеалы Брауновского корпуса, однако специфика японского языка ставит перед разработчиками корпусов свои задачи.

Исследования по японской корпусной лингвистике вне Японии практически не известны и достаточно редки. И цель данной работы — представить краткий обзор современного состояния японской корпусной лингвистики, а также анализ различных вопросов, решением которых занимаются разработчики японских корпусов.

2. Развитие японской корпусной лингвистики

Можно сказать, что зарождение корпусной лингвистики в Японии произошло в 1950-х гг. В то время методология сбора и исследования языкового материала была схожа с методологией современной корпусной лингвистики, хотя сам термин «корпус» еще не использовался. В конце 1960-х — в 1970-х гг. для исследований японского языка начинают использовать ЭВМ, собранные материалы оцифровывают. В 1990-е гг. вместе с ростом популярности персональных компьютеров появляется все больше электронных версий журналов и газет, что облегчает создание электронных баз по их материалам и дает возможность использовать их в отдельных исследованиях по японскому языку, в научный оборот входит термин «корпус». Однако все еще нет сбалансированного корпуса и признанных методов исследования.

По мнению японских исследователей, новый этап развития корпусной лингвистики в Японии начался с создания Центром разработки корпусов Национального института японского языка и лингвистики (国立国語研究所, NINJAL)[3] объемного Сбалансированного корпуса современного японского письменного языка (現代日本語書き言葉均衡コーパス, BCCWJ).

[3] http://pj.ninjal.ac.jp/corpus_center/

Кроме NINJAL, стоит отметить и Университет Киото[4], работающий над созданием различных корпусов, среди которых текстовый корпус (京都大学テキストコーパス) и веб-корпус (京都大学ウェブ文書リードコーパス), а также над созданием инструментов для автоматической обработки японского языка (морфологического анализатора Juman и его версий, парсера KNP и др.). Международному институту передовых телекоммуникационных исследований ATR[5] принадлежит внушительный список баз данных устной речи, используемых во многих исследованиях: базы спонтанной речи, база речи пожилых людей, база речи детей и др. Национальным институтом информационно-коммуникационных технологий NICT[6] разработаны различные базы данных, включая учебный корпус JLE, содержащий более 1 тыс. записей интервью с носителями японского языка, сдававших устный экзамен на знание английского языка OPI, а также параллельные корпусы (англо-японский, корейско-японский, японско-китайский и др.).

Некоторые организации взяли на себя сбор и распространение корпусов, созданных различными разработчиками. Например, на сайте Национального института информатики (国立情報学研究所)[7] представлено более 40 корпусов устных текстов, на сайте Общества языковых ресурсов GSK (言語資源協会)[8] имеется 11 корпусов письменных и устных текстов. Такие организации помогают получить доступ к необходимому для исследований материалу, ведь не все корпусы широко известны.

3. Обзор корпусов японских текстов

В Японии создается большое количество самых разных текстовых корпусов, и их описание заслуживает отдельной статьи. Здесь мы приведем только несколько наиболее важных.

3.1. Корпусы письменных текстов

Выше уже упоминался корпус BCCWJ, чей объем составляет около 100 млн единиц[9]. Корпус состоит из 3 подкорпусов: подкорпуса публи-

[4] https://nlp.ist.i.kyoto-u.ac.jp/EN/

[5] http://www.atr-p.com/products/sdb.html

[6] https://www.nict.go.jp/

[7] http://research.nii.ac.jp/src/en/list.html

[8] https://www.gsk.or.jp/en/catalog/

[9] В качестве единиц NINJAL использует SUW, о которых будет сказано ниже.

каций, библиотечного подкорпуса и подкорпуса специализированных текстов.

Объемным является и Текстовый корпус университета Киото (京都大学テキストコーパス), который состоит из 40 тыс. предложений, взятых из газеты «Майнити Синбун» за 1995 г.

Из веб-корпусов стоит отметить разработанный NINJAL Веб-корпус японского языка (国語研日本語ウェブコーパス) и Веб-корпус университета Киото (京都大学ウェブ文書リードコーパス). Первый содержит 10 млрд единиц, для его создания разработчики собирали по 100 млн веб-страниц в период с октября по декабрь 2014 г. Для создания второго корпуса выбирались по три первых предложения из 5 тыс. веб-документов (общим объемом 15 тыс. предложений) различных жанров и стилей, включая новостные и энциклопедические статьи, блоги и др.

На основе Корпуса истории японского языка (日本語歴史コーパス, CHJ), который пока находится в разработке, планируется создать диахронический корпус, который будет включать в том числе и тексты на старояпонском.

Среди корпусов детской речи выделяется выпущенный Университетом Конан Корпус Kodomo (こどもコーパス), содержащий сочинения детей в возрасте 10–11 лет. В нем можно проследить историю корректуры текстов.

В качестве примера письменных учебных корпусов приведем Корпус JEFLL (The Japanese English as a Foreign Language Learner Corpus), содержащий более 10 тыс. сочинений (свыше 600 тыс. с/у) носителей японского языка, изучающих английский язык. Была создана его версия, где написанные сочинения проверены носителями английского.

3.2. Корпусы устных текстов

NINJAL разработал и Корпус спонтанной речи (日本語話し言葉コーパス, CSJ), объем которого составил порядка 7 млн единиц (660 часов записи). Основную часть корпуса составляют «академическая речь» (запись живых академических презентаций) и «симулированная публичная речь» (студийные записи речи на повседневные темы перед небольшой аудиторией и в относительно непринужденной обстановке).

Среди корпусов устных текстов нельзя не отметить упомянутые выше базы данных устной речи ATR разного содержания и разработанный NICT Корпус JLE объемом более 1 тыс. записей.

Первым японским корпусом, содержащим записи диалектной речи многочисленных регионов Японии, считается разработанный NINJAL Корпус японских диалектов (日本語諸方言コーパス, Corpus of Japanese Dialect). Он содержит 4 тыс. часов записи диалектной речи из более чем 200 мест в 47 префектурах по всей Японии. Сбор материала осуществляло Агентство по культурным вопросам с 1977 по 1985 год.

4. Сложности при создании корпуса японских текстов

Одной из главных сложностей является обеспечение репрезентативности корпуса, что в полном объеме зачастую невыполнимо. Для получения максимально репрезентативного корпуса можно, например, уравнивать количество текстов по важным для изучаемого объекта параметрам, следя за тем, чтобы корпус отражал все значимые характеристики объекта. Подобный принцип был использован при создании Базы данных устного повседневного общения (談話資料 日常生活のことば) и Корпуса повседневного общения (日常会話コーパス, CEJC), когда при выборе информантов учитывали их пол и возраст, контролируя таким образом выборку. Разработчики Корпуса спонтанной речи (日本語話し言葉コーパス, CSJ) включили в корпус тексты разных типов, стараясь таким образом сбалансировать материал. При создании корпуса BCCWJ для обеспечения репрезентативности применялась случайная выборка текстов.

Большинство сложностей связаны с автоматической обработкой языка. Несмотря на значительные успехи, программы допускают ошибки при анализе текстов. В связи с этим исследователям приходится идти на определенные жертвы. Можно получить либо объемный корпус, но не очень тщательно размеченный, либо хорошо размеченный, но небольшой по объему. NINJAL использует следующий принцип. В нескольких корпусах разработчики выделили «Ядро», которое составляет небольшую часть корпуса, где разметка проведена наиболее детально и отредактирована вручную. Остальная часть корпуса размечена автоматически.

Сложности с автоматической обработкой языка обусловлены в первую очередь спецификой самого языка. В случае с японским языком многие из них связаны с особенностями японской письменности, среди которых:

1) больший набор символов (иероглифы, две слоговых азбуки, латиница, арабские цифры, знаки пунктуации и др.) и несовместимость имеющихся наборов символов [Костыркин 2004: 21–22];

2) различные варианты написания одного и того же слова: оно может быть записано каной, иероглифами или сочетанием иероглифа с каной, причем каждый из этих способов может иметь несколько вариантов;
3) отсутствие пробелов в тексте, что приводит к сложностям с сегментацией текста, а она, в свою очередь, выливается в проблему морфологического и синтаксического анализа единиц текста.

Сегментация японского текста зависит от выбора его базовой единицы. Японская лингвистическая традиция выделяет единицы, которые получили название *go*. В европейской японистике этот термин обычно переводят как «слово». В статьях японских авторов на английском языке можно встретить термины «словоподобные единицы» (word-like units) и «морфемоподобные единицы» (morpheme-like units) с замечанием о том, что понятие слова для японского языка плохо разработано и что в значении термина «слово» часто используется термин «морфема», хотя и понимается этот термин иначе, чем в европейской лингвистике. На наш взгляд, критика не совсем справедлива. Дело не в недостаточной проработанности понятия, а лишь в иной его трактовке. Как отмечает В. М. Алпатов, самостоятельным *go* соответствуют единицы, которые могут состоять из корней и словообразовательных показателей, несамостоятельным *go* — как служебные слова, так и аффиксы [Алпатов 2018: 59]. Таким образом, налицо несовпадение японского *go* с европейскими «словом» и «морфемой».

Другая единица, выделяемая в японском языке, — *bunsetsu*. Она соответствует самостоятельному *go* со всеми примыкающими к нему формантами, если таковые имеются. Деление текста на *bunsetsu* кажется носителям языка наиболее естественным.

Традиции японского языкознания очевидным образом прослеживаются в идеях NINJAL. Институт предложил интересное с практической точки зрения решение для обработки текстов, состоящее в сегментации текста на «короткие слова» (SUW, short unit words), некоторые из которых затем при помощи статистических методов объединяются в «длинные слова» (LUW, long unit words)[10]. LUW составляются из SUW в основном в случае составных существительных, глаголов и частиц, т. е. в тех случаях, когда вопрос о проведении границ внутри слова представляется исследователям японского языка наиболее спорным. Например, составное существительное 公害紛争処理法 («акт об

[10] Для устной речи выделяют также «средние слова», представляющие собой «короткие слова» с добавлением супрасегментных характеристик.

урегулировании конфликтов по вопросам загрязнения окружающей среды») состоит из существительных 公害 («загрязнение окружающей среды»), 紛争 («диспут, конфликт»), 処理 («урегулирование, разрешение») и 法 («закон, акт»); частица において (локатив) — из частицы に (датив), глагола おけ («быть, находиться») и вспомогательного глагола る (настоящее время)[11] [Maekawa et al. 2014: 355]. Как видно из примера, SUW — это и не слова, и не морфемы в европейском понимании этих терминов, SUW наиболее близки традиционно выделяемым базовым единицам. LUW — единицы, которые могут быть больше SUW, но меньше *bunsetsu*. *Bunsetsu* в основном используются разработчиками разных корпусов для парсинга (но, например, в проекте Universal Dependencies[12] для японского языка используют синтаксическое слово, которое не равно *bunsetsu*).

Для морфологической разметки корпусов NINJAL использует морфологический анализатор MeCab, работающий на словаре UniDic. Этот словарь был разработан относительно недавно как универсальный словарь, решающий большинство проблем его предшественников, одной из которых был непоследовательный подход к членению текста.

В упомянутой выше работе авторы дают сравнение четырех словарей (Juman, IPA, Yahoo! и UniDic) и показывают, что одно и то же составное слово разбивается этими словарями по-разному. Словарь UniDic отличается от остальных словарей тем, что проводит наиболее дробное и единообразное членение текста [Maekawa et al. 2014: 354], что служит не только более эффективному автоматическому анализу текста, но и получению более надежных данных при поиске в корпусе текстов.

В случае с автоматической сегментацией текста программа имеет дело с письменным текстом, а значит, в этом случае речь идет именно об орфографических словах, т. е. единицах между двумя пробелами или пробелом и знаком препинания. Однако перед компьютерной лингвистикой стоит и другая, не менее сложная задача, а именно распознавание и дешифровка речи. В речевом потоке слова сливаются в некоторую последовательность звуков, происходит искажение фонетического облика слов, а значит, для распознавания речи нужны более длинные единицы, как минимум, фонетические слова. Чем длиннее единица, тем более объемным должен быть словарь, но в случае с использованием более длинных единиц снижается количество ошибок в работе программы.

[11] Сохранена терминология авторов статьи.

[12] https://universaldependencies.org/introduction

На наш взгляд, для распознавания японской устной речи наилучшим решением является *bunsetsu*. В японском языке высока частота омонимии и эллиптических конструкций, соответственно, для корректного распознавания нужен контекст. Более того, *bunsetsu* часто является и фонетическим словом, имеющим единый тоновый контур. Тоновый контур *go* в японском языке может отличаться от диалекта к диалекту и меняться даже в рамках одного диалекта в зависимости от окружения единицы в предложении, от индивидуальных характеристик говорящего и даже от ситуации общения, что осложняет процесс автоматизации расшифровки устных текстов. Поэтому так важно рассматривать фонетическое слово целиком.

5. Заключение

Японская корпусная лингвистика, зародившись в 1950-х гг., стремительно развивается. Появляется все больше представительных корпусов, проводится большое число исследований на материале созданных корпусов, разрабатываются методы усовершенствования имеющихся корпусов и построения новых.

В ходе работы над созданием корпуса текстов разработчики сталкиваются со многими сложностями, некоторые из которых являются универсальными (обеспечение репрезентативности корпуса, сбор материала и т. д.), а некоторые обусловлены спецификой языка. В случае с японским языком они связаны с особенностями японской письменности, супрасегментными характеристиками японской речи, высокой степенью омонимии, разногласием по различным теоретическим вопросам (сегментного членения текста, морфологического анализа языковых единиц и пр.). Эти проблемы значительно осложняют аннотирование (а в случае с устной речью — и распознавание) японского текста. При создании корпуса текстов не обойтись без автоматической обработки текстов, но несмотря на развитие программ по обработке японского языка, машине все еще нужна помощь человека для проведения полностью корректного анализа текстов.

Литература

1. *Алпатов В. М.* (2018), Слово и части речи. М.
2. *Кострыкин А. В.* (2004), Японские корпусные проекты. Научно-техническая информация. Сер. 2, № 9, с. 20–30.

3. 後藤 斉 *[Gotoo Hitosi]* (2007), コーパス言語学と日本語研究 [Corpus linguistics and Japanese language studies]. 日本語科学 [Japanese Linguistics]. Vol. 22, pp. 47–58, URL: doi/10.15084/00002182
4. *Maekawa K., Yamazaki M., Ogiso T., Maruyama T., Ogura H., Kashino W., Koiso H., Yamaguchi M., Tanaka M., Den Ya.* (2014), Balanced corpus of contemporary written Japanese. In: Language Resources and Evaluation. Vol. 48, pp. 345–371, URL: link.springer.com/article/10.1007/s10579-013-9261-0 (дата обращения: 12.03.2021).
5. 宮島 達夫 *[Miyajima Tatsuo]* (2007), 語彙調査からコーパスへ [From vocabulary statistics to corpus-based studies]. 日本語科学 [Japanese Linguistics]. Vol. 22, pp. 29–46, doi.org/10.15084/0000218
6. 国立国語研究所要覧 2020/2021 [National Institute for Japanese Language and Linguistics: Survey and guide 2020/2021] (2020), URL: http://id.nii.ac.jp/1328/00002825/ (дата обращения: 22.05.2021).
7. Proceedings of LREC 2018 Special Speech Sessions (2018), URL: http://doi.org/10.15084/00001908

References

1. *Alpatov V. M.* (2018), Slovo i chasti rechi [Word and Parts of Speech]. Moscow.
2. 後藤 斉 *[Gotoo Hitosi]* (2007), コーパス言語学と日本語研究 [Corpus linguistics and Japanese language studies]. 日本語科学 [Japanese Linguistics]. Vol. 22, pp. 47–58, doi/10.15084/00002182
3. *Kostyrkin A. V.* (2004), Yaponskie korpusnyie proekty [Japanese Corpus Projects]. In: Nauchno-tekhnicheskaya informatsiya [Scientific and Technical Information]. Vol. 2, No. 9, pp. 20–30.
4. *Maekawa K., Yamazaki M., Ogiso T., Maruyama T., Ogura H., Kashino W., Koiso H., Yamaguchi M., Tanaka M., Den Ya.* (2014), Balanced corpus of contemporary written Japanese. In: Language Resources and Evaluation. Vol. 48, pp. 345–371, URL: link.springer.com/article/10.1007/s10579-013-9261-0 (date of access: 12.03.2021).
5. 宮島 達夫 *[Miyajima Tatsuo] (2007)*, 語彙調査からコーパスへ [From vocabulary statistics to corpus-based studies]. 日本語科学 [Japanese Linguistics]. Vol. 22, pp. 29－46, doi.org/10.15084/0000218
6. 国立国語研究所要覧 *2020/2021* [National Institute for Japa-nese Language and Linguistics: Survey and guide 2020/2021] (2020), URL: http://id.nii.ac.jp/1328/00002825/ (date of access: 22.05.2021).
7. Proceedings of LREC 2018 Special Speech Sessions (2018), http://doi.org/10.15084/00001908

Корецкая Ирина Леонидовна
Институт языкознания РАН (Россия)
Koretskaya Irina
Institute of Linguistics of the Russian Academy of Sciences (Russia)
E-mail: koretskaya_irina@iling-ran.ru

Е. А. Корсакова

E. A. Korsakova

СОСТАВЛЕНИЕ КОРПУСА НАУЧНЫХ ПУБЛИКАЦИЙ В СФЕРЕ ОПТИКИ[1]

DESIGNING A CORPUS OF SCIENTIFIC PUBLICATIONS IN OPTICS

Аннотация. В ходе данной работы были выявлены критерии отбора научных публикаций для формирования сбалансированного и репрезентативного англоязычного текстового корпуса в сфере оптики. Кроме того, были выявлены ограничения используемой методики, возможные направления ее улучшения и дальнейшие шаги по созданию корпуса. Он будет содержать более 3500 научных работ, опубликованных в период 2010–2020 годов, и позволит анализировать научные статьи, лингвистически характеризовать научные направления и овладевать современной терминологией. На базе этих материалов в дальнейшем можно разрабатывать учебно-методические пособия для курсов ESP (English for Specific Purposes).
Ключевые слова. Корпусная лингвистика, методика построения текстовых корпусов, англоязычный корпус в области оптики, «оптический корпус», английский язык для специальных целей.

Abstract. In this study, we identify the criteria for selecting scientific publications to design a balanced and representative English-language textual corpus in the field of optics. In addition, we discuss the limitations of the method used, possible directions for its improvement and further steps toward building the corpus. It will contain papers published from 2010 to 2020, and it will allow analyzing scientific articles, revealing scientific trends and mastering state-of-the-art terminology. Basing on these data, one can develop teaching/learning materials for ESP courses (English for Specific Purposes).
Keywords. Corpus linguistics, methodology for building textual corpora, English-language corpus in optics, "optical corpus", ESP.

1. Введение

Английский язык прочно занял позиции *lingua franca* в международном научном сообществе [O'Neil 2018]. Владея им, пользователь получает доступ к огромной базе актуальной научной информации. В полной мере это касается такой научно-технической области, как оптика. Темп развития современной оптики очень высок. Ежегодно в международных научных базах цитирования появляется порядка 15 000 англоязычных научных статей, посвященных оптике, отражающих последние достижения ученых в этой области. Чтобы соответствовать современному уровню, студентам, ученым и специалистам

[1] Работа выполнена при финансовой поддержке гранта Президента РФ для магистрантов № 330-М.

в области оптики необходимо своевременно получать новые знания из надежных источников. В этой связи составление текстового корпуса научных работ представляется весьма актуальной задачей, поскольку, используя пополняемые текстовые корпусы и современные инструменты цифровой лингвистики, технические специалисты смогут быстро проводить анализ текстовых данных, выявлять научные тренды и осваивать современную терминологию [Wu 2014]. Однако, несмотря на важность этой тематики, до сих пор крайне мало исследований, посвященных созданию и изучению корпусов научных текстов по оптике [Louvigne 2014; Louvigne 2016], и, насколько известно, существует всего один доступный коммерческий корпус специальных текстов по оптике (на французском языке) [Mapelli 2016]. Составление текстового корпуса —непростая задача, потому что неизбежно возникают вопросы, связанные как с методологическими аспектами, так и с практической реализацией. В данной работе мы освещаем некоторые особенности составления специализированного текстового корпуса научных публикаций в сфере оптики.

2. Материалы и методы

Поиск и отбор научных публикаций проводился с использованием поискового инструмента библиографической и реферативной базы данных Scopus. Исходные критерии выборки:

- Предметная область: оптика;
- Временной охват: 11 лет — с 2010 по 2020 гг. (этого достаточно для проведения диахронического анализа, в то же время корпус будет содержать актуальную информацию);
- Язык: английский; авторы текстов из США, Великобритании, Канады, Австралии, Ирландии, Новой Зеландии (этим ограничением в некоторой степени обеспечивается «эталонность» языка и стиля, поскольку с большой долей вероятности среди авторов публикаций есть носители английского языка);
- Принцип доступа к публикациям: открытый (это необходимо, чтобы избежать лишних расходов и дополнительной работы с копирайтингом).

3. Исследование, результаты и обсуждение

3.1. Отбор публикаций по типу документа

За основу классификации по типу документа была взята классификация, принятая в базе Scopus; был проведен поиск документов в соответствии с рассматриваемой тематикой. На рис. 1 показано распределение публикаций с ключевым словом «оптика» по типам документов (количество документов). К примеру, в двух самых больших сегментах количество публикаций с типом «статья» включает 222 276 наименований, в то время как тезисов конференций — 140 713. Остальные сегменты намного меньше.

Стоит отметить, что мы сознательно избегали привлечения в качестве источника материала монографий и учебников, поскольку это не в полной мере первоисточники и в них велика инерция представления актуальной информации. Соответственно, рассматривая элементы корпуса в диахронном аспекте, мы бы получили не совсем корректную информацию. Кроме того, количество токенов в книгах значительно больше, чем в средней статье или тезисах конференции, что сделало бы выборку неоднородной с точки зрения объема текстов.

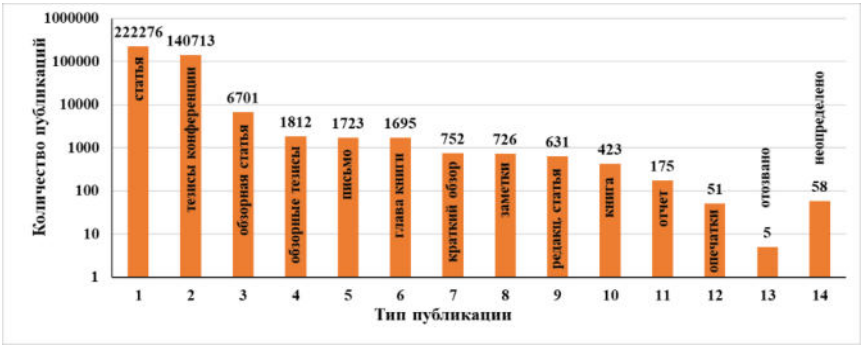


Рис. 1. **Распределение публикаций по типам документов**

С учетом вышесказанного, рассматривалось два варианта отбора публикаций по типу документа — пропорциональный отбор от каждого сегмента либо паритетная выборка из двух самых крупных сегментов. При выборе первого варианта репрезентативность корпуса оказывается выше, но пополнять корпус при этом весьма проблематично. Так как основная часть документов представлена статьями

и тезисами конференций и нет значимых отличий по используемой специальной терминологии от жанров обзора, письма и др., было решено привлечь только эти два типа документов. Аналогичное решение принято в корпусе современного американского английского языка (Corpus of Contemporary American English, COCA), в котором тексты поровну разделены между крупными жанрами [Davies 2009]. Причем этот состав относится как к корпусу в целом, так и к каждому временному периоду (году), который в нем представлен. А корпус исторического американского английского (Corpus of Historical American English, COHA), помимо этого, размечен по десятилетиям в плане жанров (например, книга, статья, обзор) и научных областей [Davies 2012]. Это означает, что исследователи могут диахронически сравнивать корпусные данные и отслеживать изменения в языке на материале однотипных текстов.

3.2. Отбор публикаций по тематической области

В области оптики исследования часто бывают междисциплинарными, и все очевиднее тенденция к усилению этого эффекта, так как мы живем в эпоху активной интеграции научных направлений и развития стыковых предметных областей. Этим объясняется многообразие представленных научных направлений. Первые десять позиций в ранжированном списке количества публикаций в той или иной области знаний (согласно данным базы Scopus), отобранных по ключевому слову «оптика», приведены в табл. 1.

Как видно из таблицы, самые крупные тематические области — физика и астрономия, инженерия и материаловедение — могут сформировать сбалансированный текстовый корпус, так как вклад этих секторов количественно близок. Поэтому мы решили ограничиться только этими областями. Таким образом, с учетом всех критериев был составлен алгоритм отбора научных публикаций для составления корпуса текстов сферы «Оптика» (см. табл. 2).

В итоге в соответствии с алгоритмом отбора получилось ограничить число публикаций до 20 464. Распределение публикаций по годам в процентном выражении варьировалось от 8 до 11 %, а в натуральном выражении от 1549 до 2307. Далее характеристики всех статей были экспортированы в таблицу Excel. Данные содержали информацию: Автор(ы), Название документа, Год, Название источника, том, выпуск, страницы, Количество цитирований, Источник и тип документа, DOI, Издатель, Ключевые слова.

Таблица 1. **Распределение публикаций в сфере оптики по тематическим областям**

№	Тематическая область	Количество публикаций
1	Физика и астрономия	239 269
2	Инженерное дело	203 075
3	Материаловедение	152 984
4	Информатика	92 131
5	Математика	67 907
6	Химия	33 458
7	Медицина	22 452
8	Науки о Земле и планетах	13 917
9	Биохимия, генетика и молекулярная биология	12 792
10	Химическая инженерия	10 161

Таблица 2. **Алгоритм отбора научных публикаций для формирования корпуса текстов сферы «Оптика»**

Алгоритм отбора публикаций	Кол-во публ.
Ключевое слово: оптика	377 741
Доступ: открытый	79 034
Тип документа: статья, тезисы конференций	75 664
Тематическая область: физика и астрономия, инженерия, материаловедение	65 455
Язык: английский	64 256
Страны: основные англоязычные страны (N = 6)	30 797
Временной охват: 11 лет, с 2010 по 2020 год	20 464

На данном этапе мы выявили ограничение методики: настройки сайта Scopus не позволяют экспортировать данные больше 2000 документов за одно скачивание. Поэтому данные экспортировали отдельно по каждому году и в те периоды, на которые приходилось больше 2000 публикаций, количество метаданных публикаций было лимитировано числом 2000.

3.3. Принцип случайности отбора публикаций

После сбора выходных данных публикаций был рассчитан объем репрезентативной выборки *CSS* по формуле:

$$CSS = \frac{SS}{1 + \frac{SS - 1}{POP}}, \quad (1)$$

где *ss* — размер выборки в общем случае при заданных точности и погрешности, *pop* — генеральная совокупность.

При точности (доверительной вероятности) 95% и погрешности (доверительном интервале) 5% *ss* составляет 384,16. Таким образом, для верхнего предела генеральной совокупности (2000 публикаций) объем репрезентативной выборки равен 322. Несмотря на то, что объем репрезентативной выборки в некоторые годы меньше 322, необходимо было применить это число публикаций как верхний предел, так как сбалансированность корпуса обеспечивается равным вкладом всех временных периодов. Для соблюдения принципа репрезентативности при отборе публикаций был использован генератор случайных чисел.

3.4. Ограничения методики и возможности ее корректировки

Оказалось, что не все публикации, отмеченные в базе Scopus как публикации открытого доступа, на самом деле доступны для скачивания. В связи с этим необходимо обрабатывать гораздо больший массив выходных данных публикаций, чем предполагалось ранее. Поэтому в дальнейшем необходимо будет решить вопрос автоматизации загрузки статей. М. Дэвис [Davies 2009] использовал VB.NET (интерфейс и язык программирования), с помощью которого создал определенный сценарий загрузки текстов. Подобная автоматизация процесса построения корпуса была бы оптимальна и в нашем случае. Однако существует проблема легитимности парсинга веб-страниц, поскольку его запрещено применять для подавляющего большинства издательских сайтов (см., например, лицензию Elsevier Text and Data Mining (TDM) [Elsevier Policies]).

4. Выводы

В рамках данной работы составлен алгоритм отбора научных публикаций для формирования сбалансированного и репрезентативно-

го англоязычного текстового корпуса в сфере оптики. Создана база данных, содержащая информацию о более чем 3500 научных работ, опубликованных в период с 2010 по 2020 гг. В соответствии с этими данными проведен отбор текстового материала за 2020 год. Выявлены ограничения предлагаемой методики и способы ее улучшения. Дальнейшие шаги по созданию корпуса будут включать решение вопросов автоматизации загрузки документов, сбора всего необходимого материала, а также разработки архитектуры корпуса.

References

1. *Davies M.* (2009), The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. In: International Journal of Corpus Linguistics. Vol. 14(2), pp. 159–190.
2. *Davies M.* (2012), Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English, Corpora. Vol. 7(2), pp. 121–157.
3. Elsevier Policies. Text and data mining. Elsevier TDM. URL: https://www.elsevier.com/about/policies/text-and-data-mining/elsevier-tdm-license (дата обращения: 01.07.2021).
4. *Louvigne S., Shi J., Sharmin S.* (2014), A Corpus-based Analysis of the Scientific RA Genre and RA Introduction, International Conference on Advanced Mechatronic Systems. No. 6911636, pp. 123–127.
5. *Louvigne S., Jie S.* (2016), Data-Driven Analysis of the Development of Linguistic Features in Research Articles on Optics, International Conference on Advanced Mechatronic Systems. No. 7813502, pp. 516–520.
6. *Mapelli V.* (2016), VERBA Polytechnic and Plurilingual Terminological Database — B — MP Optics, ISLRN: 642-374-718-061-9. In: ELRA Catalogue of Language Resources. URL: http://catalog.elra.info/en-us/repository/browse/ELRA-T0126/(дата обращения: 01.07.2021).
7. *O'Neil D.* (2018), English as the lingua franca of international publishing, World Englishes. Vol. 37(2), pp. 146–165.
8. *Wu L.-F.* (2014), Motivating college students' learning English for specific purposes courses through corpus building, English Language Teaching. Vol. 7(6), pp. 120–127.

Корсакова Елена Анатольевна
Уральский федеральный университет им. первого Президента России Б. Н. Ельцина (Россия)
Korsakova Elena
Ural Federal University named after the first President of Russia B. N. Yeltsin (Russia)
E-mail: korsakovaea@mail.ru

У. Е. Кочеткова, П. А. Скрелин

U. E. Kochetkova, P. A. Skrelin

ПАРАЛИНГВИСТИЧЕСКИЕ ЯВЛЕНИЯ ПРИ ВЫРАЖЕНИИ ИРОНИИ В РУССКОМ ЯЗЫКЕ (НА МАТЕРИАЛЕ МУЛЬТИМЕДИЙНОГО КОРПУСА ИРОНИЧЕСКИХ ВЫСКАЗЫВАНИЙ)[1]

PARALINGUISTIC PHENOMENA ACCOMPANYING IRONY EXPRESSION IN RUSSIAN (A CASE STUDE OF THE MULTIMEDIA CORPUS OF IRONIC UTTERANCES)

Аннотация. В работе рассматриваются особенности мимики и жестикуляции при выражении иронии в русском языке на материале мультимедийного корпуса, включающего синхронные аудио- и видеозаписи иронических и омонимичных им нейтральных высказываний в произнесении 10 дикторов. Результаты исследования свидетельствуют о различном характере паралингвистических явлений и их различном соотношении с интонационным членением в двух типах высказываний.
Ключевые слова. Ирония, мультимедийный корпус, интонационные характеристики, мимика, жестикуляция, лабораторная речь.

Abstract. This paper deals with the peculiarities of the mimics and gesture accompanying expression of irony in Russian language. The research is based on the multimedia corpus, which includes synchronized audio and video recordings. The recorded material consists of homonymous ironic and non-ironic utterances of various communicative types. We analyzed the paralinguistic phenomena accompanying both types of utterances in speech of 10 Russian native speakers (5 men and 5 women). The obtained results show the difference between mimics and gesture corresponding to the ironic and non-ironic utterances, as well as the their interplay with boundaries of prosodic units and the place of the intonation focus.
Keywords. Irony, multimedia corpus, prosodic boundaries, mimics, gesture, laboratory speech.

1. Введение

В настоящее время большое значение приобретают полимодальные исследования функционирования различных эмоций и коннотаций в речи. Это во многом связано с развитием диалоговых систем «человек — машина», для успешного функционирования которых необходимо автоматическое распознавание с высокой степенью надежности.

[1] Данное исследование поддержано грантом РФФИ № 20-012-00552 «Акустические характеристики иронии при реализации функциональных интонационных моделей».

Особое внимание при этом уделяется иронии, в частности, одной из ее разновидностей — антифразису. Данный тип иронии понимается как отрицание прямого лексического значения за счет интонационного оформления, а также за счет других лингвистических или паралингвистических средств. Важность правильного распознавания данной коннотации чрезвычайно высока, так как неверное истолкование высказывания, в котором конечный смысл не равен лексическому содержанию, может нанести урон коммуникации в целом.

С 80-х годов XX века началось активное изучение иронии не только как стилистической фигуры, но и как когнитивного явления, находящего свое отражение и в письменной, и в устной речи [Mishra et al. 2016]. На протяжении последних пятнадцати лет появилось большое количество работ, в которых с помощью автоматической обработки больших корпусов данных рассматривается выражение иронии в письменном тексте, как правило, на материале высказываний с различными хэш-тегами в социальной сети «Твиттер» [Mishra et al. 2016; Michael, Zahra 2019]. Для большинства новейших исследований в этой области характерно использования айтрекинга [Mishra et al. 2016].

Корпусные исследования иронии и сарказма с помощью методов машинного обучения проводились группами ученых из Индии, США, Великобритании, Испании и Индонезии. Мультимедийные корпусы, включающие как текст, так и аудиозаписи, были построены, как правило, на материале английского языка и включали в себя аудиозаписи телефонных диалогов, иронические высказывания из сериалов, представленных на канале MTV, а также из фильмов и сериалов, представленных на канале YouTube. Использованные в них подходы и полученные результаты описаны в работе индонезийских ученых [Michael, Zahra 2019].

Несмотря на большое количество исследований, посвященных выражению в речи иронии (сарказма), акустические характеристики этой коннотации до сих пор изучены лингвистами недостаточно. Отдельные сведения, полученные на материале европейских языков, можно найти в работах, появившихся в начале XXI века [Braun, Schmiedel 2018; Bryant et al. 2005; Niebuhr 2016]. Тем не менее, приведенные в них данные не являются исчерпывающими. Некоторые указания на фонетические характеристики иронии в русском языке содержатся в работе С. В. Кодзасова, в остальных же исследованиях на материале русского языка рассматриваются в большей степени лексико-грамматические способы выражения вербальной иронии и их

взаимосвязь с особыми интонационными конструкциями, как было отмечено ранее [Skrelin et al. 2020].

Описание мультимодальных корпусов, включающих одновременно записи речи и мимики (жестов), дано в работе П. Вагнер [Wagner 2014]. Однако результаты, полученные при анализе этих корпусов, несколько разнятся, что связано с различиями в выборе авторами классификации и способа аннотации просодических и паралингвистических явлений. Ни один корпус не содержит записи на русском языке.

В связи с тем, что для русского языка отсутствует полномасштабное описание фонетических характеристик иронии, а также сопутствующих жестов и мимики, нами был составлен корпус лабораторной речи, в который вошли высказывания с иронией и идентичные им по лексическому составу высказывания без иронии. При этом в корпусе представлены все коммуникативные типы высказываний, предполагающие появление различных интонационных конструкций.

Главной целью настоящего исследования стало сравнение мимики и жестов в иронических высказываниях и в омонимичных им нейтральных высказываниях, а также сравнение локализации паралингвистических явлений по отношению к границам единиц членения.

2. Материал исследования

Одновременная аудио- и видеозапись проводилась в звукоизолированной кабине на кафедре фонетики и методики преподавания иностранных языков СПбГУ. Всего в записи приняли участие 55 дикторов. Однако для настоящего исследования были отобраны записи 10 дикторов 17–30 лет (5 мужчин и 5 женщин), в речи которых наблюдалось яркое интонационное оформление высказываний с иронией. Все дикторы являлись носителями русского нормативного произношения и не имели специального актерского образования.

Каждым диктором был прочитан набор из 60 мини-диалогов или коротких текстов, состоящих из 2–4 фраз. В эти тексты вошли омонимичные целевые высказывания различных коммуникативных типов с иронией и без иронии. Например: «Целыми днями чай пьет. *Незаменимый работник,* как же… Давно пора его уволить!» и «*Незаменимый работник!* Всё делает на совесть». Короткие тексты и мини-диалоги были представлены дикторам в случайном порядке. Перед испытуемыми не ставилась задача прочитать «иронично» или «нейтрально», это следовало из читаемого ими контекста. Всего в данном исследова-

нии было рассмотрено 300 иронических и 300 нейтральных реализаций целевых фрагментов.

Аннотация паралингвистических явлений проводилась вручную с использованием программы ELAN; для осуществления фонетической аннотации (также вручную) использовались совместимые программы акустической обработки сигнала Praat и Wave Assistant.

Несмотря на то, что в специальной литературе встречаются различные подходы к описанию мимики и жестов [Wagner 2014], в настоящем исследовании мы исходили из разделения на мимические жесты, производимые лицевыми мышцами (мимика), и жесты как таковые (жестикуляция), под которыми понимаются движения, производимые различными частями тела (в данном случае головой, руками и плечами). Пантомимика и проксемика не рассматривались.

Необходимо отметить, что наравне с микропросодикой в вербальном поведении человека, на паралингвистическом уровне наблюдаются явления, обусловленные необходимостью реализации того или иного артикуляторного уклада. При этом индивидуальные особенности строения речевого аппарата могут диктовать более или менее яркие артикуляторные жесты. Другие движения, не будучи физиологически обусловлены, могут являться следствием индивидуальных поведенческих стереотипов. В настоящем исследовании не учитывались такие типичные для говорящего жесты (статические жесты), как постоянный наклон головы и другие индивидуальные черты. Фазы жестов и мимических движений также не рассматривались в рамках данного анализа, тогда как локализация данных движений по отношению к синтагменной границе, наоборот, представляла особый интерес.

3. Результаты исследования

3.1. Мимика

Основные два вида движений мимической мускулатуры — движение бровей и губ — встречались в обоих типах высказываний, однако их характер, частота, сочетаемость и дистрибуция различались.

Мимика сопутствовала 87 % иронических высказываний. Из них 46 % были произнесены с улыбкой, 43 % сопровождались подъемом либо сведением к центру бровей и 11 % соотносились с комплексной мимикой. При этом движение бровей присутствовало в высказываниях всех коммуникативных типов. Улыбка же имела ассиметричный

характер, либо движение осуществлялось только верхней или только нижней губой. В тех случаях, когда целевой фрагмент с иронией содержал огубленные гласные, губная артикуляция носила более выраженный характер, чем того требует нейтрально-нормативное произношение. Эти данные хорошо согласуются с имеющимися в литературе наблюдениями о проявлении гиперартикуляции на сегментном уровне при выражении иронии [Niebuhr 2006]. В некоторых случаях частичная огубленность возникала и на неогубленных гласных. Однако последнее наблюдение требует проверки на большем количестве материала, возможно, с использованием артикулографа.

Нейтральным высказываниям также соответствовали мимические движения, однако доля таких высказываний была меньше — 61 % от общего числа нейтральных высказываний. Движение бровями присутствовало в 52 % таких высказываний и совпадало (в 92 % случаев) с вопросительными конструкциями. Улыбка также возникала достаточно часто — в 49 % высказываний. Подобной мимикой сопровождались повествовательные и восклицательные высказывания, в которых присутствовала мелиоративная лексика.

3.2. Жестикуляция

В повествовательных и восклицательных нейтральных высказываниях дикторы часто осуществляли кивок головой, как правило, единичный и совпадающий с интонационным центром высказывания (в 67 % случаев). Также отмечалось движение головой вперед, по-видимому, характеризующее присутствие установки на продолжение общения. В вопросах без иронии возникало и движение головой в сторону, хотя подобный жест оказался достаточно редким и характеризовал лишь 4 % нейтральных высказываний.

В отличие от единичного утвердительного кивка головой в эмоционально нейтральных высказываниях, целевые фрагменты с иронией в большинстве случаев сопровождались частым повторяющимся киванием (иногда еле заметным, иногда ярко выраженным); подобное явление сопровождало 89 % иронических высказываний. Реже встречалось движение головой из стороны в сторону либо вниз (в 28 % высказываний). Последнее совпадало, как правило, с большим раствором рта на неогубленных гласных нижнего подъема и их растягиванием. В некоторых случаях это приводило и к выдвижению вперед нижней челюсти.

Движения плечами были крайне редкими в исследованном нами материале (2,5 % от общего числа иронических высказываний), что, по всей видимости, связано с условиями записи лабораторной речи.

3.3. Смех

В отличие от улыбки, которая, как правило, отражается на сегментных характеристиках высказывания, смех может приводить к изменению просодических характеристик. Кроме того, данный паралингвистический маркер (короткий смех, предшествующий целевой синтагме или следующий за ней) хорошо опознается аудиторами как маркер иронии. В данном исследовании смех наблюдался в 22 % иронических высказываний. Однако, по всей видимости, эта характеристика иронических высказываний является дикторозависимой (у 6 из 10 дикторов данная характеристика присутствовала менее чем в 5 % высказываний). В нейтральных высказываниях смех отсутствовал у всех дикторов.

3.4. Выводы

Сравнивая иронические и неиронические высказывания, можно сделать ряд выводов и наблюдений. Во-первых, для высказываний с иронией характерно одновременное использование диктором жестов и мимики, а также смеха, тогда как в нейтральных высказываниях, как правило, наблюдается лишь один тип явлений. Во-вторых, если в нейтральных высказываниях определенные жесты совпадают с интонационным центром целевой синтагмы, то в иронических высказываниях и жестикуляция, и мимика зачастую реализуются и в постцентровой части. В-третьих, целевые синтагмы с иронией, которые не сопровождались паралингвистическими явлениями, составили 4,5 %, однако и жестикуляция, и мимика появлялись в таком случае в соседних, следующих за ними синтагмах. Это же касается и интонационного оформления, которое в целевых синтагмах могло не отличаться от нейтрального, но было реализовано в следующем за целевой синтагмой отрывке либо на предшествующем лексическом маркере.

4. Заключение

Полученные в ходе исследования данные позволяют говорить о достаточно сложной картине функционирования паралингвистических явлений как в нейтральных высказываниях, так и в высказываниях

с иронией. При этом основное отличие состоит не в наличии или отсутствии данных явлений в том или ином типе речи, но в их характере, а также в их сочетаемости. Кроме того, важна и локализация данных явлений в нейтральной речи и в речи с иронией. Можно сделать вывод и о параллелизме в фонетическом и паралингвистическом оформлении отрывков с иронией, при котором яркое оформление будет соответствовать не целевой синтагме, а предшествующему или последующему контексту. При этом жестикуляция и мимика могут совпадать по времени с паузой, создавая таким образом «паралингвистический маркер» иронии, существующий отдельно от целевого отрывка так же, как и маркер лексико-семантический.

Литература

1. *Braun A., Schmiedel A.* (2018), The phonetics of ambiguity: A study on verbal irony. In: Cultures and Traditions of Wordplay and Wordplay Research. De Gruyter, Berlin, pp. 111–136.
2. *Bryant G., Fox Tree J.* (2005), Is there an Ironic Tone of Voice? Language and Speech. Vol. 48, pp. 257–277.
3. *Michael S., Zahra A.* (2019), Automatic sarcasm detection with textual and acoustic data. In: International Journal of Recent Technology and Engineering. Vol. 8, issue 4, pp. 1357–1360.
4. *Mishra A., Bhattacharyya P., Kanojia D.* (2016), Predicting readers' sarcasm understandability by modeling gaze behavior. In: AAAI Conference. Vol. 30, pp. 3747–3753.
5. *Niebuhr O.* (2016), Rich Reduction: Sound-segment residuals and the encoding of communicative functions along the hypo-hyper scale. In: 7th Tutorial and Research Workshop on Experimental Linguistics. Saint Petersburg, pp. 11–24.
6. *Skrelin P., Kochetkova U., Evdokimova V., Novoselova D.* (2020), Can we detect irony in speech using phonetic characteristics only? Looking for a methodology of analysis. In: Proceedings of the 22nd International Conference SPECOM, LNAI. Springer, Heidelberg. Vol. 12335, pp. 544–553.
7. *Wagner P.* (2014), Gesture and speech in interaction: an overview. In: Speech Communication. Vol. 57, pp. 209–232.

References

1. *Braun A., Schmiedel A.* (2018), The phonetics of ambiguity: A study on verbal irony. In: Cultures and Traditions of Wordplay and Wordplay Research. De Gruyter, Berlin, pp. 111–136.
2. *Bryant G., Fox Tree J.* (2005), Is there an Ironic Tone of Voice? Language and Speech. Vol. 48, pp. 257–277.
3. *Michael S., Zahra A.* (2019), Automatic sarcasm detection with textual and acoustic data. In: International Journal of Recent Technology and Engineering. Vol. 8, issue 4, pp. 1357–1360.

4. *Mishra A., Bhattacharyya P., Kanojia D.* (2016), Predicting readers' sarcasm understandability by modeling gaze behavior. In: AAAI Conference. Vol. 30, pp. 3747–3753.
5. *Niebuhr O.* (2016), Rich Reduction: Sound-segment residuals and the encoding of communicative functions along the hypo-hyper scale. In: 7th Tutorial and Research Workshop on Experimental Linguistics. Saint Petersburg, pp. 11–24.
6. *Skrelin P., Kochetkova U., Evdokimova V., Novoselova D.* (2020), Can we detect irony in speech using phonetic characteristics only? Looking for a methodology of analysis. In: Proceedings of the 22nd International Conference SPECOM, LNAI. Springer, Heidelberg. Vol. 12335, pp. 544–553.
7. *Wagner P.* (2014), Gesture and speech in interaction: an overview. In: Speech Communication. Vol. 57, pp. 209–232.

Кочеткова Ульяна Евгеньевна
Санкт-Петербургский государственный университет (Россия)
Kochetkova Uliana
Saint Petersburg State University (Russia)
u.kochetkova@spbu.ru

Скрелин Павел Анатольевич
Санкт-Петербургский государственный университет (Россия)
Skrelin Pavel
Saint Petersburg State University (Russia)
p.skrelin@spbu.ru

Г. И. Кустова

G. I. Kustova

ИНФИНИТИВНЫЕ КОНСТРУКЦИИ С ПРЕДИКАТИВАМИ РАЗНЫХ СЕМАНТИЧЕСКИХ КЛАССОВ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА[1]

INFINITIVE CONSTRUCTIONS WITH PREDICATIVES OF VARIOUS SEMANTIC CLASSES IN THE RUSSIAN NATIONAL CORPUS

Аннотация. Рассматриваются конструкции «предикатив ощущения/эмоциональной реакции/интерпретации + инфинитив». Для ощущений типичны инфинитивы движения и положения (*Темно ехать*; *Холодно стоять*), для эмоций – инфинитивы, обозначающие канал информации: *Больно видеть, как гибнут леса* (→ *Больно, что гибнут леса*). В контексте интерпретаций инфинитив обозначает валентность содержания: *Бессердечно разлучать мать с детьми.*
Ключевые слова. Предикатив, инфинитивная конструкция, предикаты интерпретации.

Abstract. The paper considers constructions «predicative + infinitive» with predicatives denoting sensations ([*I'm*] *cold* / *bitter* / *cramped*), emotions ([*I'm*] *sad* / *scared* [*to leave*]), interpretations ([*It is*] *heartless*). The infinitives of the perception, mental, speech verbs (informational verbs) are typical for emotional reactions: *It hurts* / *scares to see how forests are dying* ('X sees, X is scared') → 'It hurts that forests are dying (P)' (content valency P). For interpretive constructions, the subjects of the predicative and the infinitive do not coincide: *It is heartless to separate the mother from the children* – 'X separates, Y evaluates such an act as heartless'. The infinitives of perceptual and mental verbs in such a construction are either not used, or they denote a kind of action: *It is tactless to listen to private conversations.*
Keywords. Predicative, infinitive construction, interpretation predicatives.

1. Предикативы в грамматике

Грамматические свойства и синтаксические конструкции предикативов ([Сердобольская, Толдова 2014; Циммерлинг 2017; 2018; Летучий 2018]) коррелируют с семантикой. Характеристики пространства (*жарко, темно, тесно*) и физические/физиологические состояния человека (*холодно*, *больно*, *горько*) имеют только валентность субъекта (***На улице*** *темно*; ***Ему*** *больно*/*холодно*); перцептивные (*видно*, *слышно*), эмоциональные (*страшно*), ментальные (*известно*, *интересно*), оценочные (*хорошо*, *плохо*) и модальные (*нужно*, *необходимо*) предикативы кроме субъекта имеют валентность содержания (ситуация

[1] Исследование выполнено при финансовой поддержке РФФИ и Национального научного фонда Болгарии, проект № 20-512-18005.

Р), выражаемую, например, клаузой (*Слышно, **как** шумит вода*; *Неизвестно, **где** он*). Некоторые группы предикативов образуют конструкцию с подчиненным инфинитивом, которая и является предметом рассмотрения.

Инфинитив является естественным зависимым для модальных предикативов (как и для модальных глаголов): *необходимо/нужно/можно успеть*; ментальные и перцептивные предикативы, наоборот, не сочетаются с инфинитивом (**видно ехать*; **понятно читать*). Другие группы предикативов могут «втягиваться» в инфинитивную конструкцию (далее — ИК), и именно корпусные данные позволяют оценить масштабы этого процесса в русском языке. Нас будет интересовать также статус инфинитива при разных по семантике предикативах — сирконстант или актант (см. ниже). В данной работе мы рассмотрим три группы ИК — с предикативами ощущения, эмоции и интерпретации.

2. Ощущения

Большая группа предикативов обозначает физические/физиологические ощущения — реакции человека (1) на воздействие среды (*холодно*, *темно*, *сыро*, *душно*, *тесно*), назовем их пространственными, или (2) на воздействие объекта, с которым контактирует человек, — *мягко*, *жестко*; *горько*, *кисло*; назовем их контактными (ср. [Кустова 2013]).

Предикативы ощущения имеют одну валентность — субъекта-экспериенцера (*Мне душно*) и не требуют выражения источника воздействия, хотя он всегда есть (см. [Кустова 2004]).

2.1. Пространственные предикативы

2.1.1. Температура. Температурные предикативы сочетаются с глаголами местонахождения или положения в каком-то месте/пространстве/помещении, ср. *стоять*, *сидеть*, *лежать*; к ним примыкают глаголы длительного или постоянного пребывания *спать*, *ночевать*, *ждать*, *жить*, *зимовать*: *На «залавках» печи **тепло** и удобно **сидеть** в долгие зимние вечера* [«Наука и жизнь», 1979]; *Ночью шел дождь и было **прохладно спать*** [Николай II. Дневники (1913–1916)]; *Дул ветер и было **холодно стоять*** [Стругацкие. Попытка к бегству (1962)]; с глаголами движения: ***Холодно** будет **ехать*** [А. А. Фет. Мои воспоминания (1862–1889)].

Между предикативом и глаголом локализации/перемещения возникают временные отношения: ‘когда X делает V [в некотором пространстве/помещении], ему холодно/тепло’. Между ситуациями предикатива и инфинитива нет причинной связи: человеку холодно не потому, что он едет, гуляет, спит, а потому, что низкая температура воздуха/предмета. Т. е. инфинитив обозначает не семантическую валентность предикатива, а обстоятельственную рамку — ту ситуацию, во время которой человек испытывает соответствующее ощущение. Конструкции с непространственными глаголами имеют то же значение — ‘X-у холодно (в помещении/пространстве), когда он делает V’: ***Холодно заниматься****, особенно зябнут руки* [А. Гнедин. Письма (1939–1941)]; *Да пойдем-ка в избу, там* ***потеплей*** *будет нам* ***разговаривать*** [П. И. Мельников-Печерский. В лесах (1871–1874)].

Особая группа контекстов — инфинитивы глаголов информационных процессов (*смотреть*, *слушать*, *думать*, *вспоминать* и подобные); назовем их информационными глаголами. Эти конструкции интересны тем, что, воспринимая некоторую ситуацию, субъект испытывает настоящее (а не метафорическое) ощущение: *Дети поют на сыром балтийском морозе; мне* ***холодно*** *на это* ***смотреть*** [А. Дмитриев. Закрытая книга (1999)]; *Я сидел, как ледяной столб, и все тело — мне* ***холодно вспомнить*** *об этом даже сейчас — кричало тягучей болью обморожения* [С. Бабаян. Ротмистр Неженцев (1995)].

Температурные предикаты — наиболее многочисленная группа ИК, количественные данные представлены в табл. 1:

Таблица 1. **Ощущения**

ПРЕДИКАТИВ	ИНФИНИТИВ		
	Простр. глаголы	**Непростр. глаголы**	**Информац. глаголы**
Холодно	63	8	6
Жарко	31	8	1
Тепло	23	–	1
Прохладно	7	2	–
Зябко	7	3	2
Морозно	2	–	–

2.1.2. Другие пространственные предикативы. Другие предикативы, характеризующие пространство (*душно, тесно, просторно, свободно*, *пыльно*, *сыро*, *грязно*), также сочетаются с инфинитивами глаголов местонахождения и движения: *Нам было **душно и тесно жить** в каменной коробке* [М. Горький. Двадцать шесть и одна (1899)]; *В городской квартире было **тесно двигаться*** [«Октябрь», 2001]; *Тогда еще **просторнее будет гулять**, не надевая скафандр* [К. Э. Циолковский. Вне Земли (1916)]; ***Грязно**, **сыро идти** в дождь через поле да лес* [В. А. Солоухин. Капля росы (1959)]; с непространственными глаголами: *Детям здесь было **свободно играть*** [Н. А. Морозов. Повести моей жизни (1912)]; *Мама переставила лампу, чтобы было **просторней гадать*** [В. А. Каверин. Открытая книга (1949–1956)].

Предикативы освещенности сочетаются с глаголами движения (*светло/темно ехать, ездить, идти, ходить*), ср. *Мужики стали просить подождать до луны, иначе **темно ехать*** [И. А. Гончаров. Фрегат «Паллада» (1855)], но не с глаголами положения (*?темно стоять*; *?светло лежать*), так как для движения освещенность релевантна, а для положения — нет. Кроме того, характерным контекстом являются ситуации, в которых важную роль играет зрение: *Уже **темно читать*** [В. Осеева. Динка (1959)]; *И чтоб **светлей стало искать**, смахнул со стола книгу, загораживающую ночник* [В. Панова. Который час? (1941–1963)].

2.2. Предикативы контакта с объектом

2.2.1. Комфорт/дискомфорт. Предикативы ощущений от контакта с объектом сочетаются с глаголами положения и движения: *На голом полу **жестко спать*** [Ф. М. Достоевский. Идиот (1869)]; *Ему было **мягко** и уютно **сидеть** на турецком диване* [А. П. Чехов. Бабье царство (1894)]; *Я ковер, по которому **мягко ступать**…* [А. И. Цветаева. Королевские размышления (1914)].

При этом конструкции с непространственными глаголами типа *?На диване мягко смотреть телевизор*; *?В саду на скамейке жестко читать* в данный момент в НКРЯ не представлены, хотя ничего абсурдного в таких сочетаниях нет (ср. метонимическую интерпретацию: ‘мягко сидеть и смотреть телевизор’). При этом аналогичные сочетания с пространственными предикативами встречаются: *— Верно, вам **жарко рассказывать**, Нилыч. — Мне! Я даже уважаю жару* [К. М. Станюкович. На другой галс (1900)], ср. также: *холодно читать*; *В избе теплее разговаривать*.

2.2.2. Вкусовые реакции. Другой тип контакта (и канал восприятия) — вкусовые реакции: *сладко*, *горько*, *кисло*, *солоно*, *вкусно*. В собственно вкусовом значении эти предикативы встречаются в контексте глаголов еды и питья. В НКРЯ встретилось по одному примеру со *сладко и горько*: *По чистому дну струится чистая холодная вода, которую так* ***сладко пить****, когда в жару объешься спелой земляникой* [В. А. Солоухин. Капля росы (1959)]; *Обитателям Островов трудно терпеть боль и* ***горько глотать*** *лекарства* [«Трамвай», 1991]; ни одного примера с *кисло* и *солоно* (в исходном значении); несколько примеров с *вкусно*: *Шоколад так* ***вкусно есть*** [«100 % здоровья», 2003]; *Было удобно сидеть, было* ***вкусно пить*** *чай с медом* [Фридрих Горенштейн. Куча (1982)]. Ср. также с информационным глаголом: *Глядя на Женькино гастрономическое наслаждение, Тома поняла, что ей* ***вкуснее смотреть****, как ест ребенок, чем есть самой…* [Л. Улицкая. Казус Кукоцкого, 2000].

ПРИМЕЧАНИЕ. Разумеется, пространственные и контактные предикативы могут употребляться в метафорическом значении психологического состояния, ср. *кисло/солоно/сумрачно жить*; *тепло видеть*.

3. Эмоциональные реакции (эмоции)

Предикативы эмоциональных реакций имеют пропозициональную валентность P. Эта ситуация P является и причиной эмоции, и ее содержанием: *Грустно, что мы расстаемся* (эмоция изначально включает переживание и оценку, и в конструкциях с зависимой клаузой часто происходит сдвиг в сторону оценки: *Грустно/горько/больно/страшно, что P* ≈ ‘плохо, что P’).

ИК с эмоциями распадаются на две большие группы.

В первой группе ИК инфинитив обозначает ситуацию P, которая является причиной-содержанием эмоции (валентностью предиката): *…****Мне горько уезжать*** *из России. Я здесь родился, и всем, что имею за душой, обязан ей* [И. Бродский (1972)] — ‘горько, что я уезжаю’; *И нам было* ***грустно покидать*** *гимназию. Мы свыклись с ней* [К. Г. Паустовский. Книга о жизни (1946)].

Во второй группе ИК употребляются инфинитивы информационных глаголов. Такой инфинитив обозначает не ситуацию P (содержание эмоции), а ситуацию V — тот информационный канал, через который в сознание субъекта поступает или актуализируется информация о ситуации P. При этом предикатив обозначает реакцию не на

ситуацию V (*смотреть*, *слушать*, *думать*, *вспоминать* и т. д.), а на ситуацию P: ***Грустно подумать****, что все это погибло навсегда* [Л. И. Арнольди. Мое знакомство с Гоголем (1862)] — ‘грустно, что все погибло’, а не ‘грустно, что подумал’; *Ему было только* ***больно и горько слушать*** *упреки этой женщины* [В. В. Крестовский. Панургово стадо (1869)] — ‘больно и горько, что упрекает’.

Такое соотношение предикатива и инфинитива аналогично соотношению в группе ощущений: инфинитив не является семантической валентностью предикатива, валентностью является ситуация P, а инфинитив — своего рода связка, «проводник».

4. Интерпретации

В широком смысле интерпретация — это разновидность оценки (ср. [Апресян 2006]), однако это оценка, которая делается «со стороны», другим субъектом (если эту оценку осуществляет сам субъект ситуации P, он смотрит на себя и свой поступок как бы чужими глазами). Если «обычные» оценки могут относиться к любым ситуациям, то интерпретации относятся к поступкам и поведению людей. Большинство интерпретаций — отрицательные оценки (неодобрение): [*со стороны X-а*] *авантюрно, безответственно, бессердечно, бессовестно, бесстыдно, бестактно, бесчеловечно, бесчестно, глупо, жестоко, зашкварно* (сленг.), *наивно, неблагородно, невежливо, неграмотно, недобросовестно, незаконно, неконструктивно, немилосердно, неосмотрительно, неосторожно, неправильно, непредусмотрительно, неразумно, нерационально, несерьезно, нескромно, несправедливо, нечестно, нечистоплотно, неэтично, низко, преступно, самонадеянно, самоуверенно, тщеславно, целесообразно, цинично* [*делать P*] и др. Положительных оценок (одобрение) существенно меньше: *благородно, дальновидно, правильно, предусмотрительно, разумно, справедливо, человечно, этично* (с *его стороны отказаться от наследства*), ср. также: *грамотно, остроумно, умно, хитро*.

У интерпретаций, как и у других оценок, есть валентность на оцениваемую ситуацию P. Нормальный способ выражения валентности P у оценочных предикативов — зависимая клауза. Однако таких предикативов довольно мало — это универсальные оценки *хорошо* и *плохо* (они могут относиться к самым разным типам ситуаций, которые нравятся или не нравятся субъекту оценки) и их синонимы: *Хорошо* (*здорово, прекрасно, чудесно, замечательно), что вы успели на поезд*;

Плохо (*ужасно, страшно, отвратительно, чудовищно), что уничтожаются леса*.

Что касается интерпретационных оценок, то они существуют, в основном, в форме прилагательных и наречий (причем это особый тип наречий — так называемые сентенциальные, или с плавающей сферой действия, ср. [Филипенко 2003]); они могут сочетаться с глаголом *поступить* или существительным *поступок*: *X поступил благородно/подло/неосмотрительно — X совершил благородный/подлый/неосмотрительный* и т. д. *поступок*.

Что же касается интерпретационных предикативов, то в современном русском языке их больше семи десятков, но их синтаксические возможности по сравнению с универсальными оценками ограничены: они обычно не могут выражать валентность P с помощью зависимой клаузы (?*Малодушно, что ты отказался*). Поэтому инфинитивная конструкция для них единственная возможность эксплицитно выразить валентность P. Интересно, что инфинитивы перцептивных и ментальных глаголов в ИК либо не употребляются, либо обозначают особого рода поступок: *Бестактно* ***слушать*** *частные разговоры*.

Некоторые интерпретации (*преступно*, *бесчестно* и др.) употребляются в ИК, по данным НКРЯ, с XIX в. Другие начали употребляться в этой конструкции в XX в., некоторые — лишь в последние десятилетия. Благодаря корпусу можно проследить, как активизируется ИК у интерпретаций (за недостатком места в приводимой ниже табл. 2 представлена лишь часть материала).

5. Итоги корпусного исследования

В ходе рассмотрения корпусных данных обнаружились следующие закономерности. У предикативов ощущений в ИК не появляется новой валентности. Инфинитив V обозначает обстоятельства, связанные с ощущением, — ту ситуацию, в рамках которой человек взаимодействует (контактирует) с внешней средой или объектом и которая служит своего рода «проводником» воздействия среды (объекта) на человека — на его системы восприятия. Это более сложная концептуализация, которая связывает деятельность человека с пространством, в котором она происходит. В основном предикатив обозначает неблагоприятные, дискомфортные условия для деятельности и сдвигается в зону модальных значений: *темно читать* ≈ ‘невозможно’.

Таблица 2. **Интерпретации**

	1-я пол. XX в.	2-я пол. XX в.	XXI в. Осн. корп.	XXI в. Газ. корп.
безответственно	–	1	3	16
бессердечно	–	3	1	–
бессовестно	3	5	2	1
бесстыдно	–	1	–	–
бестактно	5	11	8	6
благородно	4	8	2	6
возмутительно	3	2	1	1
дальновидно	–	1	–	1
малодушно	2	–	–	1
неблагородно	4	4	2	–
неосмотрительно	1	4	3	2
нескромно	1	6	5	9
несолидно	1	3	7	13
неэтично	1	4	5	35
самонадеянно	–	–	1	7
самоубийственно	1	3	2	6
самоуверенно	2	2	–	3

Инфинитивы, присоединяемые к предикативам эмоций, тоже в большинстве случаев обозначают не семантическую валентность Р, а канал поступления информации о ситуации Р (*грустно видеть/сознавать, что Р*).

И только в группе интерпретаций картина противоположная: инфинитив выражает ситуацию, которую интерпретирует предикатив, т. е. реализует валентность содержания предикатива (*В такой компании как-то несолидно* ***лезть на рожон*** *и* ***безобразничать*** [Р] *на дороге* [РБК Дейли 2011]). Это связано с потребностью выражать валентность содержания синтаксически зависимой формой; интерпретационные слова, будучи морфологическими прилагательными и наречиями, исходно такой возможностью не обладали, и сама инфинитивная

конструкция у многих интерпретационных слов появляется в конце XX — начале XXI вв. Только корпус позволяет оценить масштаб этого явления и проследить его эволюцию в русском языке.

Литература

1. *Апресян Ю. Д.* (2006), Лексикографический тип: глаголы интерпретации. Ю. Д. Апресян и др. (ред.). Языковая картина мира и системная лексикография. М.: Языки славянских культур, с. 145–160.
2. *Кустова Г. И.* (2004), Типы производных значений и механизмы языкового расширения. М.: Языки славянских культур.
3. *Кустова Г. И.* (2013), Системные связи и системные значения экспериенциальных прилагательных: предметные и пространственные контексты. Русский язык в научном освещении. № 1(25), с. 41–69.
4. *Летучий А. Б.* (2018), Предикативы. Материалы к корпусной грамматике русского языка. Вып. III. Части речи и лексико-грамматические классы. СПб: Нестор-История, с. 136–192.
5. *Сердобольская Н. В., Толдова С. Ю.* (2014), Конструкции с оценочными предикативами в русском языке: участники ситуации оценки и семантика оценочного предиката. Acta Linguistica Petrotolitana. Труды Института лингвистических исследований. Т. X. Ч. 2. СПб: Наука, с. 443–477.
6. *Филипенко М. В.* (2003), Семантика наречий и адвербиальных выражений. М.: Азбуковник.
7. *Циммерлинг А. В.* (2017), Русские предикативы в зеркале эксперимента и корпусной грамматики. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 16. Т. 2, с. 466–481.
8. *Циммерлинг А. В.* (2018), Имперсональные конструкции и дативно-предикативные структуры в русском языке. Вопросы языкознания. № 5, с. 7–33.

References

1. *Apresjan Yu. D.* (2006), Leksikograficheskii tip: glagoly interpretatsii [Lexicographic type: verbs of interpretation]. In: Yazykovaya kartina mira i sistemnaya leksikografiya [Linguistic picture of the world and systemic lexicography]. Ed. by Yu. D. Apresjan. Moscow: Yazyki slavyanskikh kul'tur, pp. 145–160.
2. *Filipenko M. V.* (2003), Semantika narechij i adverbial'nyh vyrazhenij [Semantics of adverbs and adverbial phrases]. Moscow: Azbukovnik.
3. *Kustova G. I.* (2004), Tipy proizvodnykh znachenii i mekhanizmy yazykovogo rasshireniya [The Types of Derived Meanings and Language Extension Mechanisms]. Moscow: Yazyki slavyanskikh kul'tur.
4. *Kustova G. I.* (2013), Sistemnye svyazi i sistemnye znacheniya eksperientsial'nykh prilagatel'nykh: predmetnye i prostranstvennye konteksty [Systemic relations and systemic meanings of experiential adjectives: object and spatial contexts]. Russkii yazyk v

nauchnom osveshchenii [Russian language and linguistic theory]. № 1(25), pp. 41–69.
5. *Letuchii A. B.* (2018), Predikativy. Materialy k korpusnoi grammatike russkogo yazyka. Vyp. III. Chasti rechi i leksiko-grammaticheskie klassy [Predicatives. Materials for the corpus grammar of the Russian language. Issue III. Parts of speech and lexical and grammatical classes]. Saint Petersburg: Nestor-Istoriya, pp. 136–192.
6. *Serdobol'skaya N. V., Toldova S. Yu.* (2014), Konstruktsii s otsenochnymi predikativami v russkom yazyke: uchastniki situatsii otsenki i semantika otsenochnogo predikata [Constructions with evaluative predicatives in Russian: participants in the evaluation situation and the semantics of the evaluative predicate]. In: Acta Linguistica Petrotolitana. Trudy Instituta lingvisticheskikh issledovanii [Transactions of the Institute for Linguistic Studies]. Vol. X. Part 2. Saint Petersburg: Nauka, pp. 443–477.
7. *Zimmerling A. V.* (2017), Russkie predikativy v zerkale eksperimenta i korpusnoi grammatiki (транслит) [Russian Predicatives in the Perspective of Sociolinguistic Experiment and Corpus Grammar]. In: Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi mezhdunarodnoi konferentsii «Dialog» (транслит) [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"]. Issue 16. Vol. 2. Moscow, pp. 466–481.
8. *Zimmerling A. V.* (2018), Impersonal'nye konstruktsii i dativno-predikativnye struktury v russkom yazyke (транслит) [Impersonal constructions and dative-predicative structures in Russian]. Voprosy Jazykoznanija [Topics in the study of language перевод на англ]. No. 5, pp. 7–33.

Кустова Галина Ивановна
Институт русского языка им. В. В. Виноградова РАН (Россия)
Kustova Galina
Vinogradov Russian Language Institute of
the Russian Academy of Sciences (Russia)
E-mail: galinak03@gmail.com

А. М. Лаврентьев, Л. А. Курышева

A. M. Lavrentiev, L. A. Kurysheva

ЛИНГВИСТИЧЕСКАЯ ОБРАБОТКА ЦИФРОВЫХ ИЗДАНИЙ РУССКИХ ТЕКСТОВ XVIII ВЕКА

LANGUAGE PROCESSING IN DIGITAL EDITIONS OF RUSSIAN 18th CENTURY TEXTS

Аннотация. В докладе рассмотрены проблемы лингвистической обработки русских текстов XVIII в. на материале двух цифровых корпусов: печатного издания петровского времени «Аль Коран» и рукописной книги 1750-х гг. «Повесть о Лабелле и звере». К лингвистической обработке относятся нормализация орфографии, токенизация, морфологическая разметка и лемматизация. Работа была реализована с помощью предварительной разметки в текстовом редакторе Microsoft Word, конвертации в формат TEI и последующей автоматизированной обработки на платформе TXM, включающей применение TreeTagger и построение многоуровневой транскрипции.
Ключевые слова. Русский язык и литература XVIII в., нормализация орфографии, электронное издание, платформа TXM, разметка TEI XML, лемматизация.

Abstract. This paper deals with the problems of language processing of Russian 18th century texts that occurred in the work on digital editions of the printed translation of *Al'Quran* (1716) and a manuscript translation of *La Belle et la Bête* (*The Beauty and the Beast*, 1758). The linguistic processing includes spelling normalization, tokenization, morphological markup and lemmatization. The work was carried out using manual pre-markup with Microsoft Word, conversion to TEI XML format and further automatic processing on the TXM platform including annotation with TreeTagger and building multi-layer transcription. In *Al'Quaran* edition the spelling normalization is fully automated but only the simplest cases are dealt with, while in *La Belle et la Bête* manual pre-markup allows generating modern form for all words.
Keywords. Russian language and literature of the 18th century, spelling normalization, digital edition, TXM platform, TEI XML markup, lemmatization.

1. Проблемы создания исторических корпусов на примере русского языка XVIII в.

Для исторической лингвистики наличие качественно размеченных представительных корпусов различных этапов истории языка жизненно необходимо, так как исследователи не могут полагаться на собственную интуицию или языковой эксперимент. При этом распространенные инструменты цифровизации и лингвистической разметки не всегда адаптированы к орфографии, пунктуации и графической сегментации исторических текстов, что приводит к необходимости достаточно больших затрат времени и ресурсов для их «ручной» об-

работки. Этим объясняется относительно малый объем исторической части многих национальных корпусов. Так, в Национальном корпусе русского языка (НКРЯ) объем подкорпуса XVIII в. составляет приблизительно 6,36 млн с/у, т. е. менее 2 % от общего объема основного корпуса (по данным на март 2021 года).

С. О. Савчук и Д. В. Сичинава подробно рассматривают принципы отбора и методику обработки текстов XVIII в. для НКРЯ [Савчук и др. 2009]. Частью этой методики является «умеренная модернизация» орфографии, принятая в большинстве академических изданий. Авторы отмечают, что практика модернизации, применяемая в различных изданиях, непостоянна и не всегда последовательна.

Поиск в НКРЯ показывает, что в нем присутствуют тексты как с модернизированной, так и с оригинальной орфографией, что создает определенные неудобства для пользователей. Модернизированная орфография облегчает поиск и лингвистическую разметку, однако оригинальная орфография источников необходима для исследований в области истории морфологии, орфографии и пунктуации.

В настоящем докладе мы представляем опыт создания лингвистически размеченных цифровых изданий памятников XVIII в., в которых проблема совмещения модернизированной и оригинальной орфографии решается с помощью автоматической обработки транскрипций, в которых имеется возможность предварительной ручной разметки отдельных словоформ. Использованное программное обеспечение и сами цифровые издания распространяются свободно на условиях открытых лицензий.

Материалом для настоящего доклада послужили два проекта, непосредственно не связанные между собой. Их объединяет относительное сходство языкового материала и примененная технологическая цепочка создания и обработки цифрового издания.

2. Русский перевод «Корана» (1716 г.)

Цифровое издание первого русского перевода «Корана» является составной частью международного проекта Coran 12– 21[2], направленного на создание параллельного корпуса европейских переводов этой священной книги. В настоящее время корпус включает два издания на арабском, а также различные переводы на латинский, французский и итальянский языки, впервые опубликованные в XVI–XX в.

[2] https://coran12-21.org.

Первый полный перевод «Корана» на русский язык был осуществлен не с арабского языка, а с французского (издание Дю Рие, 1647) и был напечатан «повелением царского величества» в Санкт-Петербургской типографии в 1716 г. Согласно общепринятой вплоть до конца XX в. точке зрения, автором перевода был П. В. Постников [Быкова и др. 1955], однако более поздние исследования показали маловероятность данной атрибуции [Запольская 2002].

Электронная транскрипция текста была проведена вручную Н. В. Луговской с экземпляра, хранящегося в РГБ (шифр МК Си-2°/16-К). К сожалению, ресурсы проекта не позволили оплатить оцифровку источника. Транскрипция в максимальной степени точно воспроизводит орфографию и пунктуацию источника, явные опечатки корректируются в сносках. Дополнительно используются стили заголовков для отражения структуры документа (деление на суры) и специальные коды для указания номеров страниц и стихов.

Текст транскрипции был автоматически переведен из формата .docx в TEI XML с использованием сервиса Oxgarage[3]. Дальнейшая обработка производилась автоматически на платформе TXM [Heiden, 2010] с использованием модуля импорта XML TEI Zero и пакета скриптов XSLT, позволяющих привести структурную разметку в соответствие с рекомендациями TEI и добавить слой частично нормализованной орфографии.

Нормализация затрагивает упраздненные в результате реформы орфографии 1918 года буквы и конечный ъ. Как уже отмечалось, такая практика достаточно распространена в академических изданиях текстов XVIII в., однако в электронном издании нормализованная графика не заменяет, а дополняет оригинальную.

Помимо повышения комфорта чтения для неспециалистов нормализация орфографии позволяет использовать для автоматической морфологической разметки и лемматизации лингвистическую модель современного русского языка, созданную для программы TreeTagger [Schmid 1994]. Хотя далеко не все словоформы успешно «осовременниваются» с помощью использованного простейшего алгоритма, частеречная разметка оказывается правильной для 90 %, а лемматизация — для 83 % словоупотреблений (подсчет проведен на фрагменте объемом 500 словоупотреблений в начале текста).

[3] https://oxgarage.tei-c.org.

3. «Повесть о Лабелле и звере» (перевод 1758 г.)

Опыт более глубокой нормализации орфографии, позволяющей привести к современной форме подавляющее большинство словоупотреблений, был проделан на материале другого памятника, рукописного перевода на русский язык французской сказки La Belle et la Bête (заглавие в рукописи — «Повесть о Лабелле и звере»). Перевод был сделан и оформлен в виде подарочной рукописной книги четырнадцатилетней девочкой Хионией Демидовой для своего брата в 1758 г. (Зональная научная библиотека имени В. А. Артисевич Саратовского госуниверситета, рукопись ОРКР № 456). Оригиналом послужила сказка французской писательницы Мари Лепренс де Бомон (Marie Leprince de Beaumont), вошедшая в ее учебник для девочек Magasin des enfants (1756). Данный перевод представляет интерес как источник сведений о домашнем образовании и обучении иностранным языкам в России XVIII в., а также о том, как инокультурные реалии воспринимались подростком того времени. Подробное описание источника и проекта цифрового издания представлены в [Курышева и др. 2019], электронное издание доступно на портале лаборатории IHRIM[4].

Небольшой объем памятника позволил в сравнительно короткие сроки подготовить пилотное электронное издание с критическим аппаратом и более качественной лингвистической разметкой, чем в случае «Корана».

При этом была использована та же технологическая цепочка, что и в проекте цифрового издания «Корана»: предварительная разметка в Microsoft Word, преобразование в TEI XML и импортирование в TXM с применением пакета скриптов XSLT и автоматической разметки TreeTagger.

В случаях, когда простая замена букв и удаление конечного ъ не позволяли получить современную словоформу, соответствующая форма набиралась вручную в дополнение к форме оригинала. Это позволило создать три слоя графической формы слов: оригинальный, нормализованный и модернизированный. Модернизированный слой скрыт от читателей, однако именно он используется для лингвистической разметки и позволяет использовать современную орфографию в запросах для построения конкордансов и частотных словарей.

[4] https://txm-ihrim.huma-num.fr/txm/?command=Documentation&path=/LA-BELLE (дата доступа: 01.07.2021).

Отдельную проблему составляет сегментация текста на словоупотребления. В рукописи встречается немало примеров, когда употребление пробелов не соответствует принятой в современном русском языке норме. Например, *съ начала* (наречие ‘сначала’), *невидя*. В этих случаях простые коды из специальных символов были использованы с тем, чтобы обеспечить отображение графики источника в дипломатическом слое и нормализовать сегментацию в остальных слоях.

4. Заключение

Процесс «ручной» модернизации достаточно трудоемок и не может быть масштабирован на крупные корпусы или коллекции текстов. Основная задача издания «Повести о Лабелле и звере» — служить прототипом для создания пользовательского интерфейса и выработки методики анализа многослойных транскрипций.

Простейший алгоритм, примененный в издании «Корана», показал свою эффективность и в то же время оказался недостаточным для обеспечения высококачественной автоматической аннотации частей речи и лемматизации.

Совершенствование алгоритма автоматической модернизации орфографии может в дальнейшем существенно сократить необходимость «ручной предразметки» словоформ, которая потребуется только в случаях ошибочных или нестандартных написаний.

Мы надеемся, что предложенная методика транскрипции и разработанная технологическая цепочка смогут быть использованы в новых проектах цифровых изданий и корпусов текстов на русском языке в старой орфографии.

Цифровые издания имеют ряд преимуществ, хотя и не отменяют ценности печатных академических изданий памятников. Прежде всего, это доступность и богатство предоставляемого материала и инструментария, которые позволяют решать самые разнообразные задачи и отвечать требованиям различных категорий читателей — от узких специалистов до школьников и широкой публики. Кроме того, подобные издания открывают широкие возможности для новых исследовательских и педагогических проектов.

Литература

1. *Быкова Т. А., Гуревич М. М.* (1955), Описание изданий гражданской печати: 1708 — январь 1725 г. М.-Л.
2. *Запольская Н. Н.* (2002), Культурно-языковой статус личности и текста в петровскую эпоху (опыт прогнозирующего анализа). Славянская языковая и этноязыковая системы в контакте с неславянским окружением. М., с. 422–447.
3. *Курышева Л. А., Лаврентьев А. М.* (2019), Об электронном издании рукописной «Повести о Лабеле и звере» (1758): первый русский перевод сказки «Красавица и зверь» на демонстрационном портале платформы TXM. Сибирский филологический журнал. № 1, с. 54–61. DOI: 10.17223/18137083/66/4
4. *Савчук С. О, Сичинава Д. В.* (2009), Корпус русских текстов XVIII века в составе НКРЯ: проблемы и перспективы. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб, с. 52–70.
5. *Heiden S.* (2010), The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In: Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Nov. 2010. Sendai, Japan, Institute for Digital Enhancement of Cognitive Development. Waseda University, pp. 389–398. URL: halshs.archives-ouvertes.fr/halshs-00549764 (дата обращения: 20.05.2021).
6. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. URL: www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (дата обращения: 20.05.2021).

References

7. *Bykova T. A., Gurevich M. M.* (1955), Opisanie izdanij grazhdanskoj pechati: 1708 — janvar' 1725 g. [Description of Civil Press Publications: 1708 — January 1725]. Moscow-Leningrad.
8. *Heiden S.* (2010), The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In: Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation. Nov. 2010. Sendai, Japan, Institute for Digital Enhancement of Cognitive Development. Waseda University, pp. 389–398. URL: halshs.archives-ouvertes.fr/halshs-00549764 (date of access: 20.05.2021).
9. *Kurysheva L. A., Lavrentiev A. M.* (2019), Ob elektronnom izdanii rukopisnoj "Povesti o Labele i zvere" (1758): pervyj russkij perevod skazki "Krasavica i zver'" na demonstracionnom portale platformy TXM [The Story of Labelle and the Beast": a Digital Edition of a Manuscript the First Russian Translation of the "Beauty and the Beast" Fairy-Tale Powered by the TXM Demo Portal]. In: Sibirskij filologicheskij zhurnal [Siberian Journal of Philology]. No. 1, pp. 54–61. DOI: 10.17223/18137083/66/4
10. *Savchuk S. O, Sichinava D. V.* (2009), Korpus russkih tekstov XVIII veka v sostave NKRJa: problemy i perspektivy [A Corpus of 18th Century Russian Texts in the Framework of the RNC: Problems and Prospectives]. In: Nacional'nyj korpus russkogo jazy-

ka: 2006–2008. Novye rezul'taty i perspektivy [Russian National Corpus: 2006–2008. New Results and Prospectives]. Saint Petersburg, pp. 52–70.
11. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. URL: www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf (date of access: 20.05.2021).
12. *Zapol'skaja N. N.* (2002), Kul'turno-jazykovoj status lichnosti i teksta v Petrovskuju epohu (opyt prognoziruemogo analiza) [Cultural and Linguistic Status of Personality and Text at the Time of Peter the Great (An Experience of Predictive Analysis)]. In: Slavjansakja jazykovaja i etnojazykovaja sistemy v kontakte s neslavjanskim okruzheniem [Slavic Linguistic and Ethno-Linguistic Systems in Contact with Non-Slavic Environment]. Moscow, pp. 422–447.

Лаврентьев Алексей Михайлович
Институт истории представлений и идей нового времени НЦНИ и Высшей нормальной школы Лиона (Франция)
Lavrentiev Alexei
Institute of History of Representations and Ideas in the Modernity, National Centre for Scientific Research and Ecole Normale Supérieure de Lyon (France)
E-mail: alexei.lavrentev@ens-lyon.fr

Курышева Любовь Александровна
Институт филологии Сибирского отделения РАН (Россия)
Kurysheva Liubov
Institute of philology, Siberian Branch of the Russian Academy of Sciences (Russia)
E-mail: luba.kurysh@gmail.com

А. Н. Лапошина, Т. С. Веселовская, Л. Ю. Жильцова, О. Ф. Купрещенко, М. Ю. Лебедева

A. N. Laposhina, T. S. Veselovskaya, L. Yu. Zhiltsova, O. F. Kupreshshenko, M. Yu. Lebedeva

КОРПУСНОЕ УЧЕБНИКОВЕДЕНИЕ: В ПОИСКАХ ОБЪЕКТИВНЫХ КРИТЕРИЕВ ОЦЕНКИ УРОВНЯ УЧЕБНИКОВ ДЛЯ БИЛИНГВОВ[1]

CORPUS-BASED ANALYSIS OF RUSSIAN TEXTBOOK FOR BILINGUAL CHILDREN: TOWARDS DATA-DRIVEN TEXTBOOK EVALUATION

Аннотация. В статье предлагается способ объективизации информации об уровне сложности учебников русского языка для детей 7–10 лет со вторым родным русским (РК2Р) с помощью количественного анализа данных корпуса учебников русского языка TIRTEC. Работа содержит сравнительный анализ лексических и синтаксических показателей учебных текстов для детей РК2Р с текстами для детей, изучающих русский как родной и как иностранный.
Ключевые слова. Корпус учебников, корпусное учебниковедение, русский язык как иностранный, русский как второй родной.

Abstract. The article is devoted to the discussion of the way to obtain objective information about the level of difficulty of Russian language textbooks for children 7–10 years old with second native Russian (RSNL) by means of quantitative analysis of TIRTEC Russian language textbook corpus data. This paper contains a comparative analysis of the lexical and syntactic indicators of the educational texts for RSNL children with the texts for children learning Russian as a native and as a foreign language. The results of this analysis allow us to determine the place of each textbook on the scale of progressive complexity of the language material and can be used to clarify information about the target audience of the textbook.
Keywords. Corpus of textbooks, textbook analysis, Russian as a foreign language, Russian as a second native language.

1. Введение

Сбор репрезентативной коллекции языковых учебников и ее последующий количественный анализ способен дать ответы на целый ряд исследовательских вопросов. В частности, такие коллекции могут служить для изучения лексического состава учебника или полноты описания определенных методических категорий [Gabrielatos 2006], поиска формальных показателей, указывающих на сложность учеб-

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

ного текста [Laposhina et al. 2018; Reynolds 2016], исследования культурной информации и стереотипов, транслирующихся посредством учебных материалов [Al Jumiah 2016; Sun, Kwon 2020].

С развитием технологий сбора и обработки текстовых данных все чаще подобные исследования выполняются в русле корпусного учебниковедения (англ. corpus-based textbook analysis): создаются специальные корпусы учебников, разрабатывается методика педагогической разметки таких корпусов [Volodina 2014].

Задача классификации учебных пособий по уровню языковой сложности крайне важна при обучении языку, поскольку позволяет подобрать оптимальное пособие для учащегося исходя из его языковой подготовки, возраста, целей и пр. Одной из самых проблемных зон этой области методики преподавания РКИ можно назвать оценку уровня пособий для детей, для которых русский язык является не единственным языком повседневной коммуникации, из-за разнородности этой группы и отсутствия регламентирующих документов и стандартов для этой категории учащихся.

В настоящей работе в качестве одного из решений этой проблемы предлагается система оценивания языкового учебного материала по уровню сложности на примере анализа учебников по русскому языку для детей младшего школьного возраста с разным уровнем владения русским языком.

2. Материалы и методы

Для данного исследования была отобрана коллекция из 18 пособий по русскому языку из корпуса TIRTEC (Text-Image Russian Textbook Corpus). Корпус содержит тексты учебников русского языка для детей младшего школьного возраста с разным языковым опытом, от учебников русского как иностранного до учебников, использующихся в российской школе, снабженные педагогической разметкой, которая включает информацию об авторах и предполагаемой целевой аудитории учебника, типе текстового блока (например, *текст, упражнение, справочная информация*) и мн. др. [Лапошина и др. 2019].

Поскольку в фокусе внимания этой статьи находится группа пособий для детей со вторым родным русским (РК2Р), мы отобрали 5 пособий для данной группы учащихся (далее в таблицах обозначаются индексом 2), 3 пособия по русскому как иностранному (индекс 1), линейку из учебников русского языка для 1–3 классов русских школ

за рубежом (индекс 3) и линейку из учебников русского языка для 1–3 классов российских школ (индекс 4).

Все приведенные ниже расчеты производились с помощью программного кода на языке Python и морфологического анализатора pymystem3[2].

3. Анализ лексических показателей уровня сложности учебных материалов

Для сравнительного анализа лексического состава выбранных пособий мы произвели расчет общего объема учебника в словах, количества уникальных лексем, представленных в учебнике, а также долю слов, входящих в лексические списки, которые могут быть релевантны в данном случае.

Одним из наиболее разработанных показателей доступности лексики текстов, предназначенных для изучающих язык как иностранный, является процент лексики из лексических минимумов [Karpov et al. 2014; Reynolds 2016]. Поскольку к настоящему моменту не существует специальных лексических списков для детей, изучающих русский язык, мы использовали в качестве ориентира Лексический минимум для взрослых, начинающих изучать русский язык как иностранный (далее — ЛМ А1)[3], который содержит лексические единицы элементарного уровня — например, *город, рука, бутерброд*.

Еще одним показателем доступности лексики учебных текстов традиционно считается частотность [Chen, Meurers 2016]. Для этих подсчетов мы выбрали список из 5 тысяч самых частотных слов русского языка Нового частотного словаря русской лексики (далее — ЧС 5000)[4], однако стоит отметить, что данный список может быть не вполне релевантен для задачи оценки детской литературы — например, в первую тысячу частотных слов входят такие лексемы, как *суд, советский* и пр. Для получения информации о частотности слова по коллекции текстов, более близкой детской аудитории, мы использовали список из 5 тысяч самых частотных слов на материале корпуса детской ли-

[2] https://github.com/nlpub/pymystem3

[3] Андрюшина Н. П., Козлова Т. В. Лексический минимум по русскому языку как иностранному. Элементарный уровень. СПб, 2012.

[4] Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

тературы Деткорпус[5] (далее — Деткорпус 5000). В расчетах использовались все фрагменты учебников (тексты, упражнения, справочная информация и пр.), за исключением текстов, обращенных к учителю или родителю.

Таблица 1. **Лексический состав учебников русского языка для разных групп учащихся[6]**

Индекс	Пособие	ЛМ А1	ЧС 5000	Деткорпус 5000	Всего слов	Уникальных лексем
1	Сорока	83	92	95	5931	743
1	Шаг за шагом	80	90	89	5852	741
2	Лена и Миша	70	89	93	16 645	1603
2	Истоки	67	87	93	22 271	2631
2	У костра	62	84	88	12 599	2600
3	Дронов 3	60	83	88	14 567	1930
3	Дронов 2	58	82	88	13 886	1975
3	Дронов 1	57	81	88	13 938	2093
2	Совенок	57	78	87	6427	1042
1	Говорю-пишу	57	74	82	7299	1480
4	Канакина 1	53	78	84	11 861	1855
4	Канакина 2	51	77	82	28 304	3390
4	Канакина 3	48	75	81	38 334	4542
2	Крокодил	39	63	70	5771	1940

Табл. 1 содержит информацию о лексическом составе четырех групп анализируемых учебников. Отсортированная по убыванию процента лексики из ЛМ А1, гипотетически она должна показывать плавный рост индекса группы учебников (от 1 — русский как иностранный к 4 — русский как родной) по мере уменьшения процента

[5] detcorpus.ru.

[6] В тексте работы приводятся сокращенные авторские обозначения учебников и пособий, цифра соответствует классу. Полный список и расшифровка указаний доступны по ссылке: https://digitalpushkin.tilda.ws/istochniki

лексики из списка элементарного уровня. Однако в реальности мы наблюдаем более сложную картину. Некоторые пособия оказываются ожидаемо простыми по этому параметру (Сорока, Шаг за шагом), тогда как другие заметно сближаются с учебниками для российской школы (Говорю-пишу) или даже оказываются сложнее их (Крокодил). Если линейка учебников для российских школ (Канакина) выстраивается в логичную цепочку уменьшения процента элементарной лексики от 1 к 3 классу, то линейка для зарубежных школ (Дронов) не столь последовательна: минимальные показатели демонстрируют учебники 1 и 2 класса.

Несмотря на разный объем и состав лексических списков (объем ЛМ А1 составляет 900 лексем; лексика ЧС 5000 и Деткорпуса 5000 пересекается в 60 % случаях), результаты подсчетов по спискам достаточно сильно коррелируют друг с другом: сортировка таблицы по убыванию процента лексики из двух других списков практически не меняет порядок учебников на получившейся шкале. Для большей точности подсчетов в будущих исследованиях будет выбран какой-то один список или предложен скомпилированный из них единый список.

Информативным показателем целевой группы учебника может также стать общее количество слов и количество уникальной лексики учебника. Здесь на фоне стремительного роста объема лексики от 1 к 3 классу учебника Канакиной выделяется стабильность этого параметра по всей аналогичной линейке Дронова.

Отношение общего количества слов к количеству уникальной лексики иллюстрирует повторяемость, отработанность лексических единиц. По этому параметру нижние строчки занимают пособия для РК2Р, «Крокодил» и «У костра».

4. Анализ синтаксических показателей

Для того, чтобы проиллюстрировать возможные отличия пособий на синтаксическом уровне, приведем несколько характеристик, традиционно используемых для задачи измерения сложности и связности текста [Graesser et al. 2014; Reynolds 2016]. Формула читабельности Флеша представлена в адаптированном варианте для русского языка [Оборнева 2006]. В расчетах использовались только законченные фрагменты текстов, исключались грамматические упражнения, списки лексики и пр.

Таблица 2. **Показатели синтаксической сложности учебников русского языка для разных групп учащихся**

Индекс	Пособие	Медианное значение формулы Оборневой	Средняя длина предложения	Среднее кол-во подчинит. союзов на предложение
1	Сорока	94	4	0,2
2	Совенок	93	3,3	0,3
1	Шаг за шагом	91	3,8	0,3
1	Говорю-пишу	88	5	0,2
3	Дронов 1	81	5,7	0,4
3	Дронов 2	79	6,8	0,5
2	У костра	77	6	0,6
4	Канакина 1	75	6,6	0,5
3	Дронов 3	74	7,2	0,6
2	Истоки	74	9,7	1
2	Лена и Миша	74	9,1	1
4	Канакина 3	72	7	0,6
4	Канакина 2	71	7,3	0,7
2	Крокодил	59	11,2	0,8

Табл. 2 демонстрирует шкалу учебников, выстроенную по убыванию индекса читабельности Флеша, то есть по возрастанию уровня сложности. Несмотря на то, что все учебники, кроме пособия «Крокодил», попадают в категорию трактовки *65–100 баллов — очень легко читается* [Оборнева 2006], что достаточно закономерно для учебников для детей 7–10 лет, наглядным становится их распределение внутри этой группы. Верхние строчки занимают все 3 учебника для детей-иностранцев. Подтверждается замеченная ранее в табл. 1 существенная разница в уровне сложности учебников 1 и 2 класса линейки для российских школ Канакиной. Выявлены учебники, по нескольким параметрам структурной сложности превосходящие линейку для российских школ: «Истоки», «Лена и Миша». Пособие «Крокодил», оказавшись самым сложным по данным обеих таблиц, требует более детального анализа и, вероятнее всего, корректировки методического описания предполагаемой аудитории учебника.

5. Выводы

Таким образом, одним из подходов к решению методической проблемы отсутствия единой системы маркировки и классификации учебников русского языка как второго родного (РК2Р) может стать сравнительный анализ коллекции текстов таких учебников и их сопоставления с двумя методическими «полюсами», между которыми они находятся: учебников русского как родного и русского как иностранного. Подобное сравнение позволяет отобрать лингвистические показатели, которые могут стать основой составления протокола оценивания и маркировки учебников для билингвов.

Литература

1. *Лапошина А. Н., Веселовская Т. С., Купрещенко О. Ф.* (2019), Иллюстративно-текстовый корпус учебников русского языка для детей младшего школьного возраста: концепция и методика создания. Труды международной конференции «Корпусная лингвистика-2019», с. 63–71.
2. *Оборнева И. В.* (2006), Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис… канд. пед. наук. М., 120 с.
3. *Al Jumiah A.* (2016), Language, Power, and Ideology in High School EFL Textbooks in Saudi Arabia (Doctoral of Philosophy). University of New Mexico, Albuquerque, New Mexico.
4. *Chen X., Meurers D.* (2016), Characterizing text difficulty with word frequencies. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84–94.
5. *Gabrielatos C.* (2006), Corpus-based evaluation of pedagogical materials: If-conditionals in ELT coursebooks and the BNC. Paper presented at 7th Teaching and Language Corpora Conference, Paris, France.
6. *Graesser A. C., McNamara D. S., Cai Z., Conley M., Li H., Pennebaker J.* (2014), Coh-Metrix measures text characteristics at multiple levels of language and discourse. In: The Elementary School Journal. No. 115(2), pp. 210–229.
7. *Karpov N., Baranova J., Vitugin F.* (2014), Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST), pp. 91–100.
8. *Reynolds R.* (2016), Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 289–300.
9. *Sun W., Kwon J.* (2020), Representation of monoculturalism in Chinese and Korean heritage language textbooks for immigrant children. In: Language Culture and Curriculum. No. 33, pp. 402–416.
10. *Volodina E., Pilán I., Eide S. R., Heidarsson H.* (2014), You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In

Proceedings of the third workshop on NLP for computer-assisted language learning, pp. 128–144.

References

1. *Al Jumiah A.* (2016), Language, Power, and Ideology in High School EFL Textbooks in Saudi Arabia (Doctoral of Philosophy). University of New Mexico, Albuquerque, New Mexico.
2. *Chen X., Meurers D.* (2016), Characterizing text difficulty with word frequencies. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84–94.
3. *Gabrielatos C.* (2006), Corpus-based evaluation of pedagogical materials: If-conditionals in ELT coursebooks and the BNC. Paper presented at 7th Teaching and Language Corpora Conference, Paris, France.
4. *Graesser A. C., McNamara D. S., Cai Z., Conley M., Li H., Pennebaker J.* (2014), Coh-Metrix measures text characteristics at multiple levels of language and discourse. In: The Elementary School Journal. No. 115(2), pp. 210–229.
5. *Karpov N., Baranova J., Vitugin F.* (2014), Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST), pp. 91–100.
6. *Laposhina A. N., Veselovskaya T. S., Kupreshchenko O. F.* (2019), Illyustrativno-tekstovyj korpus uchebnikov russkogo yazyka dlya detej mladshego shkol'nogo vozrasta: koncepciya i metodika sozdaniya [Text-image corpus of Russian language textbooks for primary school: concept and method of creation]. In: Trudy mezhdunarodnoj konferencii "Korpusnaya lingvistika-2019" [Proceedings of the International Conference "Corpus Linguistics-2019"]. Saint Petersburg, pp. 63–71.
7. *Oborneva I. V.* (2006), Avtomatizirovannaya ocenka slozhnosti uchebnyh tekstov na osnove statisticheskih parametrov [Automatic assessment of the complexity of educational texts on the basis of statistical parameters]. Moscow.
8. *Reynolds R.* (2016), Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 289–300.
9. *Sun W., Kwon J.* (2020), Representation of monoculturalism in Chinese and Korean heritage language textbooks for immigrant children. In: Language Culture and Curriculum. No. 33, pp. 402–416.
10. *Volodina E., Pilán I., Eide S. R., Heidarsson, H.* (2014), You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In Proceedings of the third workshop on NLP for computer-assisted language learning, pp. 128–144.

Лапошина Антонина Николаевна
Государственный институт русского языка им. А. С. Пушкина (Россия)

Laposhina Antonina
Pushkin State Russian Language Institute (Russia)
E-mail: antonina.laposhina@gmail.com

Веселовская Татьяна Сергеевна
Государственный институт русского языка им. А. С. Пушкина (Россия)
Veselovskaya Tatyana
Pushkin State Russian Language Institute (Russia)
E-mail: veselovskayats@gmail.com

Жильцова Людмила Юрьевна
Государственный институт русского языка им. А. С. Пушкина (Россия)
Zhiltsova Lyudmila
Pushkin State Russian Language Institute (Russia)
E-mail: jiltsova.ludmila@gmail.com

Купрещенко Ольга Федоровна
Государственный институт русского языка им. А. С. Пушкина (Россия)
Kupreshchenko Olga
Pushkin State Russian Language Institute (Russia)
E-mail: ofkupr@gmail.com

Лебедева Мария Юрьевна
Государственный институт русского языка им. А. С. Пушкина (Россия)
Lebedeva Maria
Pushkin State Russian Language Institute (Russia)
E-mail: m.u.lebedeva@gmail.com

С. А. Мелещева, П. М. Эйсмонт

S. A. Melescheva, P. M. Eismont

ЭЛЕМЕНТЫ НЕВЕРБАЛЬНОГО ПОВЕДЕНИЯ В ПРОЦЕССЕ ПОРОЖДЕНИЯ НЕПОДГОТОВЛЕННОГО НАРРАТИВА РУССКОЯЗЫЧНЫМИ ДЕТЬМИ (КОРПУСНОЕ ИССЛЕДОВАНИЕ)[1]

ELEMENTS OF NON-VERBAL BEHAVIOR IN THE PROCESS OF NARRATIVE PRODUCTION BY RUSSIAN-NATIVE CHILDREN (A CORPUS STUDY)

Аннотация. На примере анализа тридцати неподготовленных устных нарративов русскоязычных детей в возрасте от 4 лет 7 месяцев до 7 лет 6 месяцев из корпуса «Кондуит» представлена классификация типов используемых детьми жестов. Наличие, интенсивность и типы жестикуляции связаны с уровнем языкового развития ребенка (объемом нарратива, типами используемых дискурсивных актов), однако возрастную разницу выявить не удалось.
Ключевые слова. Детская речь, нарратив, корпус «Кондуит», невербальное поведение, жесты.

Abstract. The paper deals with the analysis of a set of unprepared elicited oral narratives by Russian-native children at the age of 4 years 7 months old to 7 years 6 months old from the "Conduit" corpus and focuses on the classification of gestures used by children. Seven types of gestures performed by the narrators have been described, most of which imitate the gestures of the original cartoon characters. The analysis shows that both the types of gestures and their intensity, but not the age of the speaker correlate with the level of language development (the size of the narrative, the types of discourse acts).
Keywords. Language acquisition, narrative, "Conduit" corpus, non-verbal behavior, gestures.

Большинство исследований невербального поведения, особенно жестов, относится к области жестово-речевого взаимодействия и касается их общих и отличительных черт, а также их вклада в общение и передачу информации. Большинство исследователей считают, что в рамках коммуникации вербальный и невербальный компоненты составляют единое целое, что способствует их успешной интерпретации [Goldin-Meadow 2007]. Жест и язык действуют как равноправные партнеры, и говорящий может перераспределить их значимость даже в пределах одного высказывания, выделив тот или иной компонент.

В то же время многие исследователи отмечают, что роль жестов является вспомогательной, поскольку они могут как уточнить содержа-

[1] Статья подготовлена при финансовой поддержке РФФИ (проект № 20-012-00290 «Устный и письменный нарратив как вторичный текст: особенности порождения разными категориями носителей русского языка»).

ние сопровождающего высказывания, облегчить его восприятие или что-то подчеркнуть [Cienki 2019], так и изменить буквальное содержание вербального компонента, выдав неявное намерение говорящего [Kelly et al. 1999]. Тем не менее, в коммуникации речь преобладает, а количество информации, передаваемой жестами, очевидно, весьма незначительно. По мнению Краусса с коллегами [Krauss et al. 1996], информация, заключенная в жесте, может быть настолько незначительной, что в отсутствие речи она вообще бессмысленна, а при естественном речевом общении воспроизводит значение сопутствующих лексических единиц.

Невербальное поведение играет важную роль в формировании социальных навыков у детей и в развитии их речи. Сам процесс овладения языком играет важную роль в восприятии невербальной информации. Мультимодальность общения приводит к тому, что дети младшего возраста испытывают трудности при распределении внимания между вербальным и невербальным компонентами, обычно предпочитая один или другой. Дети постепенно осваивают жесты либо на основании их роли в социальных взаимодействиях, либо благодаря получению желаемой реакции. На определенном этапе развития дети обращают гораздо больше внимания на словесную информацию, чем на визуальную [Эйсмонт 2008]. По мнению многих авторов, дети используют жестикуляцию для компенсации языкового дефицита, поскольку невербальная составляющая коммуникации может представлять довольно сложную информацию в более простой форме, и на определенном уровне развития языка дети усваивают ее легче, чем речь [Novack et al. 2016]. Тем не менее, исследования Колетта с соавторами [Coletta et al. 2010] опровергают эту идею на основе анализа такого сложного речевого продукта, как спонтанный нарратив, и показывают, что при порождении неподготовленного нарратива невербальный и вербальный компоненты развиваются параллельно.

Материалом нашего исследования послужили видеозаписи детей, полученные при проведении эксперимента с целью изучения процесса порождения неподготовленных устных нарративов по методике извлеченных текстов. Перед участниками эксперимента была поставлена задача посмотреть четырехминутный фрагмент мультфильма «Как стать большим» (реж. В. Дегтярев, Союзмультфильм, 1967) в беззвучном режиме и во время просмотра рассказать о событиях, происходящих в нем. Каждый ребенок опрашивался индивидуально, ход эксперимента фиксировался на аудио- и видеоносители. Ана-

лизируемые нарративы представлены в Корпусе неподготовленных детских устных текстов «Кондуит» [Эйсмонт 2017]. Для анализа случайным образом были отобраны записи 30 детей из трех возрастных групп: в первую группу входят 10 детей возрастом от 4 лет 7 месяцев до 5 лет 6 месяцев (4 мальчика и 6 девочек), во второй группе также 10 детей в возрасте от 5 лет 7 месяцев до 6 лет 6 месяцев (3 мальчика и 7 девочек), в третьей группе — 10 детей возрастом от 6 лет 7 месяцев до 7 лет 6 месяцев (3 мальчика и 7 девочек). Все дети посещают дошкольные образовательные учреждения Санкт-Петербурга. Примерная продолжительность видеозаписи с каждым ребенком составляет 4 минуты 50 секунд. Разметка и аннотация данных осуществлялась с использованием программного обеспечения ELAN (http://www.mpi.nl/tools/elan/) по шаблону, разработанному для разметки Русского эмоционального корпуса версии 2.2 [Kotov, Budyanskaya 2012], в котором содержатся слои с информацией о тексте (содержание, структура, цель), эмоциональном состоянии участника, его жестах и мимике (см. рис. 1).

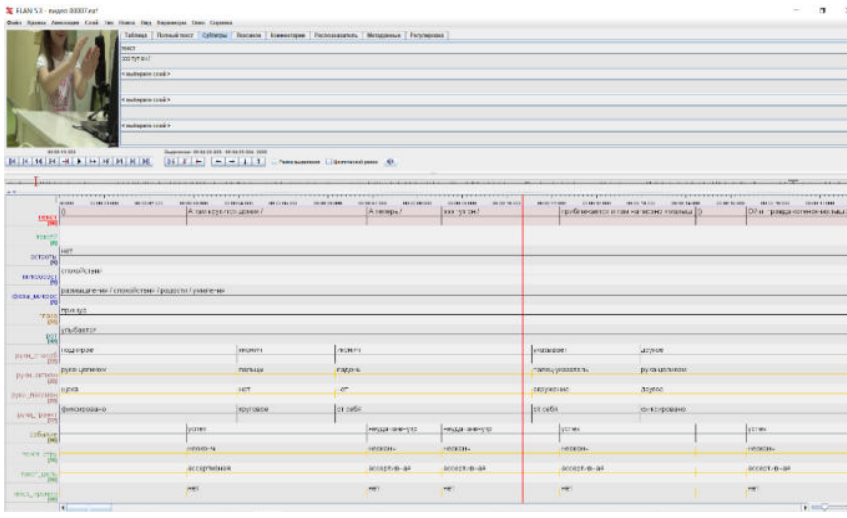


Рис. 1. **Пример разметки**

Жесты, производимые участниками, были описаны и распределены по группам, в зависимости от их формы и функции. Мы выделили 7 типов и подтипов жестов:

- дейктические, или указательные: участник рукой или головой указывает на увиденный в мультфильме объект;
    - перечисление: участник по очереди показывает на воображаемые объекты или считает, загибает пальцы и т. д.;
- иконические: участник с помощью рук воспроизводит образы и формы объектов, имитирует траектории движений персонажей;
    - закрытие: участник прикрывает рукой объект, глаза, рот и т. п.;
- манипулирование: участник манипулирует каким-либо объектом, совершает с ним любые действия;
- опорные жесты: участник подпирает руками свое тело;
- скрещивание: участник скрещивает пальцы, руки и т. п.

Многие жесты представляли собой имитацию жестов персонажей. Исследования Бейтса и Дика [Bates, Dick 2002] показали, что на ранних стадиях развития дети склонны не столько распознавать представленную невербальную информацию, сколько подражать ей, что подтверждается и нашим материалом. Также были такие жесты, как «чешет», «трет», «кусает», которые являются жестами-адаптерами и передают внутреннее волнение и неудобство говорящего.

На отдельных слоях были размечены движения глаз и губ, такие как «взгляд вбок», «поднимает брови», «улыбается», «облизывается» и т. д. С помощью этих слоев мы отмечали имитирование мимики персонажей.

Текст нарратива был расположен на отдельном слое. Для анализа объема текста и уровня речевой беглости ребенка использовался показатель MLU, рассчитанный путем деления количества слов на количество высказываний, произнесенных ребенком во время нарратива. Более высокий уровень MLU означает более высокий уровень владения языком [Rice 2010]. Данный показатель необходим для сопоставления дискурсивных способностей у детей, использующих различные типы жестов, и у детей, избегающих жестикуляции, в том числе и на прагматическом уровне, которая включает в себя изучение структуры дискурса. Независимо от того, влияет ли на рассказ собственный опыт говорящего, его грамотность или общие культурные знания, нарратив можно рассматривать как сложное языковое поведение, которое обычно сочетает в себе повествование и комментарии. Таким образом, нарратив может быть охарактеризован как дискурсивный акт, которому предшествуют, включают или сопровождают объяснения,

оценки, метадискурсы и другие типы комментирующих актов. Вслед за Колетта с соавторами [Coletta et al. 2010] мы выделили четыре типа повествовательных актов:

1. Повествование — предложения содержат описание события, увиденного в отрывке из мультфильма: участник пересказывает событие так, как оно появилось в мультфильме: «*Домик вертится, котик смотрит*».
2. Объяснение — предложение содержит информацию причинно-следственного характера: участник включает дополнительное объяснение увиденного в мультике события: «Потом она расстроилась, *что она всех зайчиков распугала*».
3. Домысливание — предложение представляет собой вывод или произвольную интерпретацию события или намерений персонажа: участник формирует собственные предположения или выводы относительно данного события, исходя из общих знаний: «и котенок решил к ним пойти в гости *потому что он знает там много зайчаток а у него ничего нет*»
4. Комментирование — предложение не относится к событиям, а представляет либо «метаповествовательный комментарий», относящийся к персонажу, действию или аспекту истории, либо «паранарративный комментарий» [McNeill, Levy 1993] — относящиеся к действию отвлеченные истории (суждение, личная оценка и т. д.): «махнул кубики в разные стороны и решил в часы мяч запинуть *озорник*».

Проведенный анализ показал следующие результаты.

Во всех возрастных группах есть дети, которые производят иконические жесты: в первой группе — 5 человек, во второй — 3 человека, в третьей — 3 человека; есть дети, которые имитируют действия персонажей (например, повторяют движение глаз, закрывают рукой рот и т. д.): в первой группе — 3 человека, во второй — 4 человека, в третьей — 3 человека; и есть дети, которые совсем не производят никаких жестов: в первой группе — 3 человека, во второй — 5 человек, в третьей — 6 человек[2]. Малый объем выборки, к сожалению, не позволяет провести более глубокий статистический анализ полученных данных.

[2] Во всех группах были дети, которые и производили иконические жесты, и имитировали действия персонажей.

Сопоставление типов используемых жестов и объема текста показало, что MLU у детей, использующих жесты или имитирующих персонажей, больше 5. Дети, не производящие иконические жесты, но имитирующие движения или речь персонажей, также имеют показатель MLU выше среднего (выше 4,5). Дети, не производящие жесты и не имитирующие персонажей, имеют довольно низкий показатель языковой продуктивности (до 4).

У детей, не использующих жесты и обладающих низким показателем MLU, основной тип дискурсивного акта — акт повествовательного типа. В их нарративе не встречается объяснений и домысливаний увиденных событий. Нарративы детей, использующих жестикуляцию, можно отнести к более сложным дискурсивным актам, таким как домысливание, комментирование, объяснение, причем такой результат не зависит от возраста участника. Также стоит отметить, что дети производят жесты независимо от наличия или отсутствия адресанта, так как перед ними стояла задача рассказать о событиях в мультфильме слушателю, который отворачивался от них и не мог видеть движения испытуемого.

Таким образом, развитая жестикуляции и мимика взаимосвязаны с дискурсивными способностями детей: чем больше ребенок использует иконических жестов или повторяет движения, увиденные в мультфильме, тем выше у него показатель языковой продуктивности и тем разнообразнее набор дискурсивных актов в его неподготовленных устных нарративах.

Литература

1. *Эйсмонт П. М.* (2008), Семантика и синтаксис спонтанного нарратива: Дис. … канд. филол. наук. [Рукопись].
2. *Эйсмонт П. М.* (2017), «Кондуит»: корпус устных детских текстов. Корпусная лингвистика-2017: Сборник материалов. СПб: Изд-во Санкт-Петербургского гос. ун-та, с. 373–377.
3. *Bates E., Dick F.* (2002), Language, gesture, and the developing brain. Developmental Psychobiology, Vol. 40, pp. 293–310.
4. *Cienki A.* (2019), Gestures and grammatical constructions. In: T. V. Romanova (ed.). *Integrative Processes In Cognitive Linguistics: Papers of International Congress on Cognitive Linguistics. May, 16–18* (Cognitive studies of language; Vol. 37). DEKOM Publishing House, pp. 120–125.
5. *Colletta J.-M., Pellenq C., Guidetti M.* (2010), Age-related changes in co-speech gesture and narrative: evidence from French children and adults. In: Speech Comm. Vol. 52, pp. 565–576.

6. *Goldin-Meadow S.* (2007), Nonverbal communication: The hand's role in talking and thinking. In: Damon W., Lerner R. M., Kuhn D., Siegler R. (eds.). Handbook of child psychology. Vol. 2.
7. *Kelly S. D., Barr D. J., Church R. B., Lynch K.* (1999), Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. In: Journal of Memory and Language. Vol. 40, pp. 577–592.
8. *Kotov A., Budyanskaya E.* (2012), The Russian Emotional Corpus: Communication in Natural Emotional Situations. In: Computational Linguistics and Intelligence Technologies. Vol. 11(18), pp. 296–306.
9. *Krauss R. M., Chen Y., Chawla P.* (1996), Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In: M. Zanna (ed.). Advances in experimental social psychology. Tampa: Academic Press, pp. 389–450.
10. *McNeill D., Levy E.* (1993), Cohesion and Gesture. Discourse Processes — DISCOURSE PROCESS. Vol. 16, pp. 363–386.
11. *Novack M. A., Wakefield E. M., Goldin-Meadow S.* (2016), What makes a movement a gesture? In: Cognition. Vol. 146, pp. 339–348.
12. *Rice M. L., Smolik F., Perpich D., Thompson T., Rytting N., Blossom M.* (2010), Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. In: Journal of speech, language, and hearing research: JSLHR. Vol. 53(2), pp. 333–349.

References

1. *Bates E., Dick F.* (2002), Language, gesture, and the developing brain. Developmental Psychobiology. Vol. 40, pp. 293–310.
2. *Cienki A.* (2019), Gestures and grammatical constructions. In: T. V. Romanova (ed.). *Integrative Processes In Cognitive Linguistics: Papers of International Congresson Cognitive Linguistics. May, 16–18* (Cognitive studies of language; Vol. 37). DEKOM Publishing House, pp. 120–125.
3. *Colletta J.-M., Pellenq C., Guidetti M.* (2010), Age-related changes in co-speech gesture and narrative: evidence from French children and adults. In: Speech Comm. Vol. 52, pp. 565–576.
4. *Eismont P.* (2008), Semantica i sintaksis spontannogo narrative [Semantics and syntax of spontaneous narrative]: PhD Thesis [manuscript].
5. *Eismont P.* (2017) "Konduit": korpus ustnyh detskih tekstov ["Konduit": a corpus of spoken child narratives]. In: Korpusnaya lingvistika-2017: sbornik materialov [Corpus linguistics-2017: book of abstracts]. Saint Petersburg, pp. 373–377.
6. *Goldin-Meadow S.* (2007), Nonverbal communication: The hand's role in talking and thinking. In: Damon W., Lerner R. M., Kuhn D., Siegler R. (eds.). Handbook of child psychology. Vol. 2.
7. *Kelly S. D., Barr D. J., Church R. B., Lynch K.* (1999), Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. In: Journal of Memory and Language. Vol. 40, pp. 577–592.

8. *Kotov A., Budyanskaya E.* (2012), The Russian Emotional Corpus: Communication in Natural Emotional Situations. In: Computational Linguistics and Intelligence Technologies. Vol. 11(18), pp. 296–306.
9. *Krauss R. M., Chen Y., Chawla P.* (1996), Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In: M. Zanna (ed.). Advances in experimental social psychology. Tampa: Academic Press, pp. 389–450.
10. *McNeill D., Levy E.* (1993), Cohesion and Gesture. Discourse Processes — DISCOURSE PROCESS. Vol. 16, pp. 363–386.
11. *Novack M. A., Wakefield E. M., Goldin-Meadow S.* (2016), What makes a movement a gesture? In: Cognition. Vol. 146, pp. 339–348.
12. *Rice M. L., Smolik F., Perpich D., Thompson T., Rytting N., Blossom M.* (2010), Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. In: Journal of speech, language, and hearing research: JSLHR. Vol. 53(2), pp. 333–349.

Мелещева Серафима Александровна
Санкт-Петербургский государственный университет (Россия)
Melescheva Serafima
Saint Petersburg State University (Russia)
E-mail: st064859@student.spbu.ru

Эйсмонт Полина Михайловна
Санкт-Петербургский государственный университет (Россия)
Eismont Polina
Saint Petersburg State University (Russia)
E-mail: p.eysmont@spbu.ru

И. С. Пименов

I. S. Pimenov

СПЕЦИФИКА АРГУМЕНТАЦИОННОГО АННОТИРОВАНИЯ НАУЧНЫХ И НАУЧНО-ПОПУЛЯРНЫХ ТЕКСТОВ

ON ARGUMENTATION ANNOTATION OF SCIENTIFIC AND POPULAR SCIENCE TEXTS

Аннотация. Рассматривается задача ручной разметки аргументации в научных и научно-популярных русскоязычных текстах. Метод построения графов аргументации реализуется через набор веб-инструментов, обеспечивающих создание тематических корпусов, визуализацию аргументативных утверждений и схем аргументации. Указываются специфические сложности в применении метода к текстам выбранных жанров. Приводятся сравнительные количественные характеристики аргументационных структур текстов разных жанров и тематик.
Ключевые слова. Аргументационное аннотирование, схемы Уолтона, граф аргументации, научные и научно-популярные тексты.

Abstract. The task of manual argumentation annotation is approached for scientific and popular science texts in Russian language. The method of building argumentative graphs is implemented with the set of web-tools for creating thematic corpora and visualizing argumentative statements and schemes. Annotation of texts is performed through the use of argumentative schemes from Walton's compendium in three distinct steps (by identifying argumentative statements, analyzing relations between them, specifying exact schemes). Specific difficulties in applying the method to texts of the chosen genres are noted. A quantitative comparison of argumentation structures is performed between scientific texts of different thematic areas and genres.
Keywords. Argumentation annotation, Walton's argumentation schemes, argumentation graph, scientific and popular science texts.

1. Введение

Особой задачей в области автоматического извлечения информации выступает распознавание в тексте аргументов, или выстраиваемых рассуждений, с определением организующих их отношений. Изучение аргументации (способов ее выражения в естественном языке, методов ее извлечения) актуально в свете следующих приложений:

1) анализ отзывов (для определения не только тональности текста, но и обоснования оценки);
2) понимание и ведение дебатов;
3) принятие решений в рекомендательных системах;
4) обнаружение радикальных мнений, вводящих в заблуждение текстов и т. д.

Задача аннотирования аргументации отличается сложностью анализируемых сущностей (пропозиций, выражаемых в сегментах текста различной длины при значительном разнообразии лексических и грамматических средств) и связей, в качестве которых выступают модели рассуждений. Близким к аргументационной разметке по сложности является аннотирование риторических структур [Pisarevskaya et al. 2017].

В данной работе описывается эксперимент по ручной разметке аргументации в научных и научно-популярных текстах. Жанровая специфика выбранного материала характеризуется значительным объемом текстов (в противовес новостным сообщениям, интернет-комментариям, эссе и микротекстам, из которых чаще всего составляются аргументационные корпусы [Daxenberger et al. 2017]), а также особенностями применяемых схем (в том числе ввиду монологической организации текстов). Помимо этого, перспектива машинной обработки аннотаций и использования коллекции для машинного обучения накладывает определенные ограничения на действия аннотатора.

Обзор литературы показывает, что корпусов научных текстов с аргументационной разметкой немного: в [Lauscher et al. 2018] описывается создание такого корпуса на материале английского языка с указанием на отсутствие аналогов.

Цель работы: адаптация методики аргументационного аннотирования к разметке научных и научно-популярных текстов с учетом применения аннотаций для машинного обучения.

2. Задача аргументационного аннотирования

Аргументационное аннотирование текста подразумевает моделирование структуры выстроенной в нем аргументации (системы связанных аргументов). Аргументы соответствуют утверждениям, сформулированным на языке пропозиций и объединенным отношениями, которые выражают поддержку или опровержение одними утверждениями других (соответственно аргументативными и конфликтными связями).

Связанные утверждения (тезисы) в составе аргумента различаются по ролям: среди них выделяется доказываемое утверждение (заключение), обосновываемое иными (посылками). Роли определяются отдельно для каждого аргумента и могут меняться при связи нескольких аргументов: заключение одного способно служить посылкой в другом.

На уровне полного текста определяется главный тезис, выражающий его основную идею.

Разметка аргументационной структуры текста предполагает уточнение семантики связей через их соотнесение со схемами аргументов (для отношений поддержки) или схемами конфликтов (при опровержениях). Схемы соответствуют типовым моделям рассуждений из фиксированного набора. В данной работе применяются схемы из компендиума Уолтона [Walton et al. 2008].

Структура аргументации полного текста представляется в соответствии с форматом AIF (Argument Interchange Format) [Rahwan, Reed 2009] посредством ориентированного графа с двумя типами вершин (информационными для утверждений и вершинами-схемами для схем аргументов и конфликтов). Так, решаемая задача состоит в представлении неструктурированного текста в виде связного аргументационного графа с помощью заданного набора схем рассуждений (связность графа означает наличие не менее одного пути между любыми двумя вершинами).

Аннотирование произведено с помощью набора веб-инструментов, разрабатываемых в ИСИ СО РАН [Сидорова и др. 2020] для создания тематических корпусов текстов, визуализации аргументативных утверждений и схем аргументации. Инструмент предоставляет возможность сохранения графа в формате .json, что позволяет проводить компьютерную обработку аннотаций.

3. Проблемы и методы аннотирования аргументации

Моделирование аргументационной структуры каждого текста из корпуса производилось вручную одним разметчиком. Разметка текстов означала решение трех подзадач, типичных для аннотирования аргументации ([Lawrence, Reed 2019]):

1) *Выявление аргументативных утверждений* через анализ их пропозиционального содержания. При этом в научных и научно-популярных текстах часть аргументов может выражаться таблицами или рисунками, как с интерпретацией на естественном языке (доступной для аннотирования), так и без нее. Неявные тезисы (не представленные эксплицитно) не выделялись при разметке текстов, предназначенных для машинного обучения.
2) *Обнаружение связанных аргументативных утверждений*. Построение связей между тезисами проводилось итеративно от

главного тезиса по обновляемому списку не присоединенных к графу утверждений: определялись напрямую атакующие или поддерживающие посылки, для которых затем выявлялись связанные тезисы в постепенно расширяемой анализируемой окрестности текста с учетом уже построенной аргументационной структуры (для выявления связей между позиционно удаленными утверждениями, которое представляет особую сложность).

3) *Уточнение характера построенных связей* за счет знаний о типовых схемах рассуждений, или детализация графа на уровне вершин-схем. Перспектива компьютерной обработки аннотаций и их использования в машинном обучении влечет унификацию разметки: ограничение числа используемых схем, обозначение точных критериев выбора одной схемы среди нескольких схожих.

4. Компьютерная обработка разметки в эксперименте

Аннотированная часть корпуса включает 110 текстов, из них 10 научных были размечены с целью дальнейшего использования в машинном обучении (5 лингвистических (*Ling*) из области лексикологии с анализом отдельных тематических групп слов, в том числе в культурологическом аспекте, и 5 по компьютерным технологиям (*Comp*) с представлением разработанных решений прикладных задач из областей анализа изображений, интеграции автоматизированных систем, разработки игровых алгоритмов). Другая часть корпуса — тексты научно-популярного жанра (*PS*), аннотированные для изучения приемов аргументации. Научные статьи взяты из корпуса Ru-RSTreebank [Pisarevskaya et al. 2017] (сведения об их риторических структурах не учитывались). Визуализация подграфа аргументации с иллюстрацией случаев реализации наиболее частотных схем приводится на рис. 1.

Сравнительные характеристики каждой из групп текстов корпуса (размер в словах N_W, число выделенных тезисов N_S, число построенных аргументативных связей N_A, количество и доля слов в тезисах N_T) указаны в табл. 1. Так, при меньшем объеме научных текстов в них выделено большее число тезисов.

В табл. 2 для каждой группы текстов указываются схемы аргументации с относительной частотой $F \geq p$ ($F = N_A^i \times 100/N_A$, N_A^i — число реализаций i-й схемы в группе, $p = 5\%$). Полужирным шрифтом вы-

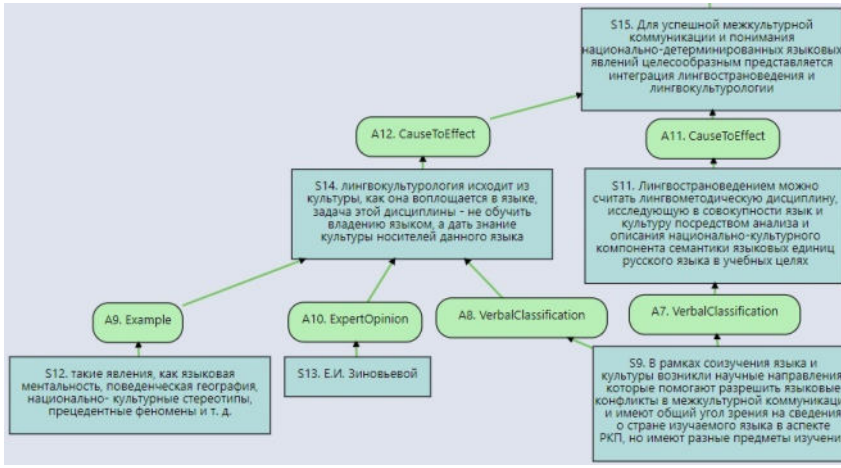


Рис. 1. **Пример фрагмента аргументационной разметки**

Таблица 1. **Количественные характеристики текстов корпуса**

	Comp	***Ling***	***PS***
N_W	1149	952	1535
N_S	44	45	20
N_A	45	45	14
N_T	662 (57,6 %)	657 (69,0 %)	316 (20,6 %)

делены частоты трех схем с наибольшими F в каждой группе текстов, курсивом — схем с $F < p$ в данной группе.

Среди аргументационных схем, наиболее часто реализуемых в текстах корпуса, можно выделить группы семантически схожих: обобщенные причинно-следственные доказательства (Cause to Effect, Correlation to Cause), рассуждения практического плана (Practical Reasoning, Positive/Negative Consequences), обоснования через свидетельство (Example, Expert Opinion, Sign).

Сравнение же размеченных научных статей по лингвистике и компьютерным технологиям позволило выявить тематическую зависимость в применении схем. Во-первых, в лингвистических текстах корпуса намного чаще реализуются рассуждения через отсылку к экспертному мнению (схема Expert Opinion) или через глубинный ана-

Таблица 2. **Встречаемость частотных аргументативных схем**

Схема	F_{Comp}	F_{Ling}	F_{PS}
Cause to Effect (*от причины к следствию*)	**15 %**	13 %	**18 %**
Correlation to Cause (*через взаимосвязь*)	14 %	**23 %**	5 %
Example (*от примера*)	**20 %**	**27 %**	**24 %**
Expert Opinion (*от мнения эксперта*)	*1 %*	7 %	11 %
Modus Ponens *(через «правило вывода»)*	–	–	**17 %**
Negative Consequences (*от осложнений*)	6 %	–	*1 %*
Positive Consequences (*от выгоды*)	5 %	–	*1 %*
Practical Reasoning *(от практической цели)*	12 %	7 %	*2 %*
Sign (*от знака к означаемому*)	*2 %*	–	6 %
Verbal Classification (*через классификацию*)	**22 %**	**20 %**	*1 %*
Суммарные доли приведенных схем:	98 %	97 %	86 %

лиз изучаемых сущностей по их внешним проявлениям (Correlation to Cause). Эти отличия объясняются тем, что в представленных текстах по компьютерным технологиям описываются разработки авторов, отчего снижается потребность в цитировании иных исследователей, а созданные программы удобно представлять от их устройства к функционированию.

Во-вторых, прикладной характер статей по компьютерным технологиям часто влечет анализ практических последствий (и положительных, и отрицательных) от возможных подходов к разработке программы (схемы Positive/Negative Consequences). Рассуждения от практической цели (Practical Reasoning) тоже более свойственны статьям по компьютерным технологиям.

Схожесть двух групп научных текстов проявляется в частом обращении к примерам (схема Example), активном использовании классификаций для анализа различных аспектов исследуемых явлений (Verbal Classification) и причинно-следственных связей общего характера (схема Cause to Effect).

Сравнение частот аргументативных схем позволяет также выявить особенности организации рассуждений в научно-популярных текстах. Во-первых, в их разметке практически не встречается схема

Verbal Classification (не требуется высокий уровень систематизации сведений). Во-вторых, доказательства в научно-популярных текстах чаще обращены к авторитетному источнику (ввиду возможности основания работ этого жанра на сопоставлении экспертных воззрений). В-третьих, научно-популярным статьям более свойственно построение общих причинно-следственных связей без глубокой дифференциации (при разметке таких отношений применялась схема Modus Ponens, понимаемая более свободно) ввиду большей обобщенности.

С другой стороны, научные и научно-популярные тексты схожи равной долей использования примеров и реализацией небольшого числа отдельных схем аргументации, что способно помочь отличать их от текстов иных жанров.

5. Заключение

Описанный эксперимент позволил выявить особенности аргументационной разметки текстов разных жанров и тематик в рамках академической коммуникации, выделить узкие места в реализации схем аргументации и выработать правила построения аргументационных структур для машинного обучения.

Литература

1. *Сидорова Е. А., Ахмадеева И. Р., Загорулько Ю. А., Серый А. С., Шестаков В. К.* (2020), Платформа для исследования аргументации в научно-популярном дискурсе. Онтология проектирования. Т. 10, № 4(38), с. 489–502.
2. *Daxenberger J., Eger S., Habernal I., Stab C., Gurevych I.* (2017), What is the Essence of a Claim? Cross-Domain Claim Identification. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2055–2066.
3. *Lauscher A., Glavaš G., Ponzetto S.* (2018), An Argument-Annotated Corpus of Scientific Publications. Proceedings of the 5th Workshop on Argument Mining, pp. 40–46.
4. *Lawrence J., Reed C.* (2019), Argument Mining: A Survey. Computational Linguisticsю Vol. 45(4), pp. 765–818.
5. *Pisarevskaya D.,* Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. (2017), Towards building a discourse-annotated corpus of Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017", pp. 194–204.
6. *Rahwan I., Reed C.* (2009), The argument interchange format. Argumentation in artificial intelligence, pp. 383–402.
7. *Walton D., Reed C., Macagno F.* (2008), Argumentation schemes. New York: Cambridge University Press, 443 p.

References

1. *Daxenberger J., Eger S., Habernal I., Stab C., Gurevych I.* (2017), What is the Essence of a Claim? Cross-Domain Claim Identification. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2055–2066.
2. *Lauscher A., Glavaš G., Ponzetto S.* (2018), An Argument-Annotated Corpus of Scientific Publications. Proceedings of the 5th Workshop on Argument Mining, pp. 40–46.
3. *Lawrence J., Reed C.* (2019), Argument Mining: A Survey. Computational Linguistics. Vol. 45(4), pp. 765–818.
4. *Pisarevskaya D.,* Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. (2017), Towards building a discourse-annotated corpus of Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017", pp. 194–204.
5. *Rahwan I., Reed C.* (2009), The argument interchange format. Argumentation in artificial intelligence, pp. 383–402.
6. *Sidorova E. A, Akhmadeeva I. R., Zagorulko Y. A., Seryj A. S., Shestakov V. K.* (2020), Platforma dlja issledovanija argumentacii v nachno-popularnom discurse [Research Platform for the Study of Argumentation in Popular Science Discourse]. Ontologija proektirovaniya [Ontology of Designing]. Vol. 10, No. 4(38), pp. 489–502.
7. *Walton D., Reed C., Macagno F.* (2008), Argumentation schemes. New York: Cambridge University Press, 443 p.

Пименов Иван Сергеевич
Новосибирский государственный университет (Россия)
Pimenov Ivan
Novosibirsk State University (Russia)
E-mail: pimenov.1330@yandex.ru

В. И. Подлесская

V. I. Podlesskaya

АППРОКСИМАТОР *ЧТО НАЗЫВАЕТСЯ*: ЭТАПЫ ПРАГМАТИЗАЦИИ В ЗЕРКАЛЕ НКРЯ

STEPPING FROM PROPOSITIONS TO PRAGAMATIC MARKERS: THE CASE OF RUSSIAN *CHTO NAZYVAETSJA* 'WHAT MIGHT BE CALLED' THROUGH THE LENS OF RUSSIAN NATIONAL CORPUS

Аннотация. На материале мультимедийного и, частично, параллельного подкорпусов НКРЯ исследованы семантические, грамматические и просодические свойства конструкции *что называется X*. Показано, что в устной речи сохраняются одновременно разные этапы перехода от конструкции с относительным придаточным [*это то*] *что называется X-ом* к прагматическому маркеру с аппроксимативным значением. Исследованная конструкция традиционно рассматривается в русистике как вводная, и наши данные подтверждают, что она удовлетворяет большинству функциональных и структурных требований к прототипической парентезе.
Ключевые слова. Устный дискурс, корпус, парентеза, русский язык, прагматикализация.

Abstract. The paper investigates Russian construction *chto nazyvaetsja* 'what might be called' in the multimedia and parallel subcorpora of the Russian Nation Corpus. We demonstrate how the construction in question is developed from the full-fledged relative clause [*eto to*] *chto nazyvaetsja X* '[this is] what is called X' to a pragmatic marker used for hedging. We show that steps of this development are retained in spoken discourse. Our results show that the construction in question meets most (though not all) of the requirements considered necessary for true parentheticals.
Keywords. Spoken discourse, corpus, parenthesis, Russian, pragmaticalization.

1. Постановка вопроса

В работе исследуются синтаксис, семантика и просодия конструкции *что называется X*:

(1) Всякий живет сибаритом…
Майков, Полонский и Фет —
Подступу к этим пиитам,
Что называется, нет!

[Н. А. Некрасов. «Что ты задумал, несчастный?..» | Мысли журналиста при чтении программы, обещающей не щадить литературных авторитетов (1860)][1]

[1] Примеры приводятся в той графической форме, в которой они даны в НКРЯ.

(2) Безнадежно. Предметов, по крайней мере,
на тебя похожих наощупь, в мире,
что называется, кот наплакал.
Какова твоя жертва, таков оракул.

[И. А. Бродский. «Если кончу дни под крылом голубки…» | Прощайте, мадемуазель Вероника (1967)]

Основной материал исследования — мультимедийный подкорпус Национального корпуса русского языка (МУРКО). В иллюстративных целях используются также и другие подкорпусы НКРЯ, прежде всего, параллельный.

Изложение будет строиться следующим образом. В разделе 2 описывается употребление этой конструкции в качестве аппроксиматора, т. е. прагматического маркера, функция которого — указать на приблизительность и контекстную обусловленность выбранной говорящим номинации. В разделе 3 показаны ступени прагматизации этой конструкции: от нейтрального пропозиционального употребления в качестве относительного придаточного (*[это то] что называется X-ом/X*) к употреблению в функции прагматического маркера. Как мы увидим, в живой устной речи различные ступени этого перехода легко сосуществуют одновременно, характеризуясь при этом различной степенью грамматических и просодических сдвигов. В заключительном разделе 4 мы введем наши данные в контекст теоретической дискуссии о лингвистическом статусе парентезы и покажем, что исследуемая конструкция удовлетворяет большинству требований, которые обычно предъявляются к парентетическим выражениям.

2. Прагматический маркер *что называется*

Конструкция *что называется* относится к классу аппроксиматоров, или маркеров приблизительной номинации [Подлесская, Стародубцева 2013]. Ее основная прагматическая функция состоит в том, чтобы сообщить слушающему, что говорящий выбрал номинацию не автоматически, ему пришлось приложить определенные усилия и что в своем выборе он опирается на авторитет сложившегося узуса, считая, что другие говорящие в подобной ситуации выбрали бы в качестве наиболее подходящей именно эту номинацию, или что он сам в подобной ситуации уже использовал эту номинацию. При этом говорящий признает, что выбранная в конце концов номинация может оказаться приблизительной.

Цитатность выбранной номинации в письменном тексте может подчеркиваться кавычками:

(3) Если бы я был, что называется, «средним французом», охочим до разряженных дам… [Владимир Набоков. Лолита (В. Набоков, 1967)]

Примечательно, что, согласно параллельному подкорпусу, в английской версии «Лолиты» В. Набоков в этом тексте не использует ни конструкций с тем или иным глаголом именования, аналогичных *что называется*, ни кавычек, зато использует такой маркер цитатности, как вставка иноязычного выражения:

(4) Had I been a frangais moyen with a taste for flashy ladies… [Vladimir Nabokov. Lolita (1955) | Владимир Набоков. Лолита (В. Набоков, 1967)]

Данные параллельного корпуса позволяют выявить и другие компоненты сформулированного выше аппроксимативного значения. Обращение к сложившемуся узусу передается в английском корреляте через обобщенно-личное *they*, обращение к собственному речевому опыту говорящего — через первое лицо:

(5) So he went into the dining-room and "glued his face" as they say, to the window. [Clive Staples Lewis. The Chronicles of Narnia. The Magician's Nephew (1955)

Дигори отправился в столовую и, что называется, прилип к оконному стеклу. [Клайв Стейплз Льюис. Хроники Нарнии. Племянник чародея (Г. А. Островская, 1991)]

(6) Да… Летчик, что называется, милостью божьей, — проворчал подполковник. [Б. Н. Полевой. Повесть о настоящем человеке (1946) | Boris Polevoi. A Story about a real man (Joe Fineberg, 1950)] [омонимия не снята] ←…→

Yes… you are what I call an airman by the grace of God, growled the lieutenant-colonel. [Boris Polevoi. A Story about a real man (Joe Fineberg, 1950)]

То, что говорящему приходится выбирать номинацию из нескольких альтернатив, передается в английском корреляте сослагательным наклонением глагола:

(7) Вот этого уж никак не могу. Андреев — это, что называется. — Ваш личный враг? [Варлам Шаламов. Потомок декабриста (1962) | Varlam Shalamov. Descendant of a Decembrist (John Glad, 1980)] [омонимия не снята] ←…→

That's one thing I'll never do. Andreev is, how should I put it…? Your personal enemy? [Varlam Shalamov. Descendant of a Decembrist (John Glad, 1980)]

Неточность выбранной номинации передается через использование аппроксиматоров, в том числе, восходящих к словам со значением 'сорт, вид', ср. *kind of*:

(8) with symmetrically wrinkled cheeks and the kind of complexion termed hemorrhoidal… [Vladimir Nabokov. Nikolai Gogol (1944) | Владимир Набоков. Николай Гоголь (Е. Голышева, 1993)]

с морщинами по обеим сторонам щек и цветом лица что называется геморроидальным… [Владимир Набоков. Николай Гоголь (Е. Голышева, 1993)]

В устной речи компонент значения, связанный с неоднозначностью выбора номинации из возможных альтернатив, проявляется в том, что использование выражения *что называется* часто совмещено с речевыми сбоями — самоисправлениями, повторами, заполненными паузами и с одновременным использованием других аппроксиматоров (e.g. *как бы, так сказать*):

(9) То есть это такие как бы игра в выборы/ что на… что называется/ про… прошла. Произошла обкатка кандидатов. [Сергей Собянин в программе «Право знать!» (2014)]

(10) Поэтому аа вы/ аа так сказать/ что называется/ на фронте всей этой работы с молодежью. [Александр Бречалов. Лекция о лидерах общественного мнения и молодежной политике в РАНХиГС (2013)

В русском языке выражение *что называется* в функции аппроксиматора появляется рано — согласно хронологии основного корпуса НКРЯ, самые ранние выдачи по запросу «*что* на расстоянии 1 от *называться* praes» датируются концом XVIII — началом XIX века, и там аппроксиматоры уверенно присутствуют наряду со случаями употреблений в формате относительных придаточных, вводимых местоиме-

нием *что*, лишенных прагматического компонента приблизительности номинации и проблем с ее выбором, ср. вводное употребление *что называется* в (11) и не вводное в (12):

(11) Батюшка, обмундировав меня что называется с ног до головы, повез меня в Сарское Село [И. М. Долгоруков. Повесть о рождении моем, происхождении и всей моей жизни, писанная мной самим и начатая в Москве, 1788-го года в августе месяце, на 25-ом году моей жизни / Части 1-2 (1788–1822)]

(12) Прелести ваши способны вливать в душу смертного радости живящие; но Любослав чужд всего, что называется радостию; в обширной стране своей он есть узник горести и томления. [В. Т. Нарежный. Славенские вечера (1809)]

В следующем разделе мы обсудим грамматические и просодические признаки, которые отличают прямое употребление от аппроксимативного, и на материале МУРКО продемонстрируем, что в устной речи эти признаки могут нейтрализоваться, что приводит к сосуществованию вариантов с разной степенью прагматизации.

3. *Что называется*: от относительного придаточного к прагматическому маркеру

Прямые, не прагматизированные употребления последовательности «*Y, что называется X*», подобные приведенному выше (12), широко встречаются и в современной устной речи, о чем свидетельствуют данные МУРКО, см., например, (13), (14):

(13) Вот эта кривая показывает рост человечества/ населения Земли от минус двухтысячного года/ значит/ того/ что называется в истории Древним миром/ где-то здесь/ оно вот так доходит до Рождества Христова. [Сергей Капица. Россия и мир в демографическом зеркале. Проект Academia (ГТРК Культура) (2012)]

(14) И более того/ наличие вот такого квадратичного закона всякий изучающий физику знает/ что это указывает на наличие того/ что называется коллективными процессами. Это уже основы молекулярной физики. [Сергей Капица. Россия и мир в демографическом зеркале. Проект Academia (ГТРК Культура) (2012)]

Прямые употребления демонстрируют следующие прототипические свойства:

(i) Наличие вершины Y, представленной предикатной или именной группой. Именная группа может быть полной или местоименной, в последнем случае чаще всего это соотносительное местоимение *то*; эта именная группа может выступать в любом падеже;
(ii) Группа (обычно именная), обозначающая номинацию X, может выступать в именительном или творительном падеже;
(iii) Внутри относительной клаузы возможны распространители — актанты и сирконстанты;
(iv) Глагол *называться* может употребляться в других формах (времени числа и проч.) и может быть заменен на другие глаголы называния (*именоваться*, *зваться* и т. п.);
(v) Группа (обычно именная), обозначающая номинацию X, тяготеет к правой периферии клаузы и является носителем фразового акцента.

Переход к аппроксимативному употреблению сопровождается полной или частичной утратой перечисленных выше грамматических и просодических признаков. Прежде всего, относительное придаточное лишается вершины Y:

(15) это задача/ которую надо решать/ что называется/ всем миром. [Круглый стол, посвященный прямым выборам мэра Москвы (2013)]

В устной разговорной речи вершина может сохраняться рудиментарно в виде соотносительного местоимения *то* (строго в единственном числе, именительном падеже), но падеж местоимения либо вовсе не лицензируется синтаксической структурой, см. (16), либо лицензируется не тот падеж, который реализуется, см. (17), где контекст требует родительного падежа местоимения, а возникает именительный (ср. правильное *во время того, что называется «ковровой бомбардировкой»*):

(16) Лужники те же/ то есть ну мы там все и живем/ и… и как-то и то/ что называется тусу-тусуемся там/ на районе/ как модно гово говорить у молодежи. [Сергей Кузнецов в передаче «Звезда на «Звезде» с Александром Стриженовым (2016)]

(17) Но во время / то/ что называется «ковровой бомбардировки» всей вот этой ближней части солнечной системы/ все кратеры на Луне основные/ они примерно одного времени. [А. Ю. Журавлев. Кто жил полмиллиарда лет назад в Сибири и Африке (2016)]

Пример (17) показывает и сдвиг в оформлении именной группы, обозначающей выбранную номинацию X: если мы имеем дело с прагматическим маркером аппроксимации, а не с прямым употреблением, то X получает тот падеж, который лицензируется извне, а не творительный или именительный, которые могли бы быть лицензированы глаголом *называться* при его полнозначном употреблении. Иначе говоря, в (17) без прагматического маркера *что называется* структура была бы с X в родительном падеже: *во время ковровой бомбардировки*; добавление маркера, даже с рудиментарным *то*, не изменило падежа X — родительного. Если бы мы имели дело с прямым употреблением, то соотносительное местоимение (Y) получило бы родительный падеж — от управляющего им *во время*, а X (*ковровая бомбардировка*) получил бы творительный или именительный от глагола *называться*: *во время того, что называется ковровой бомбардировкой*. По существу, рудиментарное сохранение соотносительного *то* сходно с разговорным неологизмом *то что* в качестве комплементайзера [Коротаев 2016; Князев 2019; Сердобольская, Егорова 2019].

Другими симптомами утраты грамматических и лексических «свобод» при прагматизации данной конструкции являются: запрет на добавление актантов и сирконстантов, запрет на употребление иных видовременных форм глагола, кроме *называется*, запрет на замену этого глагола на синонимичные.

Интересные сдвиги наблюдаются в линейно-акцентной структуре конструкции. Как было сказано выше, при прямом употреблении группа, обозначающая номинацию X, тяготеет к правой периферии клаузы и является носителем фразового акцента. Этот паттерн сохраняется и при прагматизованном употреблении: последовательность *что называется* всегда произносится атонически и непосредственно примыкает к носителю фразового акцента. В общем случае, *что называется* в качестве прагматического маркера примыкает слева к рематической составляющей. Так, в следующем примере рема — глагольная группа *носятся в воздухе*, ее акцентоноситель *(в) воздухе*, прагматический маркер *что называется* произносится без акцента:

(18) эти мысли/ что называется/ носятся в воздухе. [Технологии виртуальной реальности. Программа «Гордон» (НТВ) (2003)]

Однако при прагматизованном употреблении становится возможным также размещение группы, обозначающей номинацию X, слева, а не справа от *что называется* — порядок, запрещенный при прямом употреблении; при этом фразовый акцент на X и безударность *что называется* сохраняются. Причем маркер может примыкать справа как к рематической составляющей, см. (19), так и к тематической, см. (20):

(19) Можно ли вообще воспитывать таких людей? Или этот дар Божий/ что называется. [В. И. Арнольд, С. П. Капица. Задачи Владимира Арнольда. Передача «Очевидное — невероятное» (2009)]

(20) И если сейчас аа «с колес»/ что называется/ идут любые материалы в ээ сегодняшней газете/ то тогда был достаточно жесткий отбор. [С. Филатов в цикле В. Горшенина «Правдисты» (2012)]

Итак, мы постарались показать, что приобретение прагматического — аппроксимативного — значения сопровождается утратой лексических и грамматических «свобод» и изменениями линейно-акцентной структуры. Эти наблюдения дают возможность рассмотреть прагматизацию конструкции *что называется* в контексте современной дискуссии о лингвистической природе парентезы.

4. *Что называется* и понятие парентезы

В литературе по парентезе слова, конструкции и предложения квалифицируются как парентетические (вводные), прежде всего, по прагматической функции, которая сводится к признанию вторичности содержания вводной конструкции, к тому, что содержание вводного фрагмента не вносит вклад в пропозициональное содержание опорного высказывания (*non-at-issue content*, [Kluck, Ott, de Vries 2015; Potts 2005]). Эта функция может манифестироваться в ряде симптомов, в том числе:

(а) Парентеза предполагает вставку — она разрывает текущее высказывание на два фрагмента, между которыми сохраняется тесная связь — синтаксическая или риторическая (дискурсивная);

(б) Парентетический фрагмент не формирует узла в синтаксической структуре опорного высказывания;

(в) Парентетический фрагмент просодически выделен из опорного высказывания: его границы просодически маркированы, текущие просодические параметры (тоновый диапазон, темп, громкость) отличаются от таковых в опорном высказывании, он не имеет внутреннего коммуникативно-просодического членения;
(г) В парентетическом фрагменте имеются ограничения на лексическое многообразие и морфосинтаксис.

Обследованная нами конструкция в своем аппроксимативном значении функционально (прагматически) является вводной. Однако не все ее свойства вписываются в прототип парентезы. Безусловно, выполняется требование (г): грамматическая и лексическая вариативность аппроксиматора ограничена по сравнению с прямым употреблением. Требование (б) выполняется частично: прагматический маркер не формирует отдельного синтаксического узла, однако формально конструкция вводится типичным коннектором — относительным местоимением (*что*). В то же время, *что называется* не обязательно появляется во вставке — как мы видели, этот аппроксиматор может появляться, например, в крайней правой позиции. Просодически *что называется* примыкает к текущей составляющей, не отделяется просодическим швом. То есть требования (а) и (в) не выполняются. Таким образом, результаты нашего корпусного исследования позволяют предположить, что наиболее плодотворным подходом к описанию парентезы является многофакторный анализ, позволяющий учитывать все многообразие исследуемой зоны и выделять лингвоспецифические кластеры значений релевантных параметров.

Литература

1. *Князев М. Ю.* (2019), Экспериментальное исследование дистрибуции изъяснительного союза *то что* в русской разговорной речи. Вопросы языкознания. № 5, с. 7–40.
2. *Коротаев Н. А.* (2016), Союз *то что* в устной русской речи. Acta Linguistica Petropolitana. Т. XII, ч. 1, с. 101–106.
3. *Подлесская В. И., Стародубцева А. В.* (2013), О грамматике средств выражения нечеткой номинации в живой речи. Вопросы языкознания. № 3, с. 25–41.
4. *Сердобольская Н. В., Егорова А. Д.* (2019), Морфосинтаксические свойства ненормативных конструкций с *то что* в русской разговорной речи. Вопросы языкознания. № 5, с. 41–72.

5. *Kluck M., Ott D., de Vries M.* (2015), Incomplete parenthesis: An overview. In: M. Kluck, D. Ott, M. de Vries (eds.). Parenthesis and Ellipsis. Studies in Generative Grammar. Vol. 121. De Gruyter Mouton, pp. 1–22.
6. *Potts C.* (2005), The Logic of Conventional Implicatures. New York: Oxford University Press.

References

1. Kluck M., Ott D., de Vries M. (2015), Incomplete parenthesis: An overview. In: M. Kluck, D. Ott, M. de Vries (eds.). Parenthesis and Ellipsis. Studies in Generative Grammar. Vol. 121. De Gruyter Mouton, pp. 1–22.
2. *Knjazev M. Ju.* (2019), Eksperimental'noe issledovanie distribucii izjasnitel'nogo sojuza *to chto* v russkoj razgovornoj rechi [An experimental study of the distribution of the complementizer *to čto* in non-standard variants of Russian]. In: Voprosy Jazykoznanija [Topics in the study of language]. No. 5, pp. 7–40.
3. *Korotaev N. A.* (2016), Sojuz *to chto* v ustnoj russkoj rechi [*To čto* as a new complementizer in Russian spoken discourse]. In: Acta Linguistica Petropolitana. Vol. XII, part 1, pp. 101–106.
4. *Podlesskaya V. I, Starodubceva A. V.* (2013), O grammatike sredstv vyrazhenija nechetkoj nominacii v zhivoj rechi [Grammar of vague reference in unprepared discourse]. In: Voprosy Jazykoznanija [Topics in the study of language]. No. 3, pp. 25–41.
5. *Potts C.* (2005), The Logic of Conventional Implicatures. New York: Oxford University Press.
6. *Serdobol'skaja N. V., Egorova A. D.* (2019), Morfosintaksicheskie svojstva nenormativnyx konstrukcij s *to chto* v russkoj razgovornoj rechi [Morphosyntax of non-standard constructions with *to čto* in Colloquial Russian]. In: Voprosy Jazykoznanija [Topics in the study of language]. No. 5, pp. 41–72.

Подлесская Вера Исааковна
Российский государственный гуманитарный университет (Россия)
Podlesskaya Vera
Russian State University for the Humanities (Russia)
E-mail: vi_podlesskaya@il-rggu.ru

Г. Саяма
G. Sayama

ВЛИЯНИЕ ЧАСТОТНОСТИ НА ФОРМЫ ПАДЕЖНЫХ ОКОНЧАНИЙ *-АМ, -АМИ, -АХ* (*-ЯМ, -ЯМИ, -ЯХ*)

THE INFLUENCE OF FREQUENCY ON THE FORMS OF CASE ENDINGS *-АМ, -АМИ, -АХ* (*-ЯМ, -ЯМИ, -ЯХ*)

Аннотация. Данная статья посвящена вопросу о том, почему в русском языке дательный, творительный, предложный падежи множественного числа имеют общие формы окончаний на *-ам, -ами, -ах* (*-ям, -ями, -ях*) во всех типах склонений: в данных трех падежах не существует алломорфов окончаний в современном русском языке, за некоторыми исключениями. Мы считаем, что влияние на ныне сложившуюся парадигму оказала частота употребления этих падежей. Анализ во всех корпусах показал, что и в современном русском, и в древнерусском языках они имеют меньшую частоту употребления по сравнению с другими падежами, из чего можно предположить, что их низкая частотность привела к общим формам во всех типах склонений.
Ключевые слова. Русский язык, частотность, дательный падеж, творительный падеж, предложный падеж, древнерусский язык.

Abstract. This article is devoted to the question of why in Russian plural dative, instrumental, prepositional cases have common endings *-ам, -ами, -ах* (*-ям, -ями, -ях*) in all declensions: these 3 cases have no allomorphs of endings in modern Russian, with some exceptions. We consider that the condition is influenced by the frequency of the three cases. The analyses in all corpora showed that in modern Russian and old Russian languages they are used less than other cases. The results expect that their low frequency led to the general forms of all genders.
Keywords. Russian language, frequency, dative case, instrumental case, prepositional case, Old Russian language.

1. Введение

1.1. Формы окончаний дательного, творительного, предложного падежей во множественном числе

В русском языке в парадигмах ед. ч. сохраняется различие склонений в трех типах. Однако характерное для ед. ч. различие исчезает в д. п., т. п. и п. п. мн. ч.: эти три падежа во мн. ч. имеют общие формы окончаний *-ам, -ами, -ах* (*-ям, -ями, -ях)* вне зависимости от типов склонений [Шведова (ред.) 1980].

Как выделено жирным шрифтом в табл. 1, во всех типах падежные окончания д. п., т. п., п. п. во мн. ч. имеют общие формы.

Данная парадигма в 3-х падежах представляет необычное явление с учетом того, что р. п. во мн. ч. очень богат формами окончаний, такими, как *-ов, -ев, -ей*, *-Ø* (нулевое окончание) и др.

Таблица 1. **Пример склонений существительных во мн. ч.**

Падеж	I скл.		II/III скл.
	(Муж. род)	(Ср. род)	(Жен. род)
И. п.	столы/музеи	места/моря	карты/радости
Р. п.	столов/музеев	мест/морей	карт/радостей
Д. п.	стол**ам**/музе**ям**	мест**ам**/мор**ям**	карт**ам**/радост**ям**
В. п.	столы/музеи	места/моря	карты/радости
Т. п.	стол**ами**/музе**ями**	мест**ами**/мор**ями**	карт**ами**/радост**ями**
П. п.	стол**ах**/музе**ях**	мест**ах**/мор**ях**	карт**ах**/радост**ях**

1.2. Цель исследования

Ссылаясь на предыдущие исследования, касающиеся других славянских языков, автор [Чино , 1972] высказал гипотезу о том, что влияние на общие формы падежных окончаний д. п., т. п., п. п. во мн. ч. во всех типах оказывает частота их употребления: в древнерусском языке они имели разные формы окончаний в зависимости от типа, однако с течением времени слились в одну форму, поскольку их употребляли *наиболее редко*. Однако данная гипотеза не была доказана количественно с помощью больших языковых данных.

Для получения объективного и количественного доказательства о влиянии частотности на общие формы окончания д. п., т. п., п. п. во мн. ч. мы подсчитали частоту употребления каждого падежа с помощью данных корпусов современного и древнерусского языков. Тем самым мы проверили гипотезу о том, что причиной данного явления служит меньшая частота употребления вышеобозначенных падежей.

2. Корпусы и языковые исследования

Подходы к исследованию языка могут опираться на тексты. Большое количество текстов — корпус — служит для доказательства гипотезы о каких-либо языковых явлениях. На сегодняшний день корпусы используются в самых разных научных областях [McEnery, Hardie 2012].

С точки зрения лингвистики частотные данные, основанные на корпусах, предлагают объективное доказательство того, какие языко-

вые единицы, включая падежи, более частотны, а какие — менее (см. [Ляшевская, Шаров 2009]). Для этой цели наиболее часто обращаются к корпусам ruTenTen11 и Национальному корпусу русского языка (далее — НКРЯ, подробности см. ниже).

3. Анализ и результаты

В данном разделе анализируются частотные данные каждого падежа в ед. и во мн. ч. с синхронической и диахронической точки зрения, в качестве базы данных современного русского языка автором было выбрано два корпуса: ruTenTen11 в Sketch Engine и Основной корпус из НКРЯ. Для диахронического анализа использованы данные из подкорпуса древнерусского языка (далее — ДК) при НКРЯ.

3.1. Корпусы современного русского языка

Sketch Engine является многофункциональной программой, с помощью которой осуществляются создание своих корпусов, леммантизация, коркорданс и др. (см. [Kilgarriff et al. 2014]). В ней имеется веб-корпус ruTenTen11 размером в 14 553 856 113 словоупотреблений, что позволяет ученым исследовать частотные данные разных языковых единиц.

Функция Wordlist, содержащаяся в Sketch Engine, показывает частоту падежей в ruTenTen11. Ниже приведен результат анализа.

Таблица 2. **Доля падежей в ruTenTen11 (дата обращения к корпусу: 12. 03.2020)**

	Число	Падеж	Доля		Число	Падеж	Доля
1	Ед. ч.	Р. п.	25,0 %	7	Ед. ч.	Т. п.	5,2 %
2	Ед. ч.	И. п.	21,4 %	8	Мн. ч.	В. п.	3,4 %
3	Мн. ч.	Р. п.	11,6 %	9	Ед. ч.	Д. п.	3,2 %
4	Ед. ч.	В. п.	11,1 %	10	Мн. ч.	П. п.	1,8 %
5	Ед. ч.	П. п.	8,0 %	11	Мн. ч.	Т. п.	1,8 %
6	Мн. ч.	И. п.	6,1 %	12	Мн. ч.	Д. п.	1,4 %

Как указано в табл. 2, самые низкие ранги занимают д. п., т. п. и п. п. во мн. ч. (последовательно 1,8 %, 1,8 %, 1,4 %).

Далее нами был проведен подобный анализ с использованием частотных данных из сбалансированного подкорпуса (Основного корпуса, далее — ОК) при НКРЯ, и мы получили крайне схожий результат.

Таблица 3. **Доля падежей в ОК (дата обращения к корусу: 12.03.2020)**

	Число	Падеж	Доля		Число	Падеж	Доля
1	Ед. ч.	И. п.	17,4 %	7	Ед. ч.	Д. п.	6,9 %
2	Ед. ч.	В. п.	15,0 %	8	Мн. ч.	Р. п.	6,2 %
3	Ед. ч.	Р. п.	14,9 %	9	Ед. ч.	Т. п.	4,3 %
4	Мн. ч.	В. п.	12,5 %	10	Мн. ч.	П. п.	1,4 %
5	Мн. ч.	И. п.	11,2 %	11	Мн. ч.	Т. п.	1,2 %
6	Ед. ч.	П. п.	8,1 %	12	Мн. ч.	Д. п.	0,9 %

Как показано в табл. 3, в ОК также наиболее редко употребляются д. п., т. п. и п. п. во мн. ч. (последовательно 1,4 %, 1,2 %, 0,9 %). Кроме того, данный анализ был проведен в пяти корпусах размером в миллион словоупотреблений каждый, созданных автором на основе НКРЯ [Саяма 2018]. В итоге анализа зафиксированы те же результаты во всех пяти корпусах.

3.2. Корпус древнерусского языка

На сайте НКРЯ помещены разные подкорпусы, среди которых существует ДК общим размером в 573 252 слова. Здесь мы также провели анализ частоты каждого падежа.

Таблица 4. **Доля падежей в ДК (дата обращения к корпусу: 12.03.2020)**

	Число	Падеж	Доля		Число	Падеж	Доля
1	Ед. ч.	И. п.	23,3 %	7	Ед. ч.	П. п.	6,0 %
2	Ед. ч.	В. п.	19,2 %	8	Мн. ч.	И. п.	5,1 %
3	Ед. ч.	Р. п.	14,9 %	9	Мн. ч.	Р. п.	3,4 %
4	Ед. ч.	Д. п.	9,3 %	10	Мн. ч.	Д. п.	2,3 %
5	Ед. ч.	Т. п.	7,1 %	11	Мн. ч.	Т. п.	2,3 %
6	Мн. ч.	В. п.	6,1 %	12	Мн. ч.	П. п.	1,0 %

В ДК наблюдается тот же результат: д. п., т. п. и п. п. во мн. ч. имеют самые низкие среди падежей показатели по частоте (последовательно 2,3 %, 2,3 %, 1,0 %).

4. Выводы

В данной работе для анализа частотных данных были использованы веб-корпус (ruTenTen11), сбалансированный корпус (ОК при НКРЯ), исторический корпус (ДК), а также сравнительно маленькие по объему корпусы, созданные автором. Во всех этих корпусах количественно подтверждена гипотеза о том, что д. п., т. п. и п. п. в мн. ч. употребляются значительно реже по сравнению с другими падежами и в современном, и в древнерусском языках.

В древнерусском языке эти 3 падежа имели свои падежные формы по типам склонений, однако они были утеряны по причине малой частоты и в конце концов не сохранились. В итоге формы окончания 3-х падежей свелись к одной форме окончания жен. рода -ам, -ами, -ах (-ям, -ями, -ях), которые изначально отличались от других форм — муж. и ср. рода [Chino, 1970] (см. рис. 1).

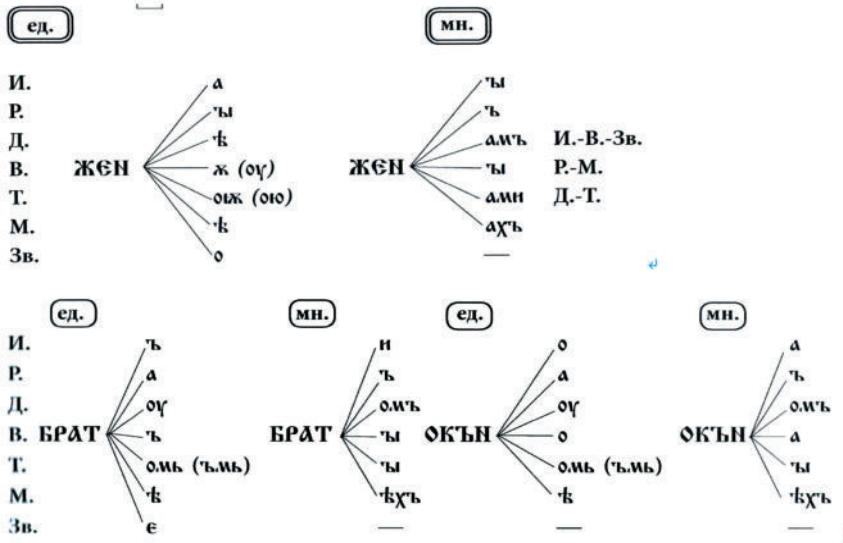


Рис. 1. **Пример склонения существительных во мн. ч. в древнерусском языке [Шулежкова 2018: 44–45]**

Как показано на рис. 1, д. п., т. п. и п. п. женского рода во мн. ч. не имеют формального сходства с другими падежными окончаниями (в отличие от муж. рода, в котором, например, окончание т. п. во мн. ч. совпадает с окончаниями других падежных форм).

Из результатов анализа данной работы получено количественное подтверждение гипотезы о корреляции частоты и падежных окончаний. Однако для более надежного подтверждения нашей гипотезы предстоит разъяснить немало вопросов. В дальнейшем данную тему необходимо обсуждать с диахронической точки зрения, ссылаясь на исторические изменения падежей других славянских языков, например, чешского и сербского. Также должны быть учтены такие факторы, как возможное фонетическое влияние на изменение падежных окончаний.

Литература

1. *Ляшевская О. Н., Шаров С. А.* (2009), Частотный словарь современного русского языка на материалах Национального корпуса русского языка. М.
2. *Саяма Г.* (2018), Влияние объема корпуса на определение наиболее часто употребляемых слов: Анализ частотных данных из пяти корпусов. Русский язык в научном освещении. № 34(1), с. 70–91.
3. *Шведова Н. Ю.* (ред.) (1980), Русская грамматика. Т. I. М.
4. *Шулежкова С. Г.* (2018), Старославянский язык, древнерусский язык и историческая грамматика русского языка. Опыт сопоставительного изучения: Учебно-методическое пособие. М.
5. *Чино Э .* (1970), Росиагомеисихэнка ниокэру икэитаи но бумпу ницуитэ -АМ, -АМИ, -АХ [Об алломорфах в склонении существительных русского языка -АМ, -АМИ, -АХ]. Росиагоросиабунгакукэнкю [Бюллетень Японской ассоциации русистов]. No. 4, pp. 70–75.
6. *Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.* (2014), The Sketch Engine: Ten years on, Lexicography. No. 1(1), pp. 7–36.
7. *McEnery T., Hardie A.* (2012), Corpus linguistics: Method, theory and practice. Cambridge, Tokyo.

References

1. *Chino E.* (1970), Roshiagomeishihenka niokeru ikeitai no bumpu nituite -AM, -AMI, -AH [About allomorphs in Russian case infrextion -AM, -AMI, -AH]. In: Roshiagoroshiabungakukenkyuu [Bulletin of the Japan Association for the Study of Russian Language and Literature]. No. 4, pp. 70–75.
2. *Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.* (2014), The Sketch Engine: Ten years on , Lexicography. No. 1(1), pp. 7–36.

3. *Ljashevskaja O. N., Sharov S. A.* (2009), Chastotnyj slovar' sovremennogo russkogo jazyka na materialah Nacional'nogo korpusa russkogo jazyka [Frequency Dictionary of the Modern Russian Language based on the Materials of Russian National Corpus]. Moscow.
4. *McEnery T., Hardie A.* (2012), Corpus linguistics: Method, theory and practice. Cambridge, Tokyo.
5. *Sayama G.* (2018), Vlijanie ob"joma korpusa na opredelenie naibolee chasto upotrebljaemyh slov: Analiz chastotnyh dannyh iz pjati korpusov [Influence of corpus size on the definition of most frequently words: Analysis of frequency data from five corpora]. In: Russkij jazyk v nauchnom osveshchenii [Russian language in academic field]. No. 34(1), pp. 70–91.
6. *Shvedova N. Ju.* (ed.) (1980), Russkaja grammatika [Russian grammar]. Vol. I. Moscow.
7. *Shulezhkova S. G.* (2018), Staroslavjanskij jazyk, drevnerusskij jazyk i istoricheskaja grammatika russkogo jazyka. Opyt sopostavitel'nogo izuchenija: Uchebno-metodicheskoe posobie [Old Slavonic language, old Russian language and historical grammar of the Russian language. Comparative experience of studying: Course materials]. Moscow.

Саяма Гота
Университет София (Япония)
Sayama Gota
Sophia University (Japan)
E-mail: sayamagouta44@gmail.com

Ю. Тао, В. П. Захаров

Yu. Tao, V. P. Zakharov

КОРПУСНЫЙ АНАЛИЗ АНТРОПОМОРФНЫХ МЕТАФОР В КИТАЙСКОМ И РУССКОМ ПОЛИТИЧЕСКОМ ДИСКУРСЕ ПО ТЕМЕ «ОДИН ПОЯС — ОДИН ПУТЬ»

CORPUS ANALYSIS OF ANTHROPOMORPHIC METAPHORS IN THE CHINESE AND RUSSIAN POLITICAL DISCOURSE ON THE TOPIC "ONE BELT ONE ROAD"

Аннотация. Антропоморфные метафоры являются типичными для политического дискурса в русском и китайском языках. В общем они похожи, однако в русских текстах они приобретают черты, не характерные для китайского языка, а именно – они длиннее и чаще встречаются.
Ключевые слова. Антропоморфные метафоры, китайский язык, русский язык, корпусный анализ.

Abstract. Anthropomorphic metaphors are stylistic devices typical of political discourse. However, in Russian texts, they acquire features that are not characteristic of the Chinese language, namely, they lengthen, inheriting an increase in number and resonance in Russian. We have studied functioning of anthropomorphic metaphors on the topic "One Belt and One Road". It was shown that in general they are similar and stable, but there are also differences, which is declared to be the influence of culture, religion and the way of thinking of the two countries and people.
Keywords. Anthropomorphic metaphors, Chinese, Russian, corpus analysis.

1. Введение

В 2013 г. Китайская Народная Республика представила мировому сообществу проект «Один пояс — один путь», понимаемый как продолжение и развитие древнего Шелкового пути. Проект предусматривает стимулирование экономического развития вовлеченных стран и укрепление культурных обменов и связей во всех областях между разными странами. Все последующие годы проект активно обсуждается в разных странах. Обращает на себя внимание тот факт, что в дискурсе «Один пояс — один путь» содержится большое количество концептуальных метафор. Представляется интересным исследовать язык публикаций по теме проекта в русской и китайской прессе. Актуальность данной статьи определяется важностью исследования закономерностей политического дискурса китайского и русского языков, что должно способствовать лучшему пониманию друг друга и сотрудничеству между двумя странами.

Метафора в политическом дискурсе как эффективное средство воздействия на читателей и слушателей, как многогранное языковое явление изучена подробно и глубоко [Charteris-Black 2005]. В широком смысле метафора является синонимом тропа. Использование концептуальной метафоры в различных типах дискурса помогает упростить понимание сложных событий или явлений. Анализ метафорики может проводиться с точки зрения стилистики, прагматики, когнитивистики и т. д. Исследование метафоры в российском политическом дискурсе, аналогично исследованиям на Западе, пережило заметный процесс перехода от стилистики к когнитивной перспективе [Чудинов 2001; Баранов, Караулов 2018]. Использованию концептуальных метафор в политических текстах посвящены и статьи в крупных китайских журналах.

Мы исследовали концептуальные метафоры в политическом дискурсе китайского и русского языков по теме «Один пояс — один путь» на основе корпусов русской и китайской периодической печати. Данный анализ совмещает когнитивную лингвистику и корпусный лингвистический подход.

В литературе выделяются различные типы метафор (антропоморфная, бытовая, военная, ботанико-земледельческая, географическая и др.). Мы исходим из классификации А. П. Чудинова [Чудинов 2001], который выделяет четыре широких типа концептуальной метафоры в политическом дискурсе: антропоморфную, социоморфную, природоморфную и артефактную.

Мы остановились на антропоморфных метафорах, в которых человек представляет мир, перенося на него принципы структурирования форм собственного организма и личности. «Создаваемая человеком метафорическая картина политического мира в значительной степени антропоцентрична: как Бог создал человека по своему образу, так и человек метафорически создает (концептуализирует) политическую действительность в виде некоего своего подобия» [Чудинов 2001: 23].

2. Методика исследования

В данной работе мы основываемся на теории концептуальных метафор Д. Лакоффа и М. Джонсон [Лакофф, Джонсон 1990] и теории метафорического моделирования, предложенной А. П. Чудиновым [Чудинов 2001]. Процесс исследования разделен на четыре этапа.

А. Создание русского и китайского корпусов по теме «Один пояс — один путь» по текстам «Российской газеты» и газеты «Женьминь жибао»; объемы корпусов см. в табл. 1.

Таблица 1. **Объемы корпусов**

	Количество слов	Количество текстов
Жэньминь жибао	92 373	153
Российская газета	94 071	152

Б. Формирование словаря ключевых слов по теме «Человек как живое существо». Словарь формировался на базе «Комплексного учебного словаря», включающего такие разделы, как «Организм человека», «Физические возможности и состояние», «Здоровье и самочувствие», «Внешний облик», «Фазы жизни» и др. [Морковкин 2004]. Сюда же было добавлено небольшое число терминов из раздела «Животный мир». Объем словаря — около 2000 слов. На основе словаря для русского языка был составлен аналогичный словарь для китайского языка.

В. Проверка наличия и подсчет частот антропоморфных ключевых слов по обоим корпусам. Выявление степени метафоричности китайского и русского дискурсов.

Г. Сравнительный анализ антропоморфных метафор в политическом дискурсе по теме «Один пояс и один путь» в китайском и русском языках.

Остановимся на двух последних этапах.

3. Сравнительный анализ антропоморфных метафор

3.1 Частотный анализ антропоморфных метафор в русском и китайском политическом дискурсе

Как уже говорилось, на втором этапе были составлены словари ключевых слов по теме «Человек как живое существо». Далее по всем этим словам проводился поиск в обоих корпусах и подсчитывалась их частота, а именно количество лексем и количество словоупотреблений для каждой лексемы. Анализ контекстов употребления показал, что в большинстве случаев мы имеем дело с метафорическим использованием этих слов.

Для определения пропорции распределения концептуальных метафор в двух дискурсах мы использовали понятие «резонанс», введен-

ное Чартерисом-Блэком [Charteris-Black 2005]. Этот показатель основан на подсчете количества метафорических ключевых слов из сферы-источника (в нашем случае «Человек как живое существо») в корпусе. Метафорические ключевые слова (МКС) — это слова, относящиеся не собственно к теме статьи, а к сфере-источнику, в нашем случае, к организму человека (иногда животного).

Метод расчета показателя заключается в перемножении количества метафорических ключевых слов (лексем) на число словоупотреблений (отсюда и название показателя — «резонанс»).

Например, в русском корпусе встретились ключевые слова «крыло» и «птичий полет», относящиеся к сфере-источнику «живое существо», которые затем в одной или нескольких метафорах встречаются в разных текстах корпуса. При этом «крыло» встречается два раза, а «птичий полет» — один раз, тогда значение показателя «резонанс» высчитывается как $2 \cdot (2 + 1) = 6$.

Таким образом вычисляется резонанс для различных типов концептуальных антропоморфных метафор в политическом дискурсе по теме «Один пояс и один путь» в «Жэньминь жибао» и «Российской газете». Суммарные результаты показаны в табл. 2.

Таблица 2. **Антропоморфные метафоры в «Жэньминь жибао» и «Российской газете»**

Корпус	Общее число МКС	Общая частота МКС	Резонанс	ipm МКС	Число употр. на одно МКС
Российская газета	63	433	27 279	4602,9	6,87
Жэньминь жибао	31	296	9176	3204,2	9,55

Как видно из приведенных цифр, резонанс, ipm метафорических ключевых слов и число антропоморфных метафор в «Российской газете» гораздо больше, чем в «Жэньминь жибао». В то же время одни и те же МКС в китайском политическом дискурсе употребляются чаще.

3.2. Сравнение метафор в двух языках

Рассмотрим отдельные антропоморфные метафоры в текстах «Российской газеты» и «Жэньминь жибао», а именно физиологические метафоры (табл. 3).

Таблица 3. **Сравнение антропоморфных физиологических метафор в «Жэньминь жибао» и «Российской газете»**

Фрейм	Жэньминь жибао	Российская газета
«Орган тела» (объект)	горло, скелет, кровь, артерия, глаз…	рука, душа, плечо, лицо, артерия, слух…
Физиологические особенности	пробуждение, астения, слабость…	здоровье, иммунитет, головная боль…
Функции органа тела или организма	держать, идти рука об руку, схватить, шагнуть, поднять голову…	идти на двух ногах / по пути, брать в руки, далеко уйти вперед…

Физиологические метафоры в китайском и российском политическом дискурсе призваны характеризовать стабильность, так как они основываются на первоначальном физическом опыте, живущем в человеческом сознании. В то же время под влиянием национальной культуры, религии, образа мышления, возможно, языкового узуса физиологические метафоры в китайском и российском политическом дискурсе различаются.

Во фрейме органа тела в китайском политическом дискурсе наиболее часто встречаются слова «артерия» и «скелет», которые используются как метафоры транспортных потоков или маршрутов. В российском политическом дискурсе чаще всего встречаются слова «рука», «плечо» и «душа».

Во фрейме физиологической особенности в китайском политическом дискурсе для описания экономической ситуации часто используются слова «пробуждение», «астения» и «слабость», тогда как в российском политическом дискурсе с экономикой метафорически связываются такие слова, как «иммунитет», «жизнеспособность», «головная боль», но в целом они встречаются заметно реже.

Метафоры, относящиеся к фрейму функции органа, в китайском политическом дискурсе гораздо многочисленнее и богаче, чем в российском дискурсе.

Другой тип — метафоры родства — и в русских, и в китайских текстах опираются на фрейм «Семья», однако в китайском языке правильнее выделить отдельный фрейм «Родственники». Здесь можно говорить о национальной специфике китайских концептов родства.

Помимо упомянутых выше, в политических текстах на русском языке присутствуют антропоморфные метафоры, связанные с поведением человека: «кормить», «выполнить обещание», «нарушить слова» и т. п., которых нет в китайском политическом дискурсе.

В итоге можно сказать, что физиологические метафоры в китайском и российском политическом дискурсе имеют сходство, но есть и различия.

Приведем пример различия в метафорах родства в двух языках.

(1) 做“一带一路”上的亲朋好友。(«Жэньминь жибао», 8 мая 2017 г.)
Перевод: Будьте друзьями и ***родственниками*** по инициативе [проекта] «Один пояс и один путь» (пер. с кит. авт.).

(2) Афганистан, будучи страной-наблюдателем ШОС, является важным членом ШОСовской ***семьи*** («Российская газета», 18 июля 2018 г., федеральный выпуск № 155(7618)).

Из примеров (1) и (2) мы видим, что фактически в одинаковых по содержанию высказываниях метафора в русском тексте отсылает читателя к понятию «семья», тогда как китайская метафора оперирует понятием «родственник».

4. Заключение

Антропоморфная метафора является типичным стилистическим приемом в политическом дискурсе. Было изучено функционирование антропоморфных метафор по теме «Один пояс — один путь» на материале псевдопараллельного корпуса статей в газетах «Жэньминь жибао» и «Российская газета». Статический анализ показал, что в русскоязычной прессе антропоморфных метафор гораздо больше.

Использование метафор, связанных с человеком, в китайском и российском политическом дискурсе показывает, что люди считают себя «прототипом» познания мира и поэтому рассматривают человеческое тело и физиологию как стандарт мышления и даже как важную «шкалу» для измерения опыта политического мира.

Благодарности

Исследование поддержано грантом Бюро национального фонда социальных и гуманитарных наук Китайской Народной Республики

№ 13BYY026 «Исследование перевода тематических текстов на основе параллельного корпуса русского и китайского языков».

Литература

1. *Лакофф Дж., Джонсон М.* (1990), Метафоры, которыми мы живем. Теория метафоры. М.: Прогресс.
2. *Морковкин В. В.* (2004), Комплексный учебный словарь. Лексическая основа русского языка. М.: АСТ.
3. *Чудинов А. П.* (2001), Россия в метафорическом зеркале: Когнитивное исследование политической метафоры (1991–2000). Екатеринбург: УрГПУ.
4. *Charteris-Black J.* (2005), Politicians and Rhetoric. In: The Persuasive Power of Metaphor. New York: Palgrave Macmillan.
5. *Musolff A.* (2004). Metaphor and political Discourse: Analogical Reasoning in Debates about Europe. Basingstoke: Palgrave Macmillan.

References

1. *Charteris-Black J.* (2005), Politicians and Rhetoric. In: The Persuasive Power of Metaphor. New York: Palgrave Macmillan.
2. *Chudinov A. P.* (2001), Rossiya v metaforicheskom zerkale: Kognitivnoe issledovanie politicheskoj metafory (1991–2000). [Cognitive exploration of political metaphor]. Ekaterinburg: UrGPU.
3. *Lakoff J., Johnson M.* (1990), Metafory, kotorymi my zhivem [Metaphors we live by.]. In: Teoriya metafory [Metaphor theory]. Moscow: Progress.
4. *Morkovkin V. V.* (2004), Kompleksnyy uchebnyy slovar'. Leksicheskaya osnova russkogo yazyka [Comprehensive educational dictionary. Lexical basis of the Russian language]. Moscow: AST.
5. *Musolff A.* (2004), Metaphor and political Discourse: Analogical Reasoning in Debates about Europe. Basingstoke: Palgrave Macmillan.

Тао Юань
Юго-Восточный университет (Китай)
Tao Yuan
South-East University (China)
E-mail: taoyuanhuahua@126.com

Захаров Виктор Павлович
Санкт-Петербургский государственный университет (Россия)
Zakharov Victor
Saint-Petersburg State University (Russia)
E-mail: v.zakharov@spbu.ru

М. Ю. Товкес

M. Yu. Tovkes

ИССЛЕДОВАНИЕ ГЕНДЕРНЫХ СТЕРЕОТИПОВ ПОЛИТИЧЕСКОГО ДИСКУРСА С ПРИМЕНЕНИЕМ КОРПУСНЫХ ТЕХНОЛОГИЙ

GENDER STEREOTYPES RESEARCH IN THE POLITICAL DISCOURCE WITH THE USE OF CORPUS TECHNOLOGIES

Аннотация. Исследование опирается на теорию и методологию гендерного конструктивизма. Эмпирическим материалом послужили 7920 записей русскоязычного сегмента «Твиттера», содержащие антропонимы, которые маркируют известных женщин-политиков. Методы корпусной лингвистики позволили выявить существенные характеристики, классифицировать их и составить стереотипный образ женщины-политика. Таким образом, в русскоязычном сегменте «Твиттера» акцентируется внимание преимущественно на фактических характеристиках образа женщины-политика.
Ключевые слова. Гендерный стереотип, гендерная идентичность, корпусная лингвистика, блог-коммуникация, «Твиттер».

Abstract. The relevance of studying the collective image of a female political leader is determined by the transformation of gender stereotypes reflected in the language. This research is based on the theory and methodology of gender constructivism, according to which gender is interpreted as a product of discourse and constructed by the communicative context. The sample includes 7920 posts in the Russian-speaking segment of Twitter containing anthroponyms and marking famous women politicians: Angela Merkel, Valentina Matvienko, Maya Sandu, Marine Le Pen, Svetlana Tikhanovskaya, Theresa May, Hillary Clinton, Yulia Tymoshenko. Methods of corpus linguistic made it possible to identify the most significant characteristics expressed in communicative contexts, to classify them and to create a stereotypical image of a female politician. The results demonstrate that in the Russian-speaking segment of Twitter there is a tendency to focus mainly on factual information, while personal qualities are considered to be less relevant.
Keywords. Gender stereotype, gender identity, corpus linguistics, blog communication, Twitter.

1. Программа исследования

В настоящее время блоги как способ электронной коммуникации активно используются в качестве трансляторов социально независимой информации благодаря свободному регулированию контента [Kenix 2009; Sánchez-Villar et al. 2017]. Подход к изучению гендерных стереотипов на основе комплексного анализа индивидуальных реакций блогеров представляется перспективным. В данной работе гендерные стереотипы понимаются как «структурированный набор представлений о персональных характеристиках мужчин и женщин» [Рябова 2008: 27].

Методология исследования опирается на гендерный конструктивизм, который осуществляет функционально-прагматический подход к языку как важнейшему творческому ресурсу и объясняет лингвистический выбор репрезентацией целевого, значимого для коммуникативного контекста аспекта гендерной идентификации [Гриценко и др. 2011: 30–31].

Использование корпусных технологий, в свою очередь, «гарантирует типичность данных и обеспечивает полноту представления всего спектра языковых явлений», а также позволяет представить языковые данные «в своей естественной контекстной форме, что создает возможность их всестороннего и объективного изучения» [Захаров и др. 2020: 12].

Цель исследования — выявить стереотипные черты образа современной женщины-политика, маркируемые участниками блог-коммуникации.

Категория «образ женщины-политика» представляет собой «результат и идеальную форму отражения политического субъекта и его характеристик в сознании человека, возникающую в условиях общественно-исторической и политической практики» [Бахтеева 2015: 100]. На восприятие образа могут оказывать влияние факторы объективного порядка (географическое положение места проживания и осуществления деятельности; временные условия и др.); факторы субъективного порядка (состояние мировой политики, политическое устройство конкретного государства, менталитет и национальные особенности региона и др.); а также субъектные факторы (влияние представителей различных сфер деятельности на восприятие женщины-политика) [Бахтеева 2015].

Эмпирическим материалом для данного исследования стал корпус текстов, включающий в себя 7920 записей (179 724 слова) в микроблоге «Твиттер» и составленный по ключевым словам *Светлана Тихановская* (3670 твитов), *Майя Санду* (1631 твит), *Тереза Мэй* (1274 твита), *Ангела Меркель* (625 твитов), *Валентина Матвиенко* (266 твитов), *Юлия Тимошенко* (178 твитов), *Хиллари Клитон* (172 твита), *Марин Ле Пен* (104 твита). Ключевыми словами в данном случае являются антропонимы, маркирующие деятельность известных женщин-политиков.

Сбор твитов производился по вышеуказанным ключевым словам и с помощью поисковых запросов *since 2018-04-15 to 2018-05-15*; *since 2020-08-01 to 2020-08-31*; *since 2020-11-01 to 2020-11-30*, язык запросов — русский. Временной период обусловлен как обсуждаемой

деятельностью женщин-политиков, так и информационными поводами — выборами глав государств в Республике Беларусь, Республике Молдова. Также следует отметить, что при анализе твитов не учитывались экстралингвистические характеристики авторов записей, поскольку значительная часть пользователей указывала никнеймы вместо настоящего имени, вследствие чего нельзя идентифицировать объективные социальные характеристики всех авторов твитов.

Вместе с тем для анализа антропонимов *Светлана Тихановская*, *Майя Санду*, *Тереза Мэй* используются функции Shuffle и Random Sample системы Sketch Engine, с помощью которых было отобрано по 600 твитов. Данные функции позволили избежать повтора записей, а также проанализировать твиты с различными антропонимами приблизительно в равных пропорциях.

Таким образом, итоговый объем выборки составляет 3145 твитов. В результате исследования будут выявлены наиболее частотно употребляемые и релевантные характеристики собирательного образа женщины-политического лидера.

Используемое программное обеспечение — AntConc, Gephi, Sketch Engine.

2. Анализ корпуса текстов

При анализе корпуса текстов был построен конкорданс для ключевых слов с помощью корпусного менеджера AntConc — «список найденных примеров (вхождений) нужного слова в минимальном контексте» [Копотев 2014: 120], который позволил выявить наиболее частотно употребляемые и значимые характеристики женщин-политиков.

Так, пользователи микроблога «Твиттер» отмечают должность женщины-политика (1015 твитов; 32,3 %). В твитах встречается как официальное наименование занимаемой должности (*Председатель Совета Федерации*[1], *Федеральный канцлер Германии*), так и разговорные варианты (*британская премьерша*).

Кроме того, в 713 случаях (22,7 %) пользователи акцентируют внимание на том, что указанные женщины являются лидерами политических партий или баллотируются на пост главы государства. Например: *Лидер «Батькивщина» Юлия Тимошенко выразила уверенность, что президент Порошенко потерял поддержку украинского общества*;

[1] Здесь и далее сохранено правописание источника.

Оппозиционный кандидат в президенты Белоруссии Светлана Тихановская призвала Европу не признавать победу Александра Лукашенко на президентских выборах. Отмечается, что предвыборная программа таких женщин-кандидатов отличается от действующего политического режима. В связи с этим употребляются следующие словосочетания: *кандидат от оппозиции*, *кандидат против коррупции*, *лидер проевропейской оппозиции*, *оппозиционный кандидат*, *лидер ультраправых* и т. д. Также блогеры отмечают, что женщина-политик будет занимать пост главы государства впервые в истории страны: *Майя Санду победила на выборах в Молдове. Президентом страны впервые станет женщина*.

Вместе с тем в твитах фиксируются рабочие визиты женщин-политиков, участие в пресс-конференциях, а также общение с коллегами или оппонентами (695 твитов; 22,1 %). Для визуализации данной тенденции была использована функция Word Sketch системы Sketch Engine. Результат на примере антропонима *Ангела Меркель* продемонстрирован на рис. 1. На изображении зафиксированы имена политиков, с которыми происходит взаимодействие — В. В. Путин, Д. Трамп, Э. Макрон и т. д., а лексемы *встретиться*, *приехать*, *сочи*, *берлин*, *разговор* маркируют рабочие поездки, встречи, телефонные разговоры.

Рис. 1. **Функция Word Sketch для антропонима** *Ангела Меркель*

Также в корпусе частотно цитируются высказывания указанных женщин-политиков (504 твита; 16 %) — *Тереза Мэй: «Мы действовали в Сирии не по приказу США»*.

В меньшей степени пользователи употребляют оценочную лексику для передачи своего отношения к деятельности указанных женщин-политиков (125 твитов; 4 %). Среди положительных маркеров отмечаются: *смелая красавица*; *беспристрастная*; *прагматичный политик с огромным стажем*; *прогрессивна*; *здравомыслящий человек*, *лицо перемен* и т. д. Для вербализации одобрительного отношения в целом употребляются лексемы *молодец*, *достойный политик* и т. д. В то же время деятельность Т. Мэй, Х. Клинтон сопровождается преимущественно негативной коннотацией. Пользователи отмечают хитрость, лицемерие, недальновидность, лживость, глупость: *Гадкая, злая и лживая Мэй…*; *Хилари Клинтон — это позор Америки*.

В ходе исследования было установлено, что в твитах, маркирующих вышеуказанных женщин-политиков, наблюдается частотное употребление так называемых глаголов говорения (1749 твитов; 55,6 %). Для визуализации данного аспекта был построен граф с помощью графопостроителя Gephi. Результаты представлены на рис. 2.

Рис. 2. **Визуализация глагольного ряда в корпусе текстов**

Данный граф отражает частотность употребления глаголов и их связь с указанными антропонимами; он включает в себя 59 узлов и 614 ребер. Наиболее релевантными узлами графа являются глаголы *выступила*, *заявила* и *призвала*. Данные глаголы преимущественно передают публичные заявления женщин-политиков. Релевантными узлами графа также являются *прокомментировала*, *считает*, *выразила*, *поздравила*, *назвала*, *высказалась* и т. д. Указанные глаголы маркируют способность женщин-политиков открыто высказывать свою точку зрения и выражать свои мысли. Частотное употребление глагола *поздравила* обусловлено информационными поводами — вступление в должность коллег-политиков, памятные события и т. д. Например, *Валентина Матвиенко поздравила Владимира Путина с убедительной победой на выборах президента России; Тереза Мэй поздравила принца Уильяма и Кейт Миддлтон с рождением сына* и т. д.

Интересно также отметить глагольный ряд, характеризующий деятельность С. Тихановской. Так, употребление конструкций *записала/выпустила видеообращение*, *запустила видео* обусловлено экстралингвистическими факторами, в частности, возрастающей ролью интернет-коммуникации и трансляцией информации в сети интернет.

3. Результаты исследования

По результатам исследования можно сделать следующие выводы:

- частотность употребления антропонимов *Светлана Тихановская*, *Майя Санду*, *Тереза Мэй* обусловлена информационными поводами — выборами на пост президента страны в Республике Беларусь, Республике Молдова, а также инцидентом, произошедшим с С. и Ю. Скрипалями в Великобритании в 2018 году, боевыми действиями в Сирии и т. д.;
- в корпусе текстов преобладают твиты, содержащие фактическую информацию: занимаемая должность женщины-политика (1015 твитов; 32,3 %), принадлежность к политическим партиям, преимущественно оппозиционным (713 твитов; 22,7 %), взаимодействие с коллегами-политиками (695 твитов; 22,1 %), а также цитирование указанных женщин-политиков (504 твита; 16 %);
- менее частотными в корпусе текстов являются записи, маркирующие личностные качества женщин-политиков и оценку их деятельности. Так, положительно оценивается способность принимать взвешенные решения, стремление к переменам. Не-

гативной коннотацией обладают записи о конкретных женщинах-политиках;
- частотное употребление так называемых глаголов говорения свидетельствует о публичном характере деятельности данных политических лидеров.

Таким образом, в русскоязычном сегменте «Твиттера» при обсуждении деятельности женщин-политиков акцентируется внимание преимущественно на фактической информации, что может свидетельствовать о том, что личностные качества не являются релевантными при конструировании стереотипного образа женщины-политического лидера в данном дискурсе.

Литература

1. *Бахтеева Е. Г.* (2015), Перспективы трансформации образа женщины-политика в общественном сознании россиян (на примере опросов жителей г. Саратов и г. Москва, 2011 г.). Известия Саратовского университета. Новая серия. Серия Социология. Политология. Вып. 1, т. 15, с. 100–104.
2. *Гриценко Е. С., Сергеева М. В., Лалетина А. О., Бодрова А. А., Дуняшева Л. Г.* (2011), Гендер в британской и американской лингвокультурах: монография. М.: ФЛИНТА: Наука.
3. *Захаров В. П., Богданова С. Ю.* (2020), Корпусная лингвистика: учебник. 3-е изд., перераб. СПб.: Изд-во С.-Петерб. ун-та.
4. *Копотев М. В.* (2014), Введение в корпусную лингвистику. Прага: Animedia Company.
5. *Рябова Т. Б.* (2008), Пол власти: гендерные стереотипы в современной российской политике. Иваново: Иван. гос. ун-т.
6. *Kenix J. L.* (2009), Blogs as Alternative. In: Journal of Computer-Mediated Communication. Vol. 14(4), pp. 790–822.
7. *Sánchez-Villar J., Bigné E., Aldás-Manzano J.* (2017), Blog Influence and Political Activism: An Emerging and Integrative Model. In: Spanish Journal of Marketing — ESIC. Vol. 21, pp. 102–116.

References

1. *Bahteeva E. G.* (2015), Perspektivy transformacii obraza zhenshchiny-politika v obshchestvennom soznanii rossiyan (na primere oprosov zhitelej g. Saratov i g. Moskva, 2011 g.) [Prospects of Transformation of the Image of Woman Politician in Public Consciousness of Russians (on the Example of Polls of Residents of Saratov and Moscow, 2011)]. In: Izvestiya Saratovskogo universiteta. Novaya seriya. Seriya Sociologiya. Politologiya [Izvestiya of Saratov University. Series: Sociology. Politology]. Issue 1, vol. 15, pp. 100–104.

2. *Gricenko E. S., Sergeeva M. V., Laletina A. O., Bodrova A. A., Dunyasheva L. G.* (2011), Gender v britanskoj i amerikanskoj lingvokul'turah: monografiya [Gender in British and American Linguacultures: monograph]. Moscow: FLINTA: Nauka.
3. *Kenix J. L.* (2009), Blogs as Alternative. In: Journal of Computer-Mediated Communication. Vol. 14(4), pp. 790–822.
4. *Kopotev M. V.* (2014), Vvedenie v korpusnuyu lingvistiku [Introduction to Corpus Linguistics]. Prague: Animedia Company.
5. *Ryabova T. B.* (2008), Pol vlasti: gendernye stereotipy v sovremennoj rossijskoj politike [Gender of Power: Gender stereotypes in Modern Russian Politics]. Ivanovo: Ivanovo State University.
6. *Sánchez-Villar J., Bigné E., Aldás-Manzano J.* (2017), Blog Influence and Political Activism: An Emerging and Integrative Model. In: Spanish Journal of Marketing — ESIC. Vol. 21, pp. 102–116.
7. *Zaharov V. P., Bogdanova S. Yu.* (2020), Korpusnaya lingvistika: uchebnik [Corpus Linguistics: Textbook]. Saint Petersburg: St. Petersburg University press.

Товкес Мария Юрьевна
Национальный исследовательский университет
«Высшая школа экономики» (Россия)
Tovkes Maria
HSE University (Russia)
E-mail: tovkes.m@yandex.ru

М.В.Хохлова, И.Д.Мамаев

M.V.Khokhlova, I.D.Mamaev

РАЗРАБОТКА БАЗЫ ДАННЫХ КОЛЛОКАЦИЙ: ОБЗОР ЗОЛОТОГО СТАНДАРТА НА ПРИМЕРЕ АТРИБУТИВНЫХ СЛОВОСОЧЕТАНИЙ[1]

BUILDING RUSSIAN COLLOCATIONS DATABASE: A CASE STUDY OF GOLD STANDARD FOR ADJ-NOUN COLLOCATIONS

Аннотация. Исследование лексической сочетаемости представляет собой важную проблему современной лингвистики. Статья посвящена разработке базы данных коллокаций для русского языка, которая включает словосочетания из словарей, а также примеры из корпусов текстов, которые были получены при помощи автоматических методов. Описываются основные принципы, которые лежат в основе данного ресурса, на примере отображения информации для атрибутивных словосочетаний, которые были извлечены из ряда словарей русского языка.
Ключевые слова. Коллокации, база данных, словари, статистические методы, русский язык.

Abstract. The study of collocability is an important task and is still highly relevant in linguistics. The present paper deals building the Russian Collocations Database which includes both verified or dictionary collocations, as well as data from text corpora extracted automatically. The authors describe adj-noun collocations that can be found in the database and in what number of dictionaries they were registered. The paper also discusses methods that were used for collocation visualization and presents the results of their evaluation against the collected data from verified resources.
Keywords. Collocations, database, dictionaries, statistical methods, Russian language.

1. Введение

Исследование лексической сочетаемости неразрывно связано с корпусами текстов, которые дают возможность изучить примеры на большом материале. Статистические методы позволяют автоматически извлечь словосочетания на основе размеченных коллекций текстов. Однако, несмотря на то, что в настоящий момент создаются разнообразные онлайн-ресурсы, традиционные словари с их тщательным подбором материала остаются важным источником сведений о лексической сочетаемости. Статья посвящена разработке базы данных сочетаемости, в которую вошли словосочетания двух типов: статистические и лексические. Первая группа представлена колло-

[1] Исследование выполнено за счет гранта Российского научного фонда (проект № 19-78-00091).

кациями, которые были получены на основе корпусов текстов, в том числе при помощи статистических методов, в то время как лексические (или словарные) коллокации включают те, которые были зафиксированы в словарях русского языка и представляют собой так называемый золотой стандарт. В статье мы обратимся именно ко второму типу и проиллюстрируем принципы создаваемого ресурса на примере атрибутивных словосочетаний, построенных по модели «прилагательное + существительное».

2. Обзор лексикографических проектов

Для русского языка существует несколько лексикографических справочников (преимущественно в печатном виде), в которых представлена информация о сочетаемости. Прежде всего это толковые словари, например, БАС [Большой академический словарь русского языка 2004–], МАС [Словарь русского языка 1999] и БТС [Большой толковый словарь русского языка 1998], в которых устойчивые словосочетания выделены в отдельные части словарных статей или получили специальную разметку.

Еще одну группу лексикографических источников представляют специализированные словари. Уникальным проектом является Толково-комбинаторный словарь русского языка [Мельчук, Жолковский 1984], в котором устойчивая сочетаемость описана при помощи лексических функций. Словарь [Регинина, Тюрина, Широкова 1980] предназначен для иностранных учащихся, в нем представлено около 3 тыс. устойчивых словосочетаний. Словарь устойчивых глагольно-именных сочетаний русского языка [Дерибас 1983] также нацелен на изучающих русский язык и содержит более 5 тыс. словосочетаний, большинство из которых представлено биграммами. Словарь усилительных словосочетаний русского и английского языков [Убин 1987] имеет две части (прямую и реверсивную, т. е. словарные статьи представлены для главных и зависимых слов) и содержит более 10 тыс. единиц. Словарь коллокаций [Борисова 1995] был первым и единственным проектом, в название которого вынесено понятие, связанное с ограниченной сочетаемостью. В нем приводятся словосочетания для 512 заголовочных единиц, а также имеется англо-русский список ключевых слов. В словаре сочетаемости слов русского языка [Денисов, Морковкин 2002] представлены 2500 словарных статей для существительных, глаголов и прилагательных. Авторами делается различие

между лексической и семантической сочетаемостью, а также дается определение синтаксической сочетаемости как некоторой валентностной рамки. Существует уникальный лексикографический проект под руководством Ю. Д. Апресяна по созданию активного словаря русского языка [Апресян 2014–2017], который включает обширную информацию о сочетаемости, отраженную отдельно в словарных статьях.

Также доступен ряд проектов, которые нацелены на описание сочетаемости и представлены в виде базы данных или в виде онлайн-систем. Если говорить об электронных ресурсах, то для русского языка НКРЯ [Национальный корпус русского языка] предоставляет ряд инструментов (n-грамм поиск со статистической оценкой, списки устойчивых слов и словосочетаний, лексические графы), а также на его основе были разработаны словари (например, [Кустова 2008]), в которых представлена ограниченная сочетаемость.

Среди остальных проектов для русского языка можно назвать FrameBank [Lyashevskaya 2010], который включает описание валентностных рамок для глаголов и конструкций. Ресурс Collocations, Colligations, Constructions [Kopotev et al. 2015] предоставляет информацию о сочетаемости на основе корпусов. В нем представлено два интерфейса в зависимости от подготовки пользователей. Еще одним проектом в данном направлении является семантический словарь Lexicograph [База данных Lexicograph].

3. Описание базы данных

В ходе исследования были рассмотрены следующие источники:

1) толковые словари (МАС и БТС);
2) словари сочетаемости [Борисова 1995; Мельчук, Жолковский 1984; Регинина, Тюрина, Широкова 1980; Убин 1987];
3) электронный словарь [Кустова 2008].

На основе данных словарей было извлечено около 17 тыс. атрибутивных словосочетаний.

При анализе толковых словарей нами были извлечены данные из заромбовой части словарных статей, так как именно в ней представлена информация об ограниченной сочетаемости (о фразеологических оборотах и об устойчивых сочетаниях). В других словарях коллокации были представлены более простым образом: списком или отдельными словарными статьями.

С проектом можно ознакомиться на сайте «База данных коллокаций»[2]. Одноименная база данных была реализована с помощью платформы для создания онлайн-ресурсов, и в ней доступны следующие возможности:

- поиск коллокаций по главному (опорному) слову;
- поиск коллокаций по коллокату;
- визуализация;
- просмотр каждой отдельной коллокации и ее лингвостатистических характеристик;
- просмотр ссылок на печатные и электронные издания словарей;
- просмотр ссылок на корпусы текстов.

Поскольку база данных может быть востребована разными группами пользователей, было решено разработать два интерфейса, которые подразумевают лингвистический и статистический поиск.

Первый тип интерфейса дает возможность просмотреть коллокации, выданные по запросу для главного слова или для коллоката. Результаты содержат список словосочетаний, в котором представлена следующая лингвистическая информация: определение лемм из Викисловаря, тип синтаксической структуры (например, прилагательное + существительное, глагол + существительное, наречие + прилагательное и пр.), ссылка на пример употребления в НКРЯ, наличие/отсутствие словосочетания в корпусах СинТагРус и Тайга, пересечение с другими словарными коллокациями. Нами также был введен словарный индекс — количество словарей, в которых зафиксирована коллокация. Соответственно, чем он больше (максимальное значение в случае атрибутивных словосочетаний равно 6), тем больше вероятность того, что словосочетание является воспроизводимым в речи и, как следствие, его необходимо запомнить (если речь идет о пользователе, изучающем русский язык). В табл. 1 представлено количество коллокаций, соответствующих каждому словарному индексу.

Таким образом, при рассмотрении словосочетаний, которые зафиксированы в нескольких словарях, можно говорить о формировании частотного списка подобных коллокаций, который, однако, будет основан не на частоте, а на воспроизводимости в словарях и, соответственно, в речи. К этим коллокациям относятся, например, сочетания с прилагательными *глубокий* (*глубокая благодарность*, *глубокий интерес*, *глубокое удовлетворение*), *острый* (*острая борьба, острая дискус-*

[2] https://collocations.spbu.ru/

Таблица 1. **Словарный индекс**

Словарный индекс	Количество коллокаций
6	1
5	25
4	145
3	668
2	4803
1	11 843

сия, острая полемика) и *широкий* (*широкий выбор, широкая известность, широкое сотрудничество*). Поскольку словарные коллокации включают единицы, длина которых превышает два слова, то для них также дополнительно показано пересечение с другими коллокациями. Например, *бросить беглый взгляд* имеет пересечение с двумя словосочетаниями: *бросить взгляд и беглый взгляд*.

Для более понятного отражения результатов предусмотрено графическое представление результатов запроса. Для визуализации коллокаций использовалась библиотека NetworkX для Python 3.7. Ребра графов представлены в шести цветах, каждый цвет соответствует общему словарному индексу:

- желтый цвет — словарный индекс равен 1;
- синий цвет — словарный индекс равен 2;
- фиолетовый цвет — словарный индекс равен 3;
- красный цвет — словарный индекс равен 4;
- коричневый цвет — словарный индекс равен 5;
- черный цвет — словарный индекс равен 6.

На рис. 1 показан пример визуализации для лексемы «буйный». Таким образом, примерами словосочетаний, которые были найдены только в одном словаре, являются *буйный спор*, *буйная головушка*, *буйный океан*, *буйное цветение*. В то время как в двух словарях были зафиксированы следующие коллокации: *буйный натиск*, *буйный смех*, *буйная фантазия*, *буйное воображение*, *буйный рост* и др.

Каждый узел в визуализированной структуре соответствует лемме, при этом цвет узла обозначает определенную часть речи. Например, в рамках рассматриваемых нами атрибутивных словосочетаний си-

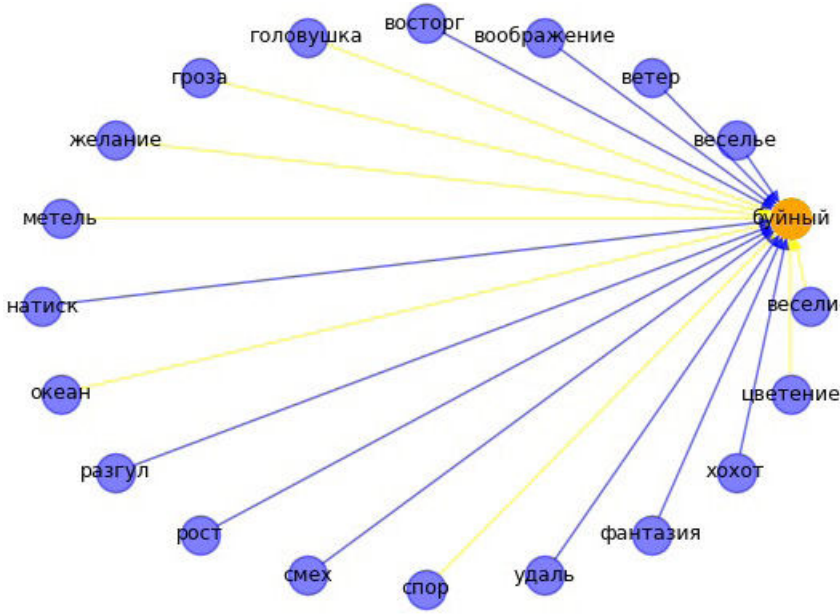


Рис. 1. **Визуализация для лексемы «буйный»**

ний цвет (на рисунке светло-фиолетовый, например, *восторг*, *океан*, *цветение*) обозначает имена существительные, а оранжевый — имена прилагательные (в данном случае только одно слово *буйный*).

С увеличением количества единиц базы данных встал вопрос о том, что с некоторыми частотными словами связано много коллокатов, поэтому такие примеры (в частности, в них могут встречаться по 100 и более слов) могут представлять сложность при восприятии пользователем. В качестве порогового значения было выбрано значение, равное 35 примерам, так как стандартная визуализация NetworkX показывает для них удобочитаемые результаты. Поэтому для слов с числом коллокатов больше порогового было принято решение демонстрировать только так называемые значимые коллокации, под которыми мы понимаем те, которые обладают словарными индексами от 2 и выше, т. е. встречающиеся не менее чем в двух словарях.

Статистический поиск подразумевает более специализированные возможности отображения результатов, которые могут быть полезны для продвинутых пользователей.

Collocate	Borisova	Melchuk	Kustova	Ubin	MAS
бурный	1	1	1	1	0
телячий	0	0	1	1	1
безумный	0	1	1	1	0
совершенный	0	1	1	1	0
неописуемый	0	1	1	1	0

Рис. 2. **Фрагмент выдачи результатов статистического интерфейса для лексемы «восторг»**

t-score	MI	MI3	log-likelihood	logDice
49.25035	8.30587	30.81239	23271.39233	6.63313
15.41538	10.34394	26.13358	2945.26428	3.93735
16.34691	5.81121	22.03869	1687.91858	3.70447
16.91275	4.72291	21.26584	1428.8988	3.39431
64.44143	12.63976	36.68033	65227.93176	8.01149

Рис. 3. **Фрагмент выдачи результатов статистического интерфейса для лексемы «восторг» (продолжение)**

Каждая словарная запись о коллокации снабжена следующей статистической информацией (см. рис. 2 и рис. 3):

- информация о вхождении той или иной коллокации в словари (всего 9 словарей);
- общий словарный индекс коллокации;
- относительная частота в ipm на основе НКРЯ и корпуса Araneum Russicum Maximum;
- значения мер ассоциации на основе Araneum Russicum Maximum (MI, MI3, log-likelihood, logDice, t-score).

Дополнительно для демонстрации значений метрик была использована инфографика (см. пример результатов выдачи для лексемы «ар-

гумент» на рис. 4). Значения разных статистических коэффициентов показаны в виде столбцов диаграмм. На рис. 4 наибольшие значения меры t-score (равные 119 и 118) соответствуют словосочетаниям *весомый аргумент* и *главный аргумент*.

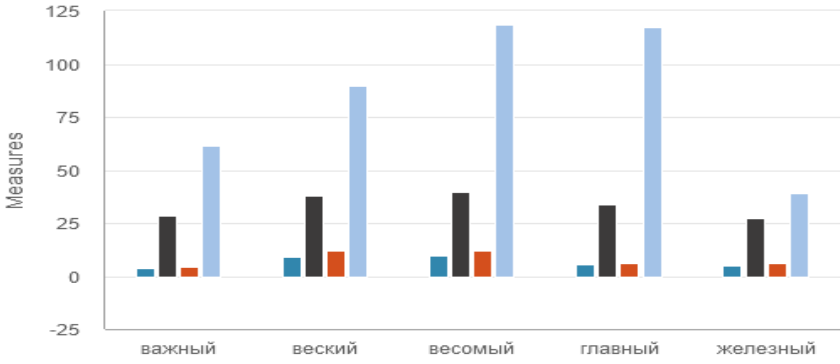


Рис. 4. **Инфографика для лексемы «аргумент»**

Различные статистические метрики могут указывать на неслучайный характер связи между компонентами словосочетаний.

4. Заключение

В заключение необходимо отметить, что база данных нуждается в пополнении, а также снабжении примерами из корпусов текстов не только в виде гиперссылок, но также в виде типовых контекстов. Планируется, что пользователи смогут отметить на их взгляд удачные коллокации или им будет доступна возможность добавить новые.

Разрабатываемая база данных может быть использована при обучении русскому как иностранному, для создания приложений, связанных с автоматической обработкой текстов, и при составлении словарей.

Литература

1. *Апресян Ю. Д.* (2014–2017), Активный словарь русского языка. Т. 1–3. М.: Языки славянской культуры.
2. База данных Lexicograph. URL: http://lexicograph.ruslang.ru (дата обращения: 23.05.2021).

3. Большой академический словарь русского языка в 30 т. (2004–), М.: Российская академия наук. Институт лингвистических исследований, Наука.
4. Большой толковый словарь русского языка: А-Я (1998). Сост., гл. ред. канд. филол. наук *С. А. Кузнецов.* СПб: Норинт.
5. *Борисова Е. Г.* (1995), Слово в тексте. Словарь коллокаций (устойчивых сочетаний) русского языка с англо-русским словарем ключевых слов. М.: Филология.
6. *Денисов П. Н., Морковкин В. В.* (2002), Словарь сочетаемости слов русского языка: 3-е изд., испр. М.
7. *Дерибас В. М.* (1983), Устойчивые глагольно-именные словосочетания русского языка. М.: Русский язык.
8. *Кустова Г. И.* (2008), Словарь русской идиоматики. Сочетания слов со значением высокой степени. М. URL: http://dict.ruslang.ru/magn.php (дата обращения: 23.05.2021).
9. *Мельчук И. А., Жолковский А. К.* (1984), Толково-комбинаторный словарь современного русского языка. Опыты семантико-синтаксического описания русской лексики. Вена: Wiener Slavistischer Almanach.
10. Национальный корпус русского языка. URL: http://ruscorpora.ru (дата обращения: 23.05.2021).
11. *Регинина К. В., Тюрина Г. П., Широкова Л. И.* (1980), Устойчивые словосочетания русского языка: Учеб. пособие для студентов-иностранцев. Под ред. Л. И. Широковой. М.
12. Словарь русского языка (1999), в 4-х т. Под ред. *А. П. Евгеньевой.* 4-е изд., стер. М.: Русский язык.
13. *Убин И. И.* (1987), Словарь усилительных словосочетаний русского и английского языков. М.: Русский язык.
14. *Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R.* (2015), CoCoCo: Online Extraction of Russian Multiword Expressions. In: The 5th Workshop on Balto-Slavic Natural Language Processing. 10–11 September 2015, Hissar, Bulgaria. Sofia: INCOMA Ltd, pp. 43–45.
15. *Lyashevskaya O.* (2010), Bank of Russian Constructions and Valencies. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, pp. 1802–1805.

References

1. *Apresyan Yu. D.* (2014–2017), Aktivnyy slovar' russkogo yazyka [Active Dictionary of the Russian language]. Vol. 1–3. Moscow: Yazyki slavyanskoy kul'tury.
2. Baza dannykh "Lexicograph" [Lexicograph Database]. URL: http://lexicograph.ruslang.ru (date of access: 23.05.2021).
3. Bol'shoy tolkovyy slovar' russkogo yazyka A-YA (1998) [Large Explanatory Dictionary of the Russian Language], *S. A. Kuznetsov* (ed.). St. Petersburg: Norint.
4. Bolshoy akademicheskiy slovar v 30 tomakh [Large academic dictionary of Russian language in 30 volumes]. (2004–2016). Moscow, ILS RAS: Nauka.
5. *Borisova E. G.* (1995), Slovo v tekste. Slovar' kollokatsiy (ustoychivykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov [Word in Texts. Diction-

ary of Russian Collocations with English-Russian Dictionary of Keywords]. Moscow: Filologiya.
6. *Denisov P.N., Morkovkin V.V.* (2002), Slovar' sochetaemosti slov russkogo yazyka [Collocability Dictionary of the Russian Language]. Moscow: Russkij jazyk.
7. *Deribas V.M.* (1983), Ustojchivyje glagol'no-imennyje slovosochetanija russkogo jazyka [Verb-Noun Collocations in Russian]. Moscow: Russkij jazyk.
8. *Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R.* (2015), CoCoCo: Online Extraction of Russian Multiword Expressions. In: The 5th Workshop on Balto-Slavic Natural Language Processing. 10–11 September 2015, Hissar, Bulgaria. Sofia: INCOMA Ltd, pp. 43–45.
9. *Kustova G.I.* (2008), Slovar' russkoj idiomatiki. Sochetanija slov so znachenijem vysokoj stepeni [Dictionary of Russian Idioms. Collocations with the meaning of high degree]. URL: http://dict.ruslang.ru/magn.php (date of access: 23.05.2021).
10. *Lyashevskaya O.* (2010), Bank of Russian Constructions and Valencies. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, pp. 1802–1805.
11. *Mel'chuk I., Zholkovsky A.* (1984), Tolkovo-kombinatornyy slovar russkogo yazyka [Explanatory Combinatorial Dictionary of Russian]. Vienna. Wiener Slavistischer Almanach.
12. Natsional'ny korpus russkogo yazyka [Russian National Corpus]. URL: http://ruscorpora.ru (date of access: 23.05.2021).
13. *Oubine I.* (1987), Slovar' usilitel'nykh slovosochetaniy russkogo I angliyskogo yazykov [Dictionary of Expressive Idioms of the Russian Language]. Moscow: Russkij jazyk.
14. *Reginina K.V., Tjurina G.P., Shirokova L.I.* (1980), Ustoychivye slovosochetaniya russkogo yazyka: Uchebnoye posobiye dlya studentov-inostrantsev [Idioms of the Russian language: Textbook manual for foreign students]. Shirokova L.I. (ed.). Moscow.
15. Slovar' russkogo yazyka v 4 tomakh [Russian Dictionary in 4 volumes] (1999), *Yevgen'yeva A.P.* (ed.). Vol. 1–4, 4th edition, revised and supplemented. Moscow: Russkiy yazyk.

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

Мамаев Иван Дмитриевич
БГТУ «Военмех» им. Д. Ф. Устинова (Россия)
Mamaev Ivan
BSTU Voenmekh named after D. F. Ustinov (Russia)
E-mail: st079541@student.spbu.ru , mamaev_id@voenmeh.ru

А. Н. Чевелева, Э. С. Клышинский

A. N. Cheveleva, E. S. Klyshinskiy

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ КОНСТРУКЦИЙ ДЛЯ ПОВЕРХНОСТНОГО СИНТАКСИЧЕСКОГО АНАЛИЗА

AUTOMATED EXTRACTION OF CONSTRUCTIONS FOR SHALLOW PARSING

Аннотация. В статье исследуется синтаксическая однозначность текстов русского языка на уровне n-грамм. На материале синтаксически размеченного корпуса СинТагРус мы показываем, что синтаксическая структура примерно 25 % связей в тексте может быть восстановлена на уровне 3- и 4-грамм с долей верных порядка 97 %. Наиболее частотными являются такие связи в именных и предложных группах.
Ключевые слова. Поверхностный синтаксический анализ, автоматическое извлечение конструкций, русский язык.

Abstract. In this article, we investigate the issue of syntactical unambiguity of Russian word n-grams. Using SynTagRus corpus, we demonstrate that about 25 % of syntactic links in such a corpus can be obviously established using merely information of part-of-speech 3- and 4-grams with about 97 % accuracy. The most frequent unambiguous connections here are connections in noun and prepositional phrases.
Keywords. Shallow parsing, automatic extraction, the Russian language.

1. Введение

Одним из вариантов проведения синтаксического анализа текста является предварительный поиск синтаксически связанных групп соседних слов (выделение синтаксических фрагментов текста, chunking) [Кудинов 2013]. В некоторых случаях подобный поиск может не принимать во внимание синтаксических связей между словами [Смирнов и др. 2013]. В случае, когда синтаксические связи не нужны, речь обычно идет о выделении терминов, описываемых определенными синтаксическими конструкциями (см., например, [Большакова и др. 2007]), именованных сущностей, извлечении фактов и решении похожих на них задач [Molina et al. 2002]. В случае определения синтаксических связей речь идет о поверхностном синтаксическом анализе. Среди прочего, проведение поверхностного синтаксического анализа позволяет ускорить работу полного синтаксического анализа, сократив количество анализируемых единиц и снизив синтаксическую неоднозначность текста. Подобный подход был популярен в начале 2000-х годов (см., например, [Ножов 2003]), однако введение нейросе-

тевых подходов отодвинуло его на второй план. Предложенные в рамках данного подхода методы все еще остаются актуальными, но для их практического применения должно быть соблюдено несколько условий: точность выделения фрагментов должна превышать точность работы существующих решений; применение результатов этапа выделения фрагментов должно повышать скорость работы системы.

Для проведения поверхностного синтаксического анализа могут использоваться различные методы. Так, в [Molina et al. 2002] используются скрытые Марковские модели, в [Кудинов 2013] — условные случайные поля, в [Большакова и др. 2007] — контекстно-свободные грамматики, а в [Смирнов и др. 2013] упоминаются как конечные автоматы, так и различные методы машинного обучения. Упомянутые методы можно разделить на две группы: эмпирические методы и методы машинного обучения. В работе [Ножов 2003] для выделения именных групп используются конечные автоматы, обладающие высокой скоростью работы. Однако создание подобных автоматов требует длительного ручного труда по поиску кандидатов в извлекаемые конструкции. Методы машинного обучения не всегда показывают сопоставимую точность и работают с меньшей скоростью, однако сами выделяют конструкции из размеченного корпуса, обеспечивая большую полноту.

Мы предлагаем метод автоматического выделения шаблонов для анализа фрагментов текста, обладающих синтаксически однозначной структурой, на основании статистической информации из синтаксически размеченных корпусов русского языка. Одной из задач являлось определение доли текстов, которые могут быть подвергнуты подобному анализу.

2. Метод выделения фрагментов с однозначными связями

На начальном этапе мы разбиваем синтаксически размеченный корпус на предложения и извлекаем из них частеречные 3- и 4-граммы с информацией о номере родительской вершины для каждого слова. При этом мы не рассматриваем опущенные слова (empty nodes), а также исключаем из числа n-грамм иностранные слова, так как они не несут практической пользы. В оставшихся n-граммах номера вершин назначаются относительно начала выбранного окна. Для слов, вершина которых оказалась снаружи n-граммы, постулируется отсутствие внутренней связи. В результате полученные конструкции можно

представить в виде строки *POS1, #host1, …, POSn, #hostn, total_entries*, где *POSi, #hosti* — часть речи и номер родительской вершины i-го слова n-граммы, а total_entries — число вхождений каждой конструкции в корпус. Синтаксические связи рассматриваются как направленные не помеченные.

Очевидно, что даже в частотных конструкциях внутренние связи при поверхностном синтаксическом анализе могут определяться неоднозначно. Например, для частотной конструкции «существительное, прилагательное, существительное» синтаксические связи не определяются однозначно: последнее существительное может как относиться к первому существительному (*обострение холодной войны*), так и иметь родительскую вершину снаружи (*(не представляет для) власти особой проблемы*). С другой стороны, неверно утверждать, что среди менее частотных конструкций не найдется удачных: так, связь частицы и наречия в шаблоне *ADP, PART, ADV* (*В не менее (правильных советских фильмах)*) определяется однозначно.

Для решения этой проблемы мы продублируем n-грамму несколько раз, оставив в каждой только одну связь, и сгруппируем строки по списку частей речи. Для каждой такой n-граммы с одной связью рассчитывается ее суммарное число вхождений. После этого во всех общих конструкциях рассматриваем все слова в отдельности: вычисляются количество конструкций с данным подчинением и доля этого подчинения среди общего числа подобных n-грамм.

Финальный этап состоял в отборе относительно частотных конструкции (50 и более вхождений), используя которые мы будем определять связь правильно с вероятностью не менее 0,97. Уровень 50 вхождений выбран исходя из эмпирических соображений, точность 0,97 выбрана с тем, чтобы конкурировать с современными синтаксическими анализаторами, UAS которых достигает 0,96 (см., например, https://spacy.io/models/ru). В список 3-грамм, подходящих для поверхностного синтаксического анализа, отбираются те, что подходят по данным критериям. Далее из списка 4-грамм удаляются те, что являются расширением контекста справа и слева для уже отобранных 3-грамм (считаем, что этот контекст избыточен). Затем каждая из 3-грамм, не прошедших порог, дополняется контекстом справа. Получившиеся 4-граммы с точностью ниже 0,97 удаляются. Из оставшихся к удачным 3-граммам добавляются те 4-граммы, которые встретились 50 и более раз. Аналогичная процедура расширения проводится для левого контекста. Наконец, из оставшихся 4-грамм,

не являющихся расширением контекста для каких-либо 3-грамм, отбираем прошедшие установленный порог по доле и числу вхождений.

3. Анализ результатов

Эксперименты проводились с синтаксически размеченным корпусом SynTagRus в формате деревьев зависимостей Universal Dependencies (https://universaldependencies.org/) с добавлением маркеров начала и конца предложений. Размер корпуса превышает 1,1 миллиона словоупотреблений. После начального этапа мы получили 10 821 3-грамму и 75 126 4-грамм. После обобщений и отбора было выделено 109 3-грамм и 347 4-грамм (из них 7 и 45, соответственно, содержали маркеры начала и конца предложения). Суммарно удачные 3-граммы встретились в тексте более 146 000 раз, 4-граммы — более 153 000. Это означает, что вместе они покрывают около 27,2 % синтаксических связей корпуса (добавление маркеров начала и конца предложения увеличивает покрытие на 4,7 %). Заметим, что подобная оценка покрытия является оценкой сверху, так как 3- и 4-граммы могут накладываться друг на друга. Если считать все 3-граммы, которые имеют два общих слова в начале и в конце, за совпадения, необходимо сократить нижнюю границу примерно на 16 000 вхождений. Для 4-грамм снижение покрытия может составить около 50 000 вхождений. Таким образом, нижняя граница покрытия может быть оценена в 21,7 %.

Наиболее частотными оказались 3-граммы, связанные с определением связи прилагательного или определителя с существительным в именной и предложных группах: прилагательное/определитель и существительное, перед которыми находится предлог, еще одно существительное, глагол, еще одно прилагательное, определитель или число (например, в *последнее время, (при) помощи специального канала, говорит организатор конференции*). Подобные конструкции описаны, например, в [Кобзарева 2007], и их корректность с заданным уровнем точности не вызывает сомнений. Еще одной частотной конструкцией стало подчинение прилагательного существительному, стоящему перед союзом (*молодой человек* и). Предлог подчиняется стоящему после него существительному, личному местоимению или имени собственному, если после них идут глагол, наречие, союз, знак препинания и некоторые другие части речи (*через месяц после, в комнату вошел, на улицу и*).

Среди прочих были выделены такие редкие конструкции, как вспомогательный глагол, подчиняющийся прилагательному (*трудно было понять*), или союз, подчиняющийся глаголу через другой глагол (и *стал смотреть*).

Среди 4-грамм самой частотной оказалась зависимость предлога от существительного через прилагательное, когда перед ними стоят еще одно прилагательное, предлог или существительное, глагол, некоторые другие части речи (*знакомой до последней трещинки, сказку в новый год*), или когда после них находятся союз, частица, знак препинания или другие части речи (*по ценным бумагам и*). Из низкочастотных конструкций неожиданно была извлечена конструкция «глагол, прилагательное, существительное, существительное», в которой последнее существительное зависит от первого (*измерять абсолютную сложность заданий*).

Наш метод не определил связь первого предлога и существительного как надежную, так как более 3 % случаев приходится на конструкции с составными предлогами или устоявшимися словосочетаниями: *по глубокому снегу*, но (*Несмотря) на свойственную возрасту (впечатлительность)*.

Исходные коды программного обеспечения для получения статистики, а также результаты расчетов приведены в репозитории https://github.com/ancheveleva/grampatterns/

4. Обсуждение результатов

Как было показано выше, от 21,7 % до 27,2 % синтаксических связей в русском тексте может быть восстановлено с использованием только информации о части речи слов. Заметим, что покрытие может быть увеличено за счет снижения требований к частоте встречаемости конструкций (их частоты распределены по закону Ципфа и имеют характерный «тяжелый хвост») или добавления 5- и 6-грамм. При этом мы соблюдаем начальное условие, по которому конструкции должны быть корректными для 97 % примеров, в которых они встречаются (с точностью до корректности разметки корпуса). Расширение до грамматических параметров должно повысить долю правильных ответов, например, при соединении прилагательных и существительных, но, как показали наши эксперименты, не является обязательным. Заметим, что точность работы метода зависит от точности предшествующего ему этапа снятия омонимии. Метод не сможет давать точные ответы без

хорошего разрешения частеречной неоднозначности, но не так чувствителен к неоднозначности леммы или грамматических параметров.

Представленные результаты могут использоваться для разработки методов поверхностного синтаксического анализа или извлечения синтаксически связанных словосочетаний (например, [Klyshinsky 2018]). Метод выгодно отличается от существующих тем, что позволяет автоматически извлечь всех кандидатов, избегая ручного поиска.

Литература

1. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Е., Морозов С. С.* (2007), Лексико-синтаксические шаблоны в задачах автоматической обработки текста. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2007», с. 70–75.
2. *Кобзарева Т. Ю.* (2007), Некоторые свойства линейной структуры именных и предложных групп. Вестник РГГУ. Серия Языкознание. № 8, с. 113–130.
3. *Кудинов М. С.* (2013), Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей. Машинное обучение и анализ данных. Т. 1, № 6, с. 714–724.
4. *Ножов И. М.* (2003), Реализация автоматической синтаксической сегментации русского предложения: Дис. … канд. технических наук. М.
5. *Смирнов И. В., Шелманов А. О.* (2013), Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов. Искусственный интеллект и принятие решений. № 1, с. 41–54.
6. *Klyshinsky E. S., Lukashevich N. Y., Kobozeva I. M.* (2018), Creating a Corpus of Syntactic Co-occurrences for Russian. In: Proc. of "Dialog-2018", pp. 317–331.
7. *Molina A., Pla F.* (2002), Shallow Parsing using Specialized HMMs. In: Journal of Machine Learning Research. No. 2, pp. 595–613.

References

8. *Bolshakova E. I., Baeva N. V., Bordachenkova E. A., Vasil'eva N. E., Morozov S. S.* (2007), Leksiko-sintaksicheskie shablony v zadachah avtomaticheskoj obrabotki teksta [Lexico-Syntactic Patterns for Automatic Text Processing]. In: (транслит) [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog-2007"]. Moscow.
9. *Kobzareva T. Yu.* (2007), Nekotorye svoystva lineynoy struktury imennykh i predlozhnykh grupp [Some Properties of Linear Structure of Noun and Prepositional Phrases]. (транслит) [RSUH Bulletin. Linguistics]. No. 8, pp. 113–130.
10. *Kudinov M. S.* (2013), (транслит) [Shallow Parsing of Russian Text with Conditional Random Fields]. In: (транслит) [Machine Learning and Data Analysis]. Vol. 1, No. 6, pp. 714–724.
11. *Smirnov I. V., Shelmanov A. O.* (2013), Semantiko-sintaksicheskiy analiz estestvennykh yazykov. Chast' I. Obzor metodov sintaksicheskogo i semanticheskogo analiza tekstov.

Iskusstvennyj intellekt i prinyatie reshenij [Semantic-syntactical analysis of Natural Languages. Part I. Overview of Syntactic and Semantic Text Analysis Methods. Artificial Intelligence and Decision-making]. No. 1, pp. 41–54.
12. *Nozhov I. M.* (2003), Realizatsiya avtomaticheskoy sintaksicheskoy segmentatsii pusskogo predlozheniya [Implementation of Automatical Syntactical Segmentation of a Russian Sentence]. PhD thesis. Moscow.
13. *Klyshinsky E. S., Lukashevich N. Y., Kobozeva I. M.* (2018), Creating a Corpus of Syntactic Co-occurrences for Russian. In: Proc. of "Dialog-2018", pp. 317–331.
14. *Molina A., Pla F.* (2002), Shallow Parsing using Specialized HMMs. In: Journal of Machine Learning Research. No. 2, pp. 595–613.

Чевелева Анастасия Николаевна
Национальный исследовательский университет
«Высшая школа экономики» (Россия)
Cheveleva Anastasia
National research university "Higher School of Economics" (Russia)
E-mail: ancheveleva@edu.hse.ru

Клышинский Эдуард Станиславович
Национальный исследовательский университет
«Высшая школа экономики» (Россия)
Klyshinskiy Eduard
National research university "Higher School of Economics" (Russia)
E-mail: eklyshinsky@hse.ru

П. Чжан
P. Zhang

КОРПУСНЫЙ АНАЛИЗ ЛЕКСИЧЕСКОГО НАПОЛНЕНИЯ КОНЦЕПТА «ГОСУДАРСТВО» В РУССКОЙ И КИТАЙСКОЙ ЯЗЫКОВЫХ КАРТИНАХ МИРА

A CORPUS-BASED LEXICAL ANALYSIS OF THE CONCEPT "STATE" IN THE RUSSIAN AND CHINESE LANGUAGE PICTURES OF THE WORLD TITLE OF THE PAPER

Аннотация. Концепт «государство» является важной составляющей частью концептосферы русского и китайского народа. Статья посвящена исследованию лексического наполнения концептов «государство» на основе сопоставимых корпусов русского и китайского языков, выявлению сходств и различий в содержании дистрибутивных тезаурусов русских и китайских слов.
Ключевые слова. Концепт, дистрибутивный тезаурус, сопоставительное исследование, языковая картина мира.

Abstract. The concept "state" is an important component of the conceptual sphere of the Russian and Chinese people. The article is devoted to the study of the lexical content of the concepts "state" on the basis of comparable corpora of the Russian and Chinese languages, identifying the similarities and differences in the content of the distribution thesauri of Russian and Chinese words.
Keywords. Concept, distributional thesaurus, comparative research, the language picture of the world.

1. Введение

Концепт является основной единицей языковой картины мира. Языковая картина мира является представлением о мире, сложившимся в обыденном сознании данного языкового коллектива. В концептах аккумулируется культурный уровень каждой языковой личности и всех носителей языка в целом. Концепт репрезентируется в языке лексемами, фразеологизмами, словосочетаниями, схемами предложений, текстами, совокупностями текстов. Сопоставление концептов разных языков позволяет нам выявлять общечеловеческие универсалии в концептосферах разных народов, и в то же время выявляет специфические, национальные, а затем групповые и индивидуальные признаки концептов и их структуризации.

2. Концепты и семантические поля

Концепт как ментальная сущность имеет национально-специфические черты, соотносимые с мировидением, культурой, обычаями,

верованиями и историей народа. Семантические поля, которые принадлежат различным языкам, отражают структурную организацию окружающего мира. Они являются способами репрезентации языковых картин мира.

Семантические поля, с помощью которых можно описывать лексическую систему языка, фактически являются «овеществленными» отражениями концептов. Впервые термин «семантическое поле» был введен Г. Ипсеном. Семантическое поле обычно определяется как совокупность языковых единиц, объединенных каким-то общим семантическим признаком, имеющих некоторый общий компонент значения. Семантический признак, лежащий в основе семантического поля, может также рассматриваться как некоторая понятийная категория (А. В. Бондарко, Л. М. Васильев, И. М. Кобозева). В. Г. Адмони считает, что семантическое поле характеризуется наличием инвентаря элементов, связанных системными отношениями. По мнению В. Г. Адмони, в семантическом поле можно выделить центральную часть — ядро, элементы которого обладают полным набором признаков, определяющих данную группировку, и периферию, элементы которой обладают не всеми характерными для семантического поля признаками, но могут иметь и признаки, присущие соседним семантическим полям.

3. Сравнительный анализ лексического наполнения концептов «государство» на основе корпусов русского и китайского языков

Цель нашего исследования — сопоставление лексического наполнения концептов в корпусах русского и китайского языков и описание их основных интегральных и дифференциальных признаков. Гипотеза исследования заключается в том, что у концептов государство/国家 существуют сходства и различия, что обусловлено универсальными и национально-культурными признаками.

Материалом исследования является специально созданный нами корпус русского языка по теме «государство» на основе текстов о государстве в русской культуре (3,9 млн токенов) и корпус китайского языка по теме «国家(государство)» на основе текстов на современном китайском языке о государстве в китайской культуре (3,6 млн токенов).

В качестве метода применяется дистрибутивно-статистический анализ, при котором используется информация о дистрибуции элементов текста и их числовых параметрах.

Инструментом исследования является «Тезаурус» в системе Sketch Engine. «Тезаурус» — это инструмент для создания автоматического дистрибутивного тезауруса, который предоставляет нам возможность выявить, какие слова имеют схожую дистрибуцию с заданными словом. В этом случае мы говорим о семантической близости или парадигматических подобий слов. Единицы семантического поля обладают общими парадигматическими и синтагматическими свойствами, что показывает их семантическую близость.

С помощью инструмента «Тезаурус» мы построили дистрибутивный тезаурус для ключевых слов «государство» и «国家» на основе корпусов русского и китайского языках по теме «государство/国家». В таблицах 1 и 2 представлены полученные данные в трех столбцах, в первом столбце приведены лексемы, в втором — коэффициент семантической близости лексем с ключевым словом, в третьем — частота лексемы в корпусе.

Таблица 1. **Дистрибутивный тезаурус для слова «государство» в русском языке**

Lemma	Score	Freq	Lemma	Score	Freq
общество	0,31	3839	организация	0,179	1926
власть	0,294	5589	класс	0,179	1776
народ	0,257	4858	политика	0,17	1657
страна	0,242	4141	демократия	0,165	1315
человек	0,241	7233	церковь	0,165	1802
система	0,227	2450	князь	0,165	4926
Россия	0,215	3515	империя	0,159	1156
право	0,212	4908	орган	0,153	2038
партия	0,21	2145	собственность	0,15	1654
республика	0,203	2265	революция	0,149	1564
жизнь	0,189	3405	развитие	0,148	2623
город	0,184	3356	Русь	0,144	1869
мир	0,182	3056	строй	0,143	950
сила	0,182	3132	хозяйство	0,141	1517
союз	0,181	1580	нация	0,141	942

Таблица 2. **Дистрибутивный тезаурус для слова «государство» в китайском языке**

Lemma	Score	Freq	Lemma	Score	Freq
社会(общество)	0,422	11 991	革命(революция)	0,222	3729
政治(политика)	0,369	11 034	王朝(династия)	0,215	2737
中国(Китай)	0,362	21 721	人民(народ)	0,213	3266
政府(правительство)	0,361	6329	**力**量(сила)	0,212	1938
人(человек)	0,327	10 611	地区(регион)	0,212	2559
民族(нация)	0,318	8993	结构(структура)	0,205	2202
制度(строй)	0,316	4994	**文**明(цивилизация)	0,203	2387
国(страна)	0,313	9764	**日**本(Япония)	0,192	4701
经济(хозяйство)	0,295	6738	**政**党(партия)	0,192	1800
权力(власть)	0,272	3560	**体**系(система)	0,192	1601
帝国(империя)	0,264	2840	**政**策(политика)	0,188	2501
关系(отношение)	0,262	4040	**政**权(власть)	0,181	1567
世界(мир)	0,256	4895	**利**益(интерес)	0,178	2035
组织(организация)	0,23	2751	统治(господство)	0,176	2176
发展(развитие)	0,23	3144	**个**人(индивид)	0,172	1888

Из сравнения элементов из двух семантических полей видно, что совпадают 17 из 30, а именно: *общество, власть, народ, страна, человек, система, партия, мир, сила, организация, политика, империя, революция, развитие, строй, хозяйство, нация.* Таким образом, семантические поля «государство» на китайском и русском языках характеризуются значительным объемом совпадающих элементов.

Можно отметить и различия между китайским и русским полями. В русском тезаурусе по коэффициенту семантической близости (score) занимают первые 5 мест слова *общество, власть, народ, страна, человек*, а в китайском тезаурусе — 社会*(общество),* 政治*(политика),* 中国*(Китай),* 政府*(правительство),* 人*(человек)*. Самые частотные слова в дистрибутивном тезаурусе для слова «государство» в русском языке — *человек, власть, право, народ, страна*, а в китайском дистрибутивном тезаурусе — 社会*(общество),* 中国*(Китай),* 政治*(политика),*

人*(человек)*, 国*(страна)*. Кроме того, в ядре русского поля существуют такие уникальные элементы, как *Россия, князь, Русь, церковь. В* ядре китайского поля есть следующие лексемы, которых нет в русском: 中国*(Китай)*, 朝代*(династия)*, 日本*(Япония)*, 世界*(мир)* и др.

Рассматриваемые различия в русском и китайском полях показывают национально-культурную специфику в русской и китайской картине мира.

4. Переводческие эквиваленты семантического поля «государство»

В этой части мы исследуем переводные эквиваленты для элементов из семантического поля «государство», рассматривая русский язык как исходный, а китайский — как язык перевода.

В исследовании используется параллельный корпус OPUS в системе Sketch Engine. С помощью инструмента «Тезаурус» в Sketch Engine мы построили дистрибутивный тезаурус для слова «государство».

Рис. 1. **Дистрибутивный тезаурус для слова «государство» по корпусу OPUS**

Инструмент «Параллельный корпус» в Sketch Engine предоставляет нам возможность составить параллельные конкордансы для заданного слова.

Для каждой лексемы из тезауруса можно найти возможные переводные эквиваленты. С помощью инструмента «Параллельный корпус» мы получили из корпуса OPUS случайную выборку с 200 параллельными конкордансами для первых 10 элементов тезауруса «госу-

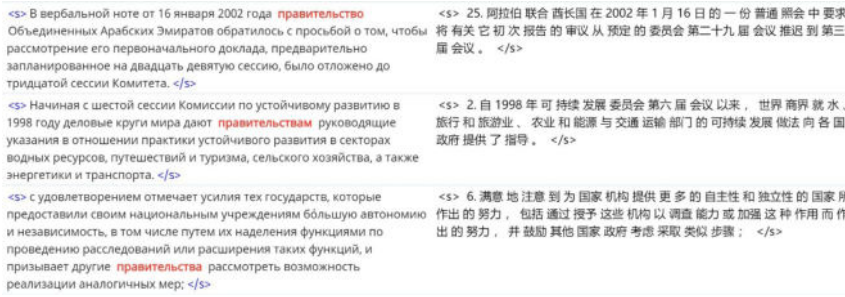


Рис. 2. **Параллельные конкордансы в системе Sketch Engine**

дарство». В табл. 3 показаны переводческие эквиваленты для первых 10 элементов тезауруса «государство» и их процентное соотношение.

Таблица 3. **Переводческие эквиваленты для первых 10 элементов тезауруса «государство»**

Лексемы	Эквиваленты в китайском языке	Процентное отношение
правительство	政府(правительство)	82 %
	国(государство)	2 %
	国家(государство)	2 %
страна	国家(государство)	57 %
	国(государство)	19 %
	境(территория)	2 %
орган	**机**构 (орган)	32 %
	机关 (орган)	10 %
	当局 (администрация)	11 %
	部门 (отделение)	4 %
	实体 (субъект)	3 %
	组织 (организация)	3 %
	局 (управление)	2 %
	政府 (правительство)	1 %
	机制 (механизм)	1 %

Лексемы	Эквиваленты в китайском языке	Процентное отношение
организация	组织 (организация)	58 %
	机构 (орган)	4 %
	举行 (проведение)	2 %
	开展(проведение)	2 %
сторона	方(сторона)	50 %
	方面(сторона)	5 %
	者(человек)	5 %
	国(страна)	2 %
учреждение	机构(учреждение)	65 %
	组织(организация)	2 %
сообщество	社会(общество)	63 %
	共同体(сообщество)	12 %
	界(круг)	5 %
	社区(община)	1 %
группа	组(группа)	31 %
	小组(группа)	24 %
	集团(группа)	7 %
	团体(коллектив)	5 %
	群体(коллектив)	5 %
	股(секция)	2 %
	队(отряд)	2 %
	界(круг)	1 %
	组织(организация)	1 %
совет	理事会(совет)	74 %
	委员会(комитет)	8 %
	会议(собрание)	2 %
	董事会(правление)	1 %
	局(управление)	1 %

Окончание табл. 3

Лексемы	Эквиваленты в китайском языке	Процентное отношение
лицо	人(человек)	78 %
	者(человек)	10 %
	个人(отдельный человек)	7 %
	人员(человек)	5 %

Эти 10 русских слов из дистрибутивного тезауруса «государство» имеют более одного китайского переводного эквивалента. Иногда эти эквиваленты довольно близки по значению русским словам, но следует отметить наличие в китайских переводах оттенков понятий, выражающихся разными словами (иероглифами). Также обращают на себя внимание различия в сочетаемости слов данного семантического поля в русском и китайском языках.

5. Заключение

В статье представлен анализ концепта «государство/国家» в русской и китайской языковых картинах мира. В русской и китайской языковых картинах мира соотносимые концепты «государство/国家» чрезвычайно важны. В семантических полях «государство/国家» имеют значительное число одинаковых элементов: *общество, власть, народ, страна, человек, система, партия, мир, сила, организация, политика, империя, революция, развитие, строй, хозяйство, нация*, это свидетельствует о принципиальном сходстве соответствующих концептов в сопоставляемых языковых картинах мира. Однако нельзя игнорировать и существующие различия лексем, например, лексемы *Россия, князь, Русь, церковь в* русском поле, лексемы 中国*(Китай),* 朝代*(династия),* 日本*(Япония),* 世界*(мир) в* китайском поле — они отражают национально-культурную специфику в русской и китайской языковых картинах мира.

Литература

1. *Аскольдов С. А.* (1980), Концепт и слово. Русская словесность. От теории словесности к структуре текста: Антология. М.

2. *Захаров В. П.* (2019), Методы автоматизированного формирования семантических полей. Структурная и прикладная лингвистика. Вып. 13. СПб, Изд-во С.-Петерб. ун-та, с. 56–79.
3. *Кобозева И. М.* (2000), Лингвистическая семантика. Эдито-риал УРСС. М.
4. *Стернин И. А.* (2001), Методика исследования структуры концепта. Методологические проблемы современной лингвистики. Воронеж.
5. *Zakharov V., Pivovarova S., Gvozdyova E., Semenova N.* (2020), Corpus methods and semantic fields: The concept of empire in English, Russian and Czech. In: A. Ronzhin, T. Noskova, A. Karpov (eds.). R. Piotrowski's Readings in Language Engineering and Applied Linguistics. PRLEAL-2019: Proceedings of the III International Conference, pp. 233–244.

References

1. *Askol'dov S. A.* (1980), Concept and word. In: Russian literature. From Literature Theory to Text Structure: An Anthology. Moscow.
2. *Kobozeva I. M.* (2000), Linguistic semantics. Edito-rial URSS. Moscow.
3. *Sternin I. A.* (2001), Methodology for researching the structure of a concept. In: Methodological problems of modern linguistics. Voronezh.
4. *Zakharov V. P.* (2019), Methods of Automated Generation of Semantic Fields. In: Structural and Applied Linguistics. Issue 13. St. Petersburg: St. Petersburg University Publishing house, pp. 56–79.
5. *Zakharov V., Pivovarova S., Gvozdyova E., Semenova N.* (2020), Corpus methods and semantic fields: The concept of empire in English, Russian and Czech. In: A. Ronzhin, T. Noskova, A. Karpov (eds.). R. Piotrowski's Readings in Language Engineering and Applied Linguistics. PRLEAL-2019: Proceedings of the III International Conference, pp. 233–244.

Чжан Пэйлинь
Санкт-Петербургский государственный университет (Россия)
Zhang Peilin
Saint Petersburg State University (Russia)
E-mail: zhangpl@yandex.ru

Научное издание

ТРУДЫ МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2021»
1–3 июля 2021 г., Санкт-Петербург

Компьютерная верстка *Ю. Ю. Тауриной*
Корректура *Ю. Б. Феофановой*

Подписано в печать 20.10.2021. Формат 60×84 $^{1}/_{16}$.
Усл. печ. л. 23,0. Тираж 50 экз. Заказ № 7398

Типография «Скифия-принт».
197198 С.-Петербург, ул. Б. Пушкарская, д. 10, лит. З.