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

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
«CORPUS LINGUISTICS–2017»

June 27–30, 2017, St. Petersburg

SAINT PETERSBURG
2017

Организационный комитет конференции
«Корпусная лингвистика–2017»

В. П. Захаров (председатель), Е. Л. Алексеева,
Л. Н. Беляева (зам. председателя), А. О. Гребенников,
О. Н. Камшилова, О. Н. Крылова, О. А. Митрофанова,
И. С. Николаев (зам. председателя), Я. К. Харабет, М. В. Хохлова

Программный комитет конференции
«Корпусная лингвистика–2017»

В. П. Захаров (председатель), И. В. Азарова, Е. Л. Алексеева,
Л. Н. Беляева, В. Бенко (Словакия), Н. В. Борисов, В. В. Бочаров,
Р. фон Вальденфельс (Норвегия), Л. А. Вербицкая,
Р. Гарабик (Словакия), Л. Л. Иомдин, Н. Н. Казанский, В. Б. Касевич,
М. В. Копотев (Финляндия), Д. А. Кочаров, О. Н. Ляшевская,
В. Матоушек (Чехия), О. А. Митрофанова, К. Пала (Чехия),
В. Петкевич (Чехия), В. А. Плунгян, Л. В. Рычкова (Беларусь),
С. О. Савчук, В. П. Селегей, Д. В. Сичинава, М. В. Хохлова,
А. Я. Шайкевич, С. А. Шаров (Великобритания), Т. Ю. Шерстинова

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

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2017»

27–30 июня 2017 г., Санкт-Петербург

САНКТ-ПЕТЕРБУРГ
2017

ББК 81.1
Т78

Ответственный редактор издания
В. П. Захаров

Труды международной конференции «Корпусная лингвистика–2017». — СПб. , 2017. — 384 с.

Т78

Сборник содержит материалы докладов, представленных на научной конференции «Корпусная лингвистика–2017» 27–30 июня 2017 г. в Санкт-Петербурге.

Создание корпусов текстов является одним из приоритетных направлений в современной компьютерной 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

Программный комитет конференции выражает искреннюю благодарность Российскому фонду фундаментальных исследований за финансовую поддержку, грант No 17-06-20225 Г

© Авторы, 2017
© Санкт-Петербургский государственный университет, 2017

ОГЛАВЛЕНИЕ

D. Batinić, S. Birzer, H. Zinsmeister

AUTOMATIC CLASSIFICATION OF RUSSIAN TEXTS FOR DIDACTIC PURPOSES

Abstract. In this paper we present the results of an automatic classification of Russian texts into three levels of difficulty. Our aim is to build a study corpus of Russian, in which a L2 student is able to select texts of a desired complexity. We are building on a pilot study, in which we classified Russian texts into two levels of difficulty. In the current paper, we apply the classification to an extended corpus of 577 labelled texts. The best-performing combination of features achieves an accuracy of 0,74 within at most one level difference.
Keywords. L2 Russian, didactic corpus, text complexity, text classification.

1. Introduction

Working with linguistic corpora is an integral part of many foreign language studies (e.g. [Römer 2008; Steinbach & Birzer 2012]). Analyzing texts which are beyond the learner's level may frustrate them and hinder the learning process, whereas reading texts beneath their proficiency may impede their improvement. We argue that the possibility of being able to select a desired level of text difficulty will bring benefits to L2 corpus users in their learning experience. Our goal is to create a Levelled Study Corpus (LeStCor) for L2 learners of Russian that involves filtering options for different complexity levels and a didactic highlighting of difficult morphosyntactic structures [Birzer & Zinsmeister 2016]. In a pilot study of automatic two-level classification on 209 texts, we obtained satisfactory results by considering both surface-oriented features adopted from general readability assessments and more linguistically informed features [Batinić et al. 2016]. In the current paper, we apply a modified classification model to an extended training corpus. In order to discriminate between the difficulty levels, we train an NLTK Naive Bayes classifier on manually labelled texts.

2. Related work

The assessment of text difficulty for native speakers has its origins in the 1920s. Surface-oriented readability measures allowed the researchers to compare different texts in an objective way. More recent approaches integrate features that address a) the lexical coverage of a text, b) parts of speech, c) syntactic structures, e) crosssentential features like the referential overlap and f) relations between clauses triggered by discourse connectives [Benjamin 2012]. Studies aimed at the difficulty level for L2 learners have the underlying hypothesis that L2 learners perceive text comprehensibility dif-

ferently than L1 students [François 2014]. [Chinkina & Meurers 2016], for example, integrated 87 linguistic features to classify English texts into three difficult levels for L2 learners. [Baranova, & Elipašheva 2014] developed a rule-based tool for analyzing the difficulty of Russian texts. Machine learning approaches exploit the strength of different features in a data-driven probabilistic way (e. g. [Xia et al. 2016]). Our work is similar in approach to [Curto et al. 2015], who studied an automatic five- and three-level classification of a small set of Portuguese texts.

3. Material

We selected 577 texts originating from the Test of Russian as a Foreign Language (TORFL, Russian: TRKI) reading and listening tasks. We also included newspaper articles from Ria Novosti[1] and labelled them as the most advanced level (Class III). The number of texts was distributed similarly across classes. A detailed view of the corpus stratification can be found in Table 1. We added the corresponding levels of the Common European Framework of References for Languages (CEFR) for comparison.

Table 1. **TRKI proficiency levels and sampling of the corpus**

Class	TRKI	CEFR	Sem	Texts	Texts/class
I	elementary	A1	1st	57	180
	basis	A2	1st	123	
II	1	B1	2nd	109	206
	2	B2	3rd	97	
II	3	C1	4th	52	191
	4	C2	indep.	25	
		C3	indep.	114	

In order to be able to apply diverse lexical and morphosyntactic features, all texts were tagged and lemmatized with TreeTagger using Russian parameter files, trained on the disambiguated version of the Russian National.[2]

[1] https://ria.ru/ (05.04.2017).

[2] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (05.04.2017).

4. Feature selection

We assumed that the most indicative feature of text difficulty consists of the proportion of basic vocabulary in texts. In order to operationalize the basic vocabulary we used vocabulary lists originating from the textbooks Dialog 1 and 2,[3] which correspond to base and elementary level of language proficiency (A1 and A2 according to CEFR). After preprocessing, the list of basic vocabulary contained 1144 lemmas. We extended the list of basic vocabulary with the list of the 5000 most frequent Russian lemmas compiled by [Sharoff 2002], which proved to be a good text difficulty predictor in our previous study. In addition, we also considered numerals, named entities, pronouns and internationalisms (gathered from Wikipedia's list of internationalisms in Russian), since they are also easily understandable to a L2 student, although not (necessarily) provided in the vocabulary or frequency lists.

With regard to other features, we measured the average number of adverbial participles, perfect participles, parts of speech (nouns, verbs, pronouns, adjectives, adverbs, adpositions, conjunctions, and particles) and abstract words per sentence. Knowing that morphosyntactic features such as participles are introduced at the intermediate proficiency levels (TRKI 1 and TRKI 2), we expected them to be highly discriminative. In order to approximate the number of abstracta in texts, we counted Russian words ending with *-изм* '*-ism*', *-остьь* '*-ness*', *-ство* '*-ship*', *-ота* '*-ness*', *-ание / -ение* (nominalized verbs). We also experimented with other features (lexical density, type/token), which, however, did not prove to be sufficiently informative.

We set the Flesch-Kincaid score adapted to Russian [Oborneva 2016] as our baseline. Flesch-Kincaid approximates the readability of a text by taking into account surface features such as the number of words, sentences and syllables in a text.

5. Results and Discussion

We performed a classification with Naive Bayes (NLTK)[4], and 10-fold cross validation. The values for all the features were set heuristically and by considering the distributions in Figure 1. The highest accuracy (0,74) was achieved by combining the features common words, abstract words, past

[3] Dialog. Lehrwerk für den Russischunterricht. Neue Generation. Bd. 1/2. (2016/2017) [Dialogue. Textbook for Russian language instruction. New Generation. Vol. 1/2.] Berlin: Cornelsen.

[4] www.nltk.org (05.04.2017).

participle and adverbial participle (thresholds: ≥ 95%, < 89% and < 85% for common words, < 0,50, > 1,30 and > 3 for abstracta, > 0 for adverbial participles, > 0 for past participles and > 0,40 for both participles together).

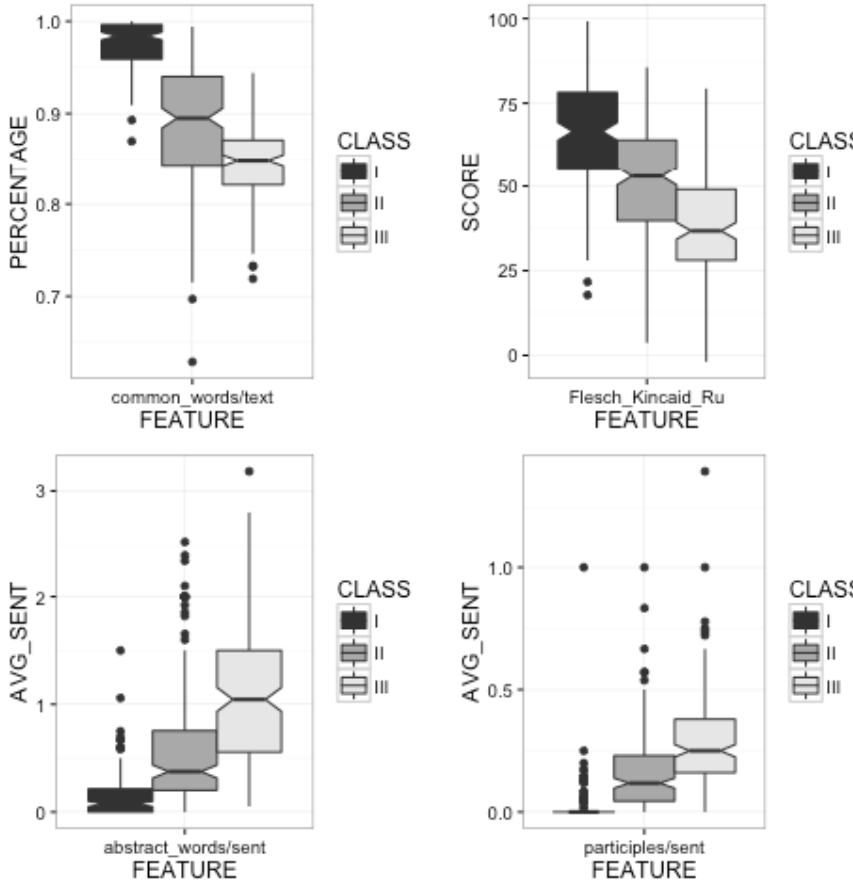

Figure 1. **Boxplots of feature distributions across classes**

As expected, the proportion of common words proved to be the most informative feature: with this feature alone the accuracy rose to 0,68. With an accuracy of 0,63, the combination of average numbers of adverbial and past participle also confirmed the assumption of being good predictors for a three-level text classification. The baseline accuracy of 0,50, which

was the highest achieved by Flesch Kincaid (threshold: > 60) was hence outperformed in a significant manner.

With the best performing combination of features, the classifier only missed within at most one level difference in all ten test sets. The erroneously predicted levels are in many cases also those, whose levels may be disputable even by human judgments.

The feature *common words* proved to be highly informative, especially for discriminating between Class I and Class II. However, differentiating between Class II and Class III based solely on vocabulary lists appears more demanding, given the fact that the vocabulary threshold between intermediate and proficient learners is difficult to estimate. Vocabulary acquisition on a high intermediate level is likely to vary from student to student and may depend on the field in which they choose to intensify their L2 study. Hence, for discriminating between intermediate and advanced levels it might be more appropriate to continue to rely on morphosyntactic features instead of gathering other vocabulary lists. It may as well be advantageous to introduce new features, such as multiword expressions or syntactic formulae, with which a proficient learner should be familiar.

As much as a readability measure such as Flesch-Kincaid may be considered to be a useful indicator of reading comprehension for both native speakers and L2 learners, one must not rely on it entirely when selecting appropriate texts for language learning purposes. The readability score does not in fact measure the level of text difficulty in terms of linguistic features, which prove to be well suited for text classification directed to L2 learners. Measures that only rely on surface features may easily fail in texts with a dialog-like structure, in which the sentences may be short, but the vocabulary may be exigent. On the contrary, passages that may appear unreadable because of long words and sentences might be easily understood by an adult L2 learner if the vocabulary and syntactic structures are familiar.

6. Conclusion

We conducted an automatic classification of Russian texts into three levels of difficulty. The classifier achieved an accuracy of 0,74 with the best predictors consisting of lexical and morphosyntactic features. In a future study, we aim to extend the set of features in order to consider different syntactic and multiword phenomena.

References

1. *Baranova J., Elipašheva T.* (2014), Sozdanie vspomogatel' nogo informacionnogo resursa dlja analiza učebnych tekstov narusskom jazyke. [Creating an auxiliary information resource for the analysis of Russian as a Foreign Language reading texts.]. In: Anis'kin N. (ed.), Čelovek v informacionnom prostranstve. Jar., JaGPU, pp. 232–246.
2. *Batinić D., Birzer S., Zinsmeister H.* (2016), Creating an extensible, levelled study corpus of Russian. Proceedings of KONVENS 2016, Bochum, pp. 38–43.
3. *Benjamin R.* (2012), Reconstructing readability. In: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review, 24, pp. 63–88.
4. *Birzer S., Zinsmeister H.* (2016), The utility of colour markup in texts for learners of Russian. In: Proceedings of the 8th International Biannual Conference "Applied Linguistics in Research and Education". St Petersburg, pp. 139–145.
5. *Chinkina M., Meurers D.* (2016), Linguistically Aware Information Retrieval. In: Providing Input Enrichment for Second Language Learners. Proceedings of BEA. San Diego, pp. 188–198.
6. *Curto P., Mamede N., Baptista B.* (2015), Automatic text difficulty classifier. In: Assisting the selection of adequate reading materials for European Portuguese teaching. Proceedings of CSEDU 2015, Lisboa, pp. 36–44.
7. *François T.* (2014), An analysis of a French as a Foreign Language corpus for readability assessment. In: Proceedings of NEALT, Linköping, pp. 13–32.
8. *Oborneva I. V.* (2006), Avtomatizirovannaja ocenka složnosti učebnych tekstov na osnove statističeskich parametrov. Avtoreferat dissertacii na soiskanie učënoj stepeni kandidata pedagogičeskich nauk. [Automatic rating of the degree of difficulty of textbook texts on the bases of statistical parameters.] Available at: http://naukapedagogika.com/pedagogika-13-00-02/dissertaciyaavtomatizirovannaya-otsenka-slozhnosti-uchebnyh-tekstov-naosnove-statisticheskih-parametrov (05.04.2017).
9. *Römer U.* (2008), Corpora and language teaching. In: Lüdeling A., Kytö M. (eds.), Corpus Linguistics. An International Handbook (vol. 1). Berlin, pp. 112–130.
10. *Sharoff S.* (2002), Meaning as use. In: Exploitation of aligned corpora for the contrastive study of lexical semantics. Proceedings of LREC 2002, Las Palmas, pp. 447–452.
11. *Steinbach A., Birzer S.* (2011), Authentisches Sprachmaterial schnell gefunden. Das Potenzial russischer Textkorpora im Russischunterricht [Authentic texts at your disposal: the potential of text corpora for Russian as a Foreign Language classes]. In: Praxis Fremdsprachenunterricht, 2, pp. 7–10.
12. *Xia M., Kochmar E., Briscoe T.* (2016), Text readability assessment for second language learners. Proceedings of BEA'16, San Diego, pp. 12–22.

Dolores Batinić
Institut für Deutsche Sprache (Germany)
E-mail: batinic@ids-mannheim.de

Sandra Birzer
University of Innsbruck (Austria)
E-mail: Sandra.Birzer@uibk.ac.at

Heike Zinsmeister
Universität Hamburg (Germany)
E-mail: heike.zinsmeister@uni-hamburg.de

V. Benko, A. Butašová

TEACHING CORPUS LINGUISTICS WITH ARANEA WEB CORPORA[1]

Abstract. Our paper describes our experience in introducing the new subject *Introduction to Corpus Linguistics* for the students of language-related programmes at our University. We describe both the technical infrastructure, and the pedagogical aspects related to the subject.

1. Introduction

The first stage of the Aranea Project [Benko 2014; 2016; Benko & Zakharov 2016] has been targeted to creation of a family of dozen+ Gigaword web corpora for languages spoken in Slovakia and its neighbouring countries, as well as for the main foreign languages taught at Slovak universities. This stage is next to completed and the Aranea family currently contains corpora for 18 languages in (usually) two sizes, with some languages having also region-specific variants.

In parallel to building the corpora, works have been done to introduce this resource into teaching within the programmes of foreign language and translation studies at our University. After first four semesters of teaching, we would like to summarize some experiences with the newly introduced subject *Introduction to Corpus Linguistics.*

2. The *Aranea* Corpus Portal

While building the Aranea corpora needed a considerable hardware infrastructure (servers with a lot of RAM and free disk space), the corpus portal itself could be maintained with a moderate hardware configuration. In our case, a new hardware has been recently assigned to our Project — a quad-core virtual machine with 4 GB of main memory and 2 TB of disk space within the University sever cloud. The portal runs the *NoSketch Engine*[2] corpus manager under the *Ubuntu Linux* operating system. The decision for this corpus manager has been mainly motivated by its user-friendliness, rich set of features and ability to cope with very large (i. e., larger than 2 Gigaword) corpora. It is, however, worth noticing that the competing *CQPweb*[3] system would be more user-friendly for the system administrator.

[1] This work has been partially financed by the Slovak KEGA Grant Agency, Project No. K-16-022-00.

[2] https://nlp.fi.muni.cz/trac/noske

[3] http://cwb.sourceforge.net/cqpweb.php

The need to migrate to the virtual system has been accelerated by the crash of the data disk array at our (8-years old) own server. Our initial worries concerning the performance on the virtual machine were not approved and the overall speed of query operations seems to be even higher than those on the "real" machine.

The *Appendix I* shows the home page of our Portal.[4]

Our Portal offers two modes of operation. The *Guest Access* mode (without a password) allows users to work with the smaller (100 Megaword) editions of all corpora, while the *Full Access* mode requires a (free) registration by name and e-mail address. Besides having more corpora at hand, registered users can also profit from some extra corpus manager features, such as saving the default display parameters, creation of subcorpora, etc.

The Guest mode can be conveniently used during the first lessons of a course, as no previous setup it required to start querying the corpus. Though smaller corpora are only available, this is usually not an issue for corpus linguistics beginners. Moreover, 100 Megaword corpora are not really small, are they?

3. The Computer Lab

The minimal configuration for teaching a practically oriented corpus subject is a classroom with a good wireless connection where students can connect their own laptops or tablets. In optimal case, however, a computer lab with workstations having large screens is preferred. It is also important that the projector conveying the contents of the teacher's monitor is able to do it in full resolution and project it at a sufficiently large screen.

Our computer lab contains 20+1 *MS-Windows* workstations with 21″ screens. We have decided to use machines in "all-in-one" configuration requiring less table space than traditional desktop computers. As the corpus manager is fully accessible via a web browser, the only special arrangement was installation of different keyboard layouts for the respective languages. If needed, however, a virtual keyboard, such as that accessible on various web sites[5], can be used.

4. Syllabus of the Course

To make maximal use of the corpora at hand, our new subject has been designed as a series of "hands-on" workshops, with most of the lessons con-

[4] http://unesco.uniba.sk/guest/index.html

[5] http://translit.net/

sisting of small research tasks to be performed by the students themselves. It has been also decided that the syllabus should be created as "language-independent" as possible, which would enable mixed-language groups.

The overall course syllabus is divided into three parts.

4.1 The first part (3–4 lessons) is a "theory-less" introduction into corpus query procedures in an "annotation-agnostic" way. During this stage, the students are shown the main differences between "linguistic querying" provided by corpus managers, and "information querying" provided via search engines. The main idea conveyed here is that "*Google is a very bad concordancer*" [Sharoff, 2006] and the ability to cope with morphological forms is really crucial for linguistic querying (not only) for languages with rich morphology.

The topics covered in the first part are:

1. Typing characters with foreign diacritics
2. Aranea portal in Guest mode
3. Basic queries: word form, lemma, phrase, "Simple query"
4. Frequency distributions and Context search

After the first part of the course, students are able to perform queries without having to know the details of more sophisticated search tools, such as Corpus Query Language.

For example, the context search can be conveniently used for looking for idioms. A query at *Fig. 1* will look for an expression containing keywords "*бисер*" ("pearl") and "*свинья*" ("swine") in any morphological form within a window of 5 words left/right.

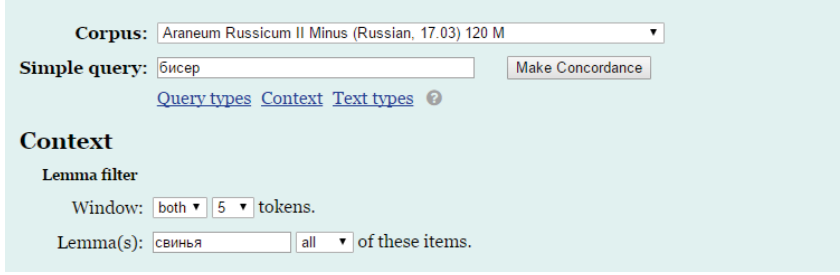

Fig 1.

Fig 2 shows the result of this query operation.

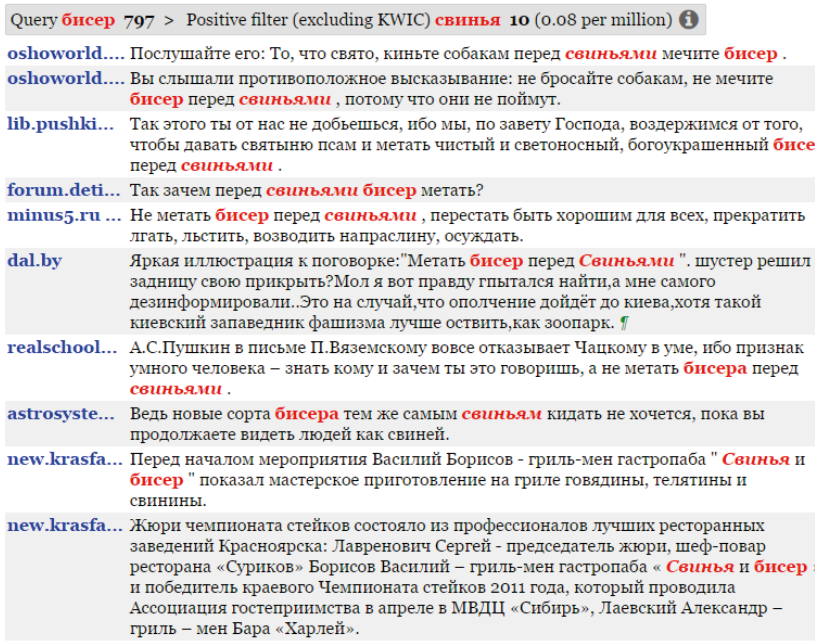

Fig. 2.

4.2. The intermediate lesson is a lecture introducing the basic concepts of corpus linguistics, Web as Corpus (*WaC*) technology [Kilgarriff, 2001; Kilgarriff and Grefenstette, 2003] and corpus annotation, both external and internal (linguistic).

4.3. The remaining lessons progressively cover topics as follows: Corus Query Language (*CQL*)

1. Morphosyntactic annotation, Slovak *SNK tagset*[6]
2. *Araneum Universal Tagset* (*AUT*)[7]
3. "Native" tagsets for other (foreign) languages
4. Regular expressions and their use in corpus queries
5. Collocations and statistical association measures
6. Parallel corpora: *InterCorp*[8] and *Treq*[9]

[6] http://korpus.sk/morpho_en.html
[7] http://aranea.juls.savba.sk/aranea_about/aut.html
[8] https://ucnk.ff.cuni.cz/intercorp/
[9] http://treq.korpus.cz/

4.4. The course is completed by a final assignment having a form of a "crowdsourcing" project. Each student is given a spreadsheet containing 1,000 tokens derived from a frequency list of word forms from the latest version of *Araneum Slovacum* that have not been recognized by the morphological analyzer. Their task is to lemmatize and/or correct the PoS tag for the respective items in the table. As each data file is being processed by two independent annotators, it can be later evaluated and used to amend the morphological lexicon during the next round of tagging.

5. The Textbook

While the previous text describes the already materialized results, the creation of a textbook is currently "work in progress".

The need for a new textbook is dictated not only by absence of any Slovak educational resource on the topic, but also lack of suitable textbook in (say) English, that would cover:

- Introduction into corpus linguistics in a compact form
- Problems of morphosyntactic annotation of morphologically-rich languages, such as Slovak
- Problems of creation and using web corpora
- Direction to use the NoSketch Engine corpus manager

We would like to take inspiration from the unique book of James Thomas [Thomas, 2016]. The planned publication is to appear both in paper and electronic form.

References

1. *Benko V.* (2014), Aranea: Yet another Family of (Comparable) Web Corpora. In: Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8–12, 2014, Proceedings. Ed. P. Sojka et al. Cham; Heidelberg; New York; Dordrecht; London: Springer, 2014, pp. 247–256. ISBN 978-3-319-10816-2.
2. *Benko V.* (2016), Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA), 2016, pp. 4245–4248. ISBN 978-2-9517408-9-1.
3. *Benko V., Zakharov V. P.* (2016), Very large Russian corpora: new opportunities and new challenges. Rec. Alexej Vladimirovič Bajtin, Igor Michajlovič Boguslavskij. In: Kompjuternaja lingvistika i intellektuaľnyje technologii: po materialam meždunarodnoj konferencii "Dialog" (2016), vypusk 15 (22). Otv. red. A. A. Belkina. Moskva: Rossijskij gosudarstvennyj gumanitarnyj universitet, 2016, pp. 79–93. ISSN 2221-7932.

4. *Kilgarriff A.* (2001), Web as corpus. In: P. Rayson, A. Wilson, T. McEncry, A. Hardic and S. Klioja (eds.) Proceedings of the Corpus Linguistics 2001 Conference, Lancaster (29 March — 2 April 2001). Lancaster: UCREL, pp. 342–344.
5. *Kilgarriff A., Grefenstette G.* (2003), Introduction to the Special Issue on the Web as Corpus. In: Computational Linguistics. E-ISSN 1530-9312, 2003, vol. 29, no. 3, pp. 333–347.
6. *Sharoff, S.* (2006), Creating General-Purpose Corpora Using Automated Search Engine Queries. In: WaCky! Working Papers on the Web as Corpus. ISBN 88-6027-004-9, Bologna: Gedit Edizioni, 2006. pp. 63–98.
7. *Thomas J.* (2016), Discovering English with Sketch Engine, 2nd. editoion. Versatile, 2016.

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

Anna Butašová
Comenius University in Bratislava (Slovakia)
E-mail: anna.butasova@uniba.sk

Comenius University in Bratislava
UNESCO Chair in Plurilingual and Multicultural Communication

Aranea Project Main NoSketch Engine Site (Guest Access)

Free registration is required for work with the ***Maius*** and ***Maximum*** class of corpora.
To register, please fill in and submit **this form**.

Language	**Aranea Corpora**	**Minus** 120 M	**Maius** 1,20 G	**Maximum**
Arabic (not tagged yet)	Araneum Arabicum	ℹ Q	ℹ Q *	
Bulgarian	Araneum Bulgaricum	ℹ Q	ℹ Q	
Chinese (simplified script)	Araneum Sinicum	ℹ Q	ℹ Q	
Czech	Araneum Bohemicum	ℹ Q	ℹ Q	5,17 G ℹ Q
Dutch	Araneum Nederlandicum	ℹ Q	ℹ Q	
English	Araneum Anglicum	ℹ Q	ℹ Q	
English (*African TLDs*)	Araneum Anglicum Africanum	ℹ Q	ℹ Q	
English (*Asian TLDs*)	Araneum Anglicum Asiaticum	ℹ Q	ℹ Q	
Finnish	Araneum Finnicum	ℹ Q	ℹ Q	
French	Araneum Francogallicum	ℹ Q	ℹ Q	
French (*African TLDs*)	Araneum Francogallicum Africanum	ℹ Q	ℹ Q *	
Georgian (not tagged yet)	Araneum Georgianum	ℹ Q		
German	Araneum Germanicum	ℹ Q	ℹ Q	
Hungarian	Araneum Hungaricum	ℹ Q	ℹ Q	
Italian	Araneum Italicum	ℹ Q	ℹ Q	
Polish	Araneum Polonicum	ℹ Q	ℹ Q	
Portuguese	Araneum Portugallicum	ℹ Q	ℹ Q	
Russian	Araneum Russicum	ℹ Q	ℹ Q	13,7 G ℹ Q
Russian (*Russia-only TLDs*)	Araneum Russicum Russicum	ℹ Q	ℹ Q	
Russian (*non-Russia TLDs*)	Araneum Russicum Externum	ℹ Q	ℹ Q	
Slovak	Araneum Slovacum	ℹ Q	ℹ Q	2,68 G ℹ Q
Spanish	Araneum Hispanicum	ℹ Q	ℹ Q	
Swedish	Araneum Suedicum	ℹ Q	ℹ Q	
Language	**Other Corpora**	**Minus** 120 M	**Maius** 1,20 G	**Maximum**
Arabic (not tagged yet)	Ajdir Arabicum	ℹ Q *		
Croatian	Zagrabia Croatica (hrWaC)	ℹ Q	ℹ Q	
Slovene	Aemona Slovena (ccGigafida)	ℹ Q		

* Parvum (< 120 M) and Medium (< 1,2 G) class corpora are only available for some languages.

Appendix I.

PARALLEL KEYWORD ANALYSIS: RUSSIAN ELEMENTS IN ENGLISH NADSAT AND FRENCH NADSAT

Abstract. This paper explores two corpus methodologies, keywords analysis and parallel corpora, with relation to an international translingual project to investigate the invented language, Nadsat, in the renowned dystopian novella *A Clockwork Orange*, by the late English novelist Anthony Burgess. In particular, it will report on the differing usages of Russian-derived lexis in the English and French versions of the text.
Keywords. Parallel Corpora, Translation, Nadsat, Anthony Burgess.

Anthony Burgess (1917–1993) was an English writer, composer and journalist who is nowadays best known for his influential dystopian text, *A Clockwork Orange* (1962), which achieved international renown after it was adapted for cinema in 1971 by the American director Stanley Kubrick. The novella has since been translated more than 50 times into 32 different languages.

One of the most striking aspects of the text of the original novella is the introduction of 'Nadsat', an invented slang or argot spoken by the protagonist, Alex, and his associates, known as 'droogs'. This futuristic slang was created by Burgess from a series of linguistic components, the most significant of which is the Russian language. Burgess spoke Russian fluently and set a number of his novels partly or entirely in Russia, including *Honey for the Bears, Any Old Iron* and *Napoleon Symphony*. His stated aim in building the invented language of Nadsat around a lexis of Anglicised Russian loanwords was to generate, during the Cold War era, "a dialect which drew on the two chief political languages of the age."

However, Russian is far from the only component in Burgess's Nadsat. Linguistic analysis has identified a rich melange of components, including multiple other languages, speech forms, truncations and archaisms, within Burgess's Nadsat (Vincent and Clarke, 2016). The resulting deculturated anti-language (Halliday) therefore poses significant challenges to translators, who are tasked with attempting to recreate, either through close tracking of the original or else via creative invention (Malamatidou, 2016) the connotational impact of Burgess's invented slang. The presence of this anti-language, or in-group cant, in the text prompted our use of keyword analysis to help identify the components of Nadsat. This is since keyword analysis is a corpus methodology created precisely for this sort of task.

We also made use of a parallel corpus including the English and French versions of the novel, which enabled us to identify exactly how these keywords were used in the two versions. Though the French translation has been the subject of some critical analysis (Bogic, 2010), this study is the first to generate comparative data derived from a English-French parallel corpus, focussing on the extent of Nadsat components deriving from Russian lexis in the French translation by Georges Belmont and Hortense Chabrier as it compares to the prevalence of Russian-derived Nadsat in the original text.

This research should be seen in the context of the multi-university research project to construct a series of parallel corpora of *A Clockwork Orange* in a number of languages, in order to identify how translators address the challenge of translating out of a source language which is itself invented and which has no acculturation in the real world. The aim of developing these corpora is to provide a framework within which strategies of interlingual translation can be examined in relative isolation.

As a keen linguist, Burgess was well-qualified to approach the issue of inventing an art language such as Nadsat. A lifelong philologist, he produced linguistics textbooks such as *Language Made Plain* (1964) and *A Mouthful of Air* (1992), as well as other art languages such as 'Ulam', the reconstruction of proto-Indo-European created for Jean-Jacques Annaud's 1981 film *Quest for Fire,* and an invented proto-Greek for his translation of Sophocles's *Oedipus the King*. This background linguistic knowledge may account for the success of Nadsat and the attention it has attracted. One of the key successes of Nadsat, which has led to its intrusion into popular culture, is the manner in which the reader of the text is 'brainwashed' into learning a small but notable Russified lexis, thus mirroring the brainwashing theme of the novella itself. The author's claims in this regard have been tested repeatedly in terms of vocabulary acquisition, and found to be substantially sound (Saragi, Nation et al.)

The strikingly Russified lexis of Nadsat was noted not only by the novella's earliest readers, but also by subsequent scholars, almost to the detriment of the other components of Burgess's anti-language, despite some of these elements being not merely evident, but also commented upon within the text. One glossary compiler went so far as to only consider Russian-derived terms as Nadsat, dispensing with the other Nadsat components.

In order to compare how Nadsat functions in translation, two tasks had to be achieved. The second of these is the construction of a parallel translation corpus of existing translations, of which there are more than fifty

into over thirty target languages. This construction is ongoing. The first task, however, was to define Nadsat linguistically. Existing attempts at definition, conducted primarily by literary critics, proved inconsistent and inadequate, and so a comprehensive categorisation was conducted, based on lexis origin, semantic function and vocabulary construction.

In total, seven categories were identified, including compounds, archaisms, babytalk, truncations, rhyming slang and word play. Unsurprisingly, the largest category of Nadsat is what we have termed 'core' Nadsat, that is, the essentially Russian-based relexicalisation of English. This category consists of 219 headwords, of which 10 items are either derived from other languages or of uncertain etymology. The count is based on lemmatisation, and identical forms which realise different parts of speech are also treated as one entry on the list.

In total, the other six categories account for 136 Nadsat terms, compared to 209 core Nadsat terms derived from Russian, which goes some way to explaining the critical focus on Russian relexification within Nadsat. Additionally, a number of terms within these smaller categories, such as the truncation *veck,* meaning 'person', which is derived from the core Nadsat term *chelloveck,* or the babytalk term *malchickiwick,* derived from core Nadsat *malchick,* meaning 'boy', can ultimately be traced back to Russian also. A number of compound Nadsat terms and wordplay Nadsat terms additionally incorporate Russian-derived Nadsat words.

This process of categorisation has permitted us to linguistically identify the perameters of Nadsat in comparison to other art languages. With no dedicated grammar of its own, such as Dothraki or Tolkien's various invented languages possess, Nadsat cannot simply be embedded in a target language translation. Equally, with no acculturation accreted around Nadsat other than that which exists in-text, translators are challenged to replicate both the creative intent and the 'brainwashing' function of the original linguistic invention.

The 1972 translation, *L'Orange Mécanique*, by Burgess's friend Georges Belmont and his partner Hortense Chabrier, is notable both for its longevity (it remains the sole extant French translation of the novella) and the likelihood that Burgess was privy to at least some of its construction, thereby giving it a dimension of indirect authorial authority that other translations lack.

Nadsat items in the English text had been identified by an initial keyword frequency comparison against a reference corpus, and we utilised a similar approach, using the French Ten Ten corpus, to filter for likely

Nadsat terms in the Belmont and Chabrier text. We then filtered this list against a number of French dictionaries to remove unusual but legitimately French terms from our keyword list. The resulting keyword list of likely French Nadsat terms amounted to 207 in total, significantly lower than the 345 Nadsat terms identified in the Burgess text. Following rationalisation of lemma variants, it was established that 77 out of the remaining 143 terms were of Russian origin. This marks a significant reduction in the dominance of Russian-derived lexis from 68.5% of the Nadsat in the English text to a mere 53.8% in the French text. So not only are there fewer overall Nadsat terms in the French translation, but also fewer of those Nadsat terms are derived from Russian. The next stage of analysis will be a comparison of the density of Nadsat usage in both the English and French texts. However, it is already possible to conclude that the Belmont-Chabrier text is notably less Russified than the original.

One possible explanation for this distinction between French-Nadsat and English-Nadsat is the introduction of Anglicised relexicalisation into the French translation to perform some of the alienation functions of Russian-based relexicalisation in the original novella. This methodology is additionally available to other translators of the novella, and can be most notably identified in the second 'Version A' (where 'A' stands for 'angielski' — English) translation of the three extant Polish translations by Robert Stiller. Stiller has translated *A Clockwork Orange* on three separate occasions, in each case generating a new version of Polish-Nadsat based largely on a relexicalisation from a different originating language. The 'Version A' text utilises English, whereas the 'Version R' text, like Burgess's original, utilises Russian.

Howeer, what distinguishes the Belmont-Chabrier text in this regard is the plurality of relexicalisation in the construction of French-Nadsat. Though Burgess had allowed a small number of terms derived from languages other than English and Russian in constructing Nadsat, this small component is much lower than 5 % of the total Nadsat lexis, and is derived from multiple languages including French and Malay. By contrast, the French translation has an identifiable Nadsat component derived from English amounting to nearly 10% of the whole. This is significantly smaller than the Russian-based cohort of French-Nadsat terms, but indicates a translation strategy predicated less on replicating the brainwashing methodology of Burgess's original and more on implying multiple waves of linguistic and cultural colonisation in the Francophone version of the dystopia.

References

1. *Belmont, G., and H. Chabrier*. (1972), L'Orange Mécanique [A Clockwork Orange]. Paris, Robert Laffont.
2. *Bogic, A*. (2010), Anthony Burgess in French Translation: Still 'As Queer as a Clockwork Orange.' Canadian Association for Translation Studies. Young Researchers Selected Papers. Available at: http://act-cats.ca/activities-events/young-researchers/.
3. *Burgess, A*. (1963), A Clockwork Orange. Edited by S. E. Hyman. New York, Norton.
4. *Clarke, J. and B. Vincent* (2017), The Language of A Clockwork Orange: A Corpus Stylistics Approach to Nadsat. Language and Literature.
5. *Malamatidou, S.* (2017), Creativity in translation through the lens of contact linguistics: a multilingual corpus of A Clockwork Orange. The Translator.
6. *Saragi, T., P. Nation, and G. Meister.* (1978), Vocabulary Learning and Reading. In: System 6: 72–78.

Clarke Jim
Coventry University (UK)
E-mail: ab8840@coventry.ac.uk

F. Fischer, T. Orlova, D. Skorinkin,
G. Palchikov, N. Tyshkevich

INTRODUCING RusDraCor, A TEI-ENCODED RUSSIAN DRAMA CORPUS FOR THE DIGITAL LITERARY STUDIES

Abstract. We describe the creation of a corpus of Russian-language drama, comprising hundreds of texts from the end of the 18th century to the first third of the 20th century. Texts are encoded in the XML-based markup standard TEI, the focus is on extra-linguistic, structural annotations, although additional annotation layers can be added easily.
Acknowledgements. The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017–2018 (grant № 17-05-0054) and by the Russian Academic Excellence Project "5-100".
Keywords. TEI, XML, Extra-Linguistic Annotation, Russian Drama, Source Selection, Representativeness, Network Analysis.

1. Demand for a Russian Drama Corpus

In this talk, we will describe the creation of a corpus of Russian drama, spanning from the end of the 18th century to the first third of the 20th century. The corpus will be encoded in accordance with the TEI guidelines (http://www.tei-c.org/Guidelines/), a 30-year old XML standard comprising around 550 elements, specifically defined for digital editions and the demands of the emerging field of Digital Humanities. The creation of the corpus is an ongoing effort; a first set of more than 100 plays will be released at the conference.

The immediate purpose for a machine-readable corpus of that kind is a research project on the social network analysis of literary texts and the automated data analysis for the identification and characterisation of structural features of hundreds of Russian dramas, a research strand which received greater attention after Moretti's tentative network analysis of "Hamlet" (Moretti 2011). An example for the extraction and visualisation of a character network is shown in Figure 1, using Pushkin's historic play "Boris Godunov" as showcase.

2. Related Works

RusDraCor has predecessors like "Shakespeare His Contemporaries" (510 dramas from the Shakespeare era, cf. Mueller 2014), "Théâtre Classique" (940 plays from the 17th and 18th century, curated by Paul Fièvre) and the DLINA corpus (465 German-language plays from 1730 till 1930, cf.

Fischer et al. 2016). These corpora are encoded in TEI and were all derived from existing sources. They have all proved their usefulness for the digital literary studies (Fischer et al. 2016; Glorieux 2016; Xanthos et al. 2016) and it is high time to add a Russian-language collection to this family of drama corpora.

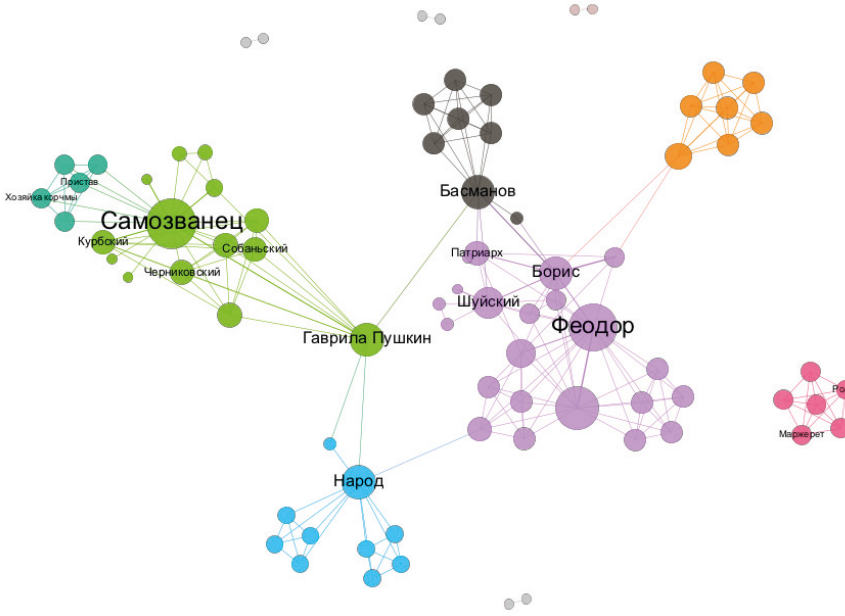

Fig. 1. **Character network from Pushkin's "Boris Godunov" (1825), extracted from a TEI-encoded version of the drama, visualised with Gephi.**

3. Transformation Procedure

The aim of the project is to offer a collection of at least 500 original Russian-language dramas spanning ca. 150 years. In order to do so, we assembled a systematic overview of existing sources, building on the offerings of open libraries like the Russian Virtual Library (rvb.ru), Vikiteka (ru.wikisource.org), ilibrary.ru and lib.ru. All of these texts are encoded in a non-standardised form (*.txt and *.html files). The overview was conducted as a table, which served as a positive list for our web-scraping tool. A snippet from the overview is shown in the following table:

Author	Title	Year	rvb.ru	vikiteka	ilibrary	lib.ru
Фонвизин	Недоросль	1782	[link]	[link]	n/a	[link]
Грибоедов	Горе от ума	1825	n/a	[link]	[link]	[link]
Пушкин	Борис Годунов	1825	[link]	[link]	[link]	[link]
…	…	…	…	…	…	…

This compilation of texts was followed by format conversions, making use of a set of XSLT rules and the Python library Beautiful Soup. Data validation, correction and enriching was done in the oXygen editor, using a customised version of the standard stylesheet of the TEI consortium for highlighting specific XML elements. Methodically, we are following the approach of the Würzburg-based CLiGS junior research group (Schöch et al. 2017).

4. Annotation Layers

Since the emphasis of our initial research project was on interactions between characters in literary texts, RusDraCor mainly provides metadata and non-linguistic annotations (like author; date of origin, publication, premiere; act, scene and speech divisions; character names and IDs; information on gender and social status of speakers, etc.). This enables us to build subcorpora comprising texts by century, decade, author, character, gender etc., and to apply large-scale analyses like speech distribution (cf. Yarkho 1997), topic modeling and stylometry. Further annotations of named entities and parts of speech (cf. Mueller 2014) are intended. RusDraCor will be released under a free license, derivation and enrichment efforts from third parties are welcome.

Bibliography

1. *Fischer, F., Göbel, M., Kampkaspar, D., Trilcke, P.* (2016), Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930. DH2016, Kraków. Available at: http://dh2016.adho.org/abstracts/360.
2. *Glorieux, F.* Dramagraphie 0.2. Online source, April 4th, 2016. Available at: http://resultats.hypotheses.org/749.
3. *Moretti, F.* (2011), Network Theory, Plot Analysis. In: New Left Review 68, 80–102.
4. *Mueller, M.* (2014), Shakespeare His Contemporaries: Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment. In: Digital Humanities Quarterly. 8.3.

5. *Schöch, C., Henny, U., Calvo Tello, J.* (2017), Cligs/textbox: Spring is coming release. Data set, Zenodo, March 10th. Available at: http://doi.org/10.5281/zenodo.376666.
6. *Xanthos, A.* et al. (2016), Visualising the Dynamics of Character Networks. Proceedings, DH2016. Jagiellonian University & Pedagogical University, Kraków, 417–419.
7. *Yarcho, B. I.* (1997), Raspredelenie reci v pjatiaktnoj tragedii (K voprosu o klassicizme i romantizme). Primecanija M. V. Akimovoj; s predisloviem M. I. Šapira. In: Philologica 4, 8/10; 201–288.

Frank Fischer (*ffischer@hse.ru*)
Tatyana Orlova (*tkorlova@edu.hse.ru*)
German Palchikov (*ggpalchikov@edu.hse.ru*)
Daniil Skorinkin (*dskorinkin@hse.ru*)
Natasha Tyshkevich (*natalie.tysh@gmail.com*)
National Research University Higher School of Economics, Moscow

S. Heiden

ANNOTATION-BASED DIGITAL TEXT CORPORA ANALYSIS WITHIN THE TXM PLATFORM

Abstract. This paper presents new developments in the TXM textual corpora analysis platform (http://textometrie.org) towards direct text annotation functionalities. Some annotations are related to a web based external historic ontology called SyMoGiH and others to co-reference between words and to word properties like part of speech or lemma. The paper discusses the methodological stakes of unifying in a single framework those annotations with the traditional ones already available in TXM corresponding to the XML markup of the text sources and to the linguistic annotations automatically added to texts by NLP tools.
Keywords. TXM software, textual corpora analysis, XML, TEI, NLP, annotation.

TXM [Heiden *& al.*, 2010][1] is a software platform offering textual corpora analysis tools. It is delivered as a standard desktop application for Windows, Mac and Linux or as a web portal server application.

Its analysis tools combine qualitative types of tools (close reading) like word lists, concordancing or text edition navigation with synthetic quantitative types of tools (distant reading) like factorial analysis, clustering or statistical cooccurrents analysis.

To be able to work on texts, the platform imports the corpus sources through the following general workflow:

— first the "base text" of each text is established: this operation is called "digital philology" and its results is best represented in the XML format following the encoding recommendations of the Text Encoding Initiative [TEI Consortium, 2016][2].
— then natural language processing (NLP) tools are applied to the texts to automatically add linguistic information like token and sentence boundaries, grammatical category (part of speech), lemma, etc.[3] There is no standard representation of the results of those tools, only de facto standards.

From the point of view of TXM, NLP tools results are seen as annotations added to the XML-TEI representation of texts. And the XML tags can themselves be seen as annotations added to the base text (or raw text),

[1] http://textometrie.org

[2] http://www.tei-c.org

[3] named entities, syntactic structure of sentences, co-reference relations between words are planned.

typically edited with the help of specialised XML editor (like OxygenXML). TXM then implements a traditional "text source encoding and annotation" to "analysis tools applied to the texts" workflow. The text analysis tools use text annotations to offer their services and produce their result. The workflow is unidirectional and the whole of it must be passed through again completely if any annotation needs to be corrected.

We have started to design and develop new possibilities to encode annotations directly in texts from within TXM through some tools results display views.

The first service is based on the annotation of concordance pivots: any sequence of words composing the pivots can be annotated with any historical semantic category coming from the SyMoGiH ontology framework [Beretta, 2016][4]. In this architecture, the SyMoGiH web platform hosts the ontology of historic facts and knowledge and TXM links the identifiers of those data to text spans for further analysis. The TXM internal management of the annotations is equivalent to a re-import of the current representation of the texts annotated. After re-import (after saving annotations) the new annotations are available for all TXM tools to work on like any original "annotation" of the texts (XML based internal structures, word properties, etc.). This development is done in partnership with the LARHRA research laboratory in history[5].

The second service is based on the annotation of text editions: any sequence of words in a text edition page can be annotated with Analec type annotation units and those units can be linked together by other Analec type annotations (relations and schemas). Analec type annotations are designed to help encoding co-reference chains in texts. The model has been developed in the Analec software [Landragin *& al.*, 2012][6], and it is being integrated into TXM for a project funded by the French National Research Agency (ANR) called DEMOCRAT[7].

The third service will be based on the annotation of concordance pivots words: any word present in the pivots of a concordance will be able to be annotated with properties. The primary goal of that service is to annotate and correct grammatical properties and lemma of single words.

---

[4] http://symogih.org/?lang=en

[5] http://larhra.ish-lyon.cnrs.fr

[6] http://lattice.cnrs.fr/Telecharger-Analec

[7] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-15-CE38-0008

This development is done for a project funded by the ANR and Deutsche Forschungsgemeinschaft (DFG) called PaLaFra[8] <http://palafra.org>.

Finally we are developing the possibility to directly edit the XML sources from within TXM through an internal XML editor and we integrate more into TXM the NLP tools framework (at least the syntactic parsing part) for a project funded by the ANR called PROFITEROLE[9].

All those different annotation services integrated into TXM will build a comprehensive annotation-based digital text corpora analysis platform. From an epistemological point of view, the integration of an annotation model into the platform should help its users to better define what comes from the source corpus they analyze and what comes from their own or others interpretation work. The annotations would represent manual (user), semi-automatic (machine+user) or automatic (machine) interpretation results used for further analysis and interpretation work in the same environment.

This work was funded by the ANR under grant numbers ANR-15-CE38-0008 and ANR-16-CE38-0010, and co-funded by the ANR and the DFG under grant numbers ANR-14-FRAL-0006.

References

1. *Beretta, F.* (2016), The symogih.org project: an ontology for collaboratively producing, sharing and curating historical data. In: 36th joined meeting of the CIDOC CRM SIG and ISO/TC46/SC4/WG9 and the 29th FRBR — CIDOC CRM Harmonization meeting, with representatives of the CIDOC CRM SIG and the IFLA FRBR Review Group will take place on 1–4 August 2016, at FORTH, in Heraklio, Crete., Aug 2016, Héraklion (Crete), Greece. Available at: http://new.cidoc-crm.org/sig-meetings.halshs-01423606.
2. *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. Available at: halshs.archives-ouvertes.fr/halshs-00549764.
3. *Landragin, F., Poibeau, T. & Victorri, B.* (2012), ANALEC: a New Tool for the Dynamic Annotation of Textual Data. In: Eighth International Conference on Language Resources and Evaluation, Istanbul, Turquey, 2012, pp. 357–362.

[8] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-14-FRAL-0006

[9] http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-16-CE38-0010

4. TEI Consortium (2016), eds. Guidelines for Electronic Text Encoding and Interchange. 15th December 2016. Available at: http://www.tei-c.org/Guidelines/P5.

Heiden Serge
École normale supérieure de Lyon (France)
E-mail: slh@ens-lyon.fr

Milena Hnátková, Vladimír Petkevič,
Hana Skoumalová

MULTIWORD EXPRESSIONS IN CZECH: BETWEEN LEXICON AND GRAMMAR[1]

Abstract. The present study characterizes a project entitled *Between Lexicon and Grammar* (2016–2018) dealing with a typology of multiword expressions/units (MWEs) in Czech and the development of a representative lexical database/lexicon of MWE entries reflecting this typology.

1. Introduction

In this study, basic features of the project entitled *Between Lexicon and Grammar* (2016–2018) dealing with multiword units (MWE) in Czech is presented. Its primary objective is the development of a detailed multifaceted typology of MWEs and a lexical database of representative 7,000 MWE entries reflecting this typology. The project is a follow-up of a previous project (2013–2015) entitled *The Grammar-Based Treebank of Czech* (cf. Skoumalová et al. 2014, Petkevič et al. 2015a, 2015b, Skoumalová) that was devoted to automatic parsing driven by a formal HPSG-like grammar of Czech, and its second objective is to improve parsing of Czech by integrating relatively petrified chunks of text constituted by MWEs contained in the lexical database. We focus on types of structures that are on the boundary between the lexicon (containing mainly specific and idiosyncratic features of lexemes) and general grammar.

2. The Data

The lexical database is to contain typologically as diverse MWEs as possible. The MWEs are:

1. extracted from the corpora of written contemporary Czech developed in the Czech National Corpus project (cf. corpora such as SYN2010, SYN, SYN2015), primarily on the basis of two measures of MWEs' fixedness: **obligatoriness** and **proximity** (cf. Cvrček 2014), possibly accompanied by other statistical association measures (cf. Evert 2004, Pecina 2010);
2. taken from other sources, mainly from the electronic version of the SČFI dictionary (cf. Čermák et al. 1983–2009). The MWEs contained in this paper dictionary have already been transferred to electronic

[1] Work on this paper was supported by grant number GAČR 16-07473S.

MWE repositories. A special software (FRANTA) was used to identify these MWEs in corpora of synchronic Czech. Moreover, one of the disambiguation modules, phras.rl, exploits these MWEs in order to morphologically disambiguate them in the process of automatic rule-based disambiguation.

3. taken from other sources.

Relevance of the MWEs to be included in the lexical database is supported by information on their usage/frequency in the abovementioned corpora.

3. Database entry

The description of every MWE in the database consists of two parts: global one, which describes the MWE as a whole, and local one, which describes the single positions (words) in the MWE.

3.1. Global description

The global description contains the description of the MWE as a whole. It consists of:

The lemma of the MWE. The form of the MWE as listed in a phraseological dictionary, e. g. *zaklepat bačkorama* ('kick the bucket', lit. 'shake the slippers').

Definition of the MWE. Explanation of the meaning.

Basic part-of-speech pattern. The part-of-speech pattern is derived from the MWE's syntactic tree (achieved by automatic parsing) and expressed as a sequence of extended POS codes, e. g.: *s ničím se nemazlit*: (R_{Instr} P_{Instr} Refl V), 'make quick work of something'; *sedět modelem*: (V N_{Ins}), 'pose model (for someone)'.

Syntactic structure as a dependency/phrase tree. Both syntactic structure and syntactic functions are expressed: *připravit o rozum*: *připravit* (o-SurfHead *rozum*-DeepHead)Obj 'drive someone mad'.

Word order variability. Three kinds of variability are distinguished: 1. fixed word order (*do třetice*, 'third time'); 2. free word order; 3. some chunks are free, some are fixed (e. g. parts constituted by prepositional phrases); the fixed parts of (fragments of) MWE are marked as such.

Type of MWE according to three different categorizations. The MWE typology adopted was primarily generally inspired by the *PARSEME (PARSing and Multi-word Expressions)* project (cf. http://typo.uni-konstanz.de/parseme). The main objective of the project is the development of a complex typology of MWEs not yet existing for Czech. MWEs are primarily

classified according to the following aspects: usage/global type, idiomaticity and syntactic type:

- **Usage/global type** characterizes a MWE as being one of the following kinds:
  1. term: *primitivní funkce* 'primitive function', *psí víno* 'woodbine'
  2. proverb: *Komu se nelení, tomu se zelení*, 'No pains, no gains'
  3. saying/locution: *dobrý večer* 'good evening'
  4. citation: *Sedm statečných*, 'The Magnificent Seven'; *Poslední Mohykán* 'The Last of the Mohicans'.
  5. comparison/simile: *vůl jako anděl*, lit. 'blockhead like an angel';
  6. other.

- **Idiomaticity** describes the degree of MWE's idiosyncrasis/anomaly on the following levels of linguistic description: morphological, syntactic, semantic, lexical, pragmatic. For instance, an MWE is described as syntactically anomalous if it contains a morphologically inappropriate word form in a syntagm. E. g. in the MWE *kluk* ***pitomá*** 'imbecile chap', the adjective *pitomá* 'imbecile' is a non-inflected feminine form, but here it is used as an expressive form that morphologically does not agree with the masculine *kluk* 'chap'.

  1. **Morphological idiomaticity / idiosyncrasy.** This type of idiosyncrasy concerns non-standard morphological forms, e. g.: *chca nechca* 'nolens volens'; the standard form is *chtě nechtě* with the same meaning.

  2. **Syntactic idiomaticity / idiosyncrasy** describes syntactic anomalies of the whole MWE:
     - zeugma: *Byli pro i proti návrh**u***. 'They were both for and against the proposal'
     - syntactic contamination: *On si* ***toho*** *cení* (instead of the correct usage: *On si toho váží.*) 'He appreciates it.'
     - anacoluthon: *Člověk, když si nedá pozor, hned se mu něco stane.* 'A man, if he is not careful, immediately something happens to him.'
     - attraction: *Je širší než* ***delší*** 'He is wider than taller', *v řadě* ***případech*** 'in many cases', ***padni*** *komu* ***padni*** 'Hit or miss'
     - aposiopesis: *Já bych tě nejradši…* 'Most of all I would like to…'
     - ellipsis: *Nevím, co* (***dělat***) *dřív.* 'I do not know what to do first.'
     - translatese constructions: MWEs appearing in translations only: *Je to vo tom, že…* 'It is about … that.'
     - other: *kluk* ***pitomá*** 'imbecile chap' (see above).

3. **Semantic idiomaticity / idiosyncrasy.** The MWE is semantically non-compositional, but the literal reading may be possible. Every MWE is marked whether it occurs in the literal meaning often (e. g. *mít holý zadek*, 'have naked bottom/be poor'), rarely (e. g. *kočičí hlavy*, 'cats' heads/cobblestones') or never (e. g. *nasadit komu psí hlavu*, 'heap dirt upon sb', lit. 'put dog's head on sb'). A special type of a semantic idiom is a **semantic contamination** — an odd combination of two idioms: *mlsný **jazýček** na vahách* (from: *mlsný **jazýček*** '(be) very fussy about one's food' and ***jazýček** na vahách* 'hold the balance'); *sypat si **máslo** na hlavu* (from: *sypat si **popel** na hlavu* 'wear sackcloth and ashes' and *mít máslo na hlavě* 'not to be without blemish').

4. **Statistical idiomaticity.** MWE is a fixed collocation, but semantically compositional (its components cannot be substituted by synonyms), *očitý svědek*, 'eyewitness'. This group contains also terms, compound prepositions (e. g. *v souvislosti s*, 'in connection with') and compound conjunctions (*i když*, 'even though').

5. **Lexical idiomaticity** concerns the MWEs containing:
   - monocollocable word forms: *do **třetice*** 'third time', *křížem **krážem*** 'crisscross', *učinit **zadost*** 'do justice'; such words are often components of terms: *kysličník **osmičelý*** 'osmium tetroxide';
   - almost monocollocable word forms (= associated with a very limited set of words), ***úhlavní** nepřítel* 'arch enemy'; ***zorný** úhel* 'viewpoint, ***zorné** pole* 'field of vision'; ***pitná** voda* 'drinking water', ***pitný** režim* 'drinking habits';
   - word(s) that are only negative: *nedílný* 'integral', *nezcizitelný* 'inalienable';
   - words loaned from a foreign language: *Mírnyx týrnyx* (German: *mir nichts dir nichts*);
   - macaronic structure: *baj voko* ((*v*)*oko=eye*; 'by guesstimate'), *per hubam* (*huba=mouth*; 'orally', 'by word of mouth');
   - translatese: words appearing in translations only (e. g. *pár babek,* '*couple of bucks*').

6. **Pragmatic idiomaticity.** MWEs display pragmatic idiomaticity if they are used in specific situations. E. g. when a man asks a woman to dance with him, he says: *Smím prosit*?, lit. 'May I ask?'

- **Syntactic type** concerns a categorization of MWEs as to their core syntactic structure. The following kinds of constructions are distinguished

mainly according to syntactic head words: nominal phrase (noun-headed), adjectival phrase (adjective-headed), verb phrase, sentential (simple sentence), compound or complex (compound and complex sentences), prepositional phrase, interjectional phrase etc.

Possible transformations of the MWE. The following kinds of variability / transformations are described:

- passivization: MWE can/cannot be passivized;
- topicalization: MWE can/cannot be topicalized;
- nominalization / substantivization: MWE can/cannot be nominalized;
- adjectivization;
- reflexivization;
- irrelevant.

Valency of the entire MWE: (*nést zodpovědnost*) + za_{Acc} NP_{Acc}.

MWE fragments and roots. Some MWE occur in texts as fragments (e. g. *až naprší* instead of the full form *až naprší a uschne*, 'never', lit. 'after it rains and dries again'). They will be matched with data by fuzzy matching. In some cases, creative authors change the phraseme in such a way that it can only be recognized after two or more characteristic words. We call such sets of words roots and they help us to find such modified MWEs in texts. For example, *pýcha předchází pád*, 'pride comes before a fall' has two characteristic words: *pýcha*, 'pride', and *pád*, 'fall'. After searching the corpus for these two words within a sentence, we found … *za pýchou jak stín kluše pád*, '… after pride like a shadow a fall trots'.

3.2. Local description

This part of the database entry contains description of all positions in the MWE:

Morphological disambiguation. Every word form that is not subject to inflection in the given MWE is fully morphologically disambiguated, i. e. it is assigned an unambiguous lemma and a tag. If a lexeme appearing in a given MWE can be inflected, its morphology is only partially accounted for (e. g. only a part of speech, gender of a noun or verbal aspect are specified).

Morphological variability. Every word form that is morphologically variable in the MWE is described as to its morphological variability via morphological categories (gender, number, case, tense, aspect, degree…) and their respective values. E. g., the MWE *košilatý vtip* '*a bawdy joke*' can appear in any of the seven cases, i.e. it displays case variability. In the MWE *hodit flintu do žita* 'throw in the towel', the verb *hodit* 'throw' can be inflected

as expressing 1st, 2nd or 3rd value of person, singular or plural, three tenses (past, present, future) and variant values of aspect (perfective/imperfective).

Variants. Some elements in MWEs can appear in variants, e. g. in the MWE *hodit / házet / zahodit / zahazovat* flintu do žita, 'throw towel', the verb *hodit* can appear in various forms (e. g. (im)perfective variants) as indicated.

Stylistic/Register marker marks every position in the MWE with respect to the register to which the wordform belongs. It distinguishes between the following values: standard; non-codified; colloquial; dialect; other.

Internal modifiability of the position. The MWE element can be internally syntactically modified: *uhodily (**třeskuté**) mrazy*, 'the (bitter) frost started'.

Possibility of negation/affirmation of a component (verb, adjective, noun or adverb) is accounted for, e. g.:

- negation is possible: ***vsadit** vše na jednu kartu,* 'put all one's eggs in one basket' vs. ***nevsadit** vše na jednu kartu,* 'not to put all one's eggs in one basket';
- negation is impossible: ***ber**, kde **ber,*** lit. 'take where take', 'take wherever you can';
- affirmation is impossible: *v **neposlední** řadě,* lit. 'in the non-final row', 'last but not least'.

4. MWE Integration into parsing

In the previous project Grammar-Based Treebank, a treebank of Czech was built using a dependency parser, a formal grammar and a valency lexicon (*Vallex*, http://ufal.mff.cuni.cz/vallex). In the current project we identify MWEs in the data and integrate them with the rest of the parsed dependency structures, using the MWE lexical database to support the parser in two ways (cf. Jelínek 2016). A subset of MWEs that (i) have a fixed word order, (ii) are contiguous, (iii) have a well-defined syntactic structure and (iv) have limited possibilities to be modified by external elements is processed as follows. Before the parsing proper, each MWE of this type is converted into a single token representing the whole MWE. After the parsing, the token is reconverted into its constituent parts and assigned a correct syntactic structure specified in the MWE database/lexicon. The other, larger set of MWEs (having free word order, being externally modifiable etc.) are automatically identified in the text already parsed and, if necessary, their syntactic structure is corrected in order to comply with the structure assigned to this MWE in the database.

5. Conclusion

A project focusing on a multifaceted typology of MWEs was presented. The typology mainly focuses on the description of morphological, syntactic, semantic, lexical and pragmatic idiomaticity. Also, a MWE lexical database reflecting this typology is being developed. The MWEs contained in the database are extracted from corpora of synchronic Czech and they will be used, i.a., for the improvement of parsing of Czech.

References

1. *Cvrček V.* (2014), Kvantitativní analýza kontextu (Quantitative analysis of context). Praha, Nakladatelství Lidové noviny.
2. *Čermák F. et al.* (1983–2009), Slovník české frazeologie a idiomatiky 1–4 (Dictionary of Czech Phraseology and Idiomatics, SČFI 1–4), Praha, Academia / Leda.
3. *Evert S.* (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD dissertation, IMS, University of Stuttgart. Published in 2005. Available at: http://www.collocations.de.
4. *Jelínek T.* (2016), Partial Accuracy Rates and Agreements of Parsers: Two Experiments With Ensemble Parsing of Czech. In: Brejová B. (ed.), ITAT 2016: Information Technologies–Applications and Theory Proceedings. Tatranské Matliare, Slovensko, 42–47.
5. *Pecina, P.* (2010), Lexical association measures and collocation extraction. In: Language Resources and Evaluation, 44 (1–2), 137–158.
6. *Petkevič V., Skoumalová H.* (2015a), The utilisation of valency dictionaries in creating a large Czech treebank. In: Prace Filologiczne, LXVII: 261–277.
7. *Petkevic˘ V., Rosen A., Skoumalová H.* (2015b), The grammarian is opening a treebank account. In: Prace Filologiczne, LXVII: 239–260.
8. *Skoumalová H., Rosen A., Petkevič V., Jelínek T., Vítovec P., Znamenáček J.* (2014), A grammar-licensed treebank of Czech. In: Henrich V., Hinrichs E., de Kok D., Osenova P., Przepiórkowski A. (eds.), International Workshop on Treebanks and Linguistic Theories (TLT13) (December 12–13, 2014, Tübingen, Germany). University of Tübingen, Tübingen 2014, 218–229. Available at: http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf.

Milena Hnátková,
Vladimír Petkevič,
Hana Skoumalová
Charles University, Prague (Czech Republic)
E-mail: first_name.surname@ff.cuni.cz

T. Iakovleva

AUTOMATIC DETECTION OF NEOLOGISMS IN RUSSIAN NEWSPAPER CORPORA WITH NEOVEILLE

Abstract. Neoveille is a web platform that automatically detects new words and monitors word usage change in seven languages [Cartier 2016, 2017]. The platform allows to select corpora, to automatically detect neologisms, to describe them linguistically and to follow their life-cycle. This paper focuses on corpus-based automatic neologism identification in Russian and describes broad tendencies in novel word formation processes. We focus on borrowings and morpho-semantic novel items.
Keywords. Neologisms, natural language processing, corpus linguistics, Russian, word formation.

1. Presentation of the "Neoveille" platform

In the context of globalization, a growing number of studies focus on how English influences the morphological, syntactic and orthographic systems of various languages, including Russian [Galtseva 2014; Rochtchina 2012; Rybushkina 2015]. These studies mainly examined borrowings which were reported to be the largest group of neologisms in modern Russian. To the best of our knowledge, Russian neologisms that are partially or fully composed of native (as opposed to borrowed) linguistic elements, received less attention [Zhdanova and Raciburskaya 2015].

The "Neoveille" platform [2016, 2017], supported by the IDEX-ANR grant, automatically detects new formal and semantic neologisms, regardless of whether they are composed of foreign lexis or of native linguistic items. Although neologism detection platforms such as NEOROM exist for Latin languages [see Humbley 2008 for review], the Néoveille platform is the first of its kind to encompass typologically different languages (Chinese, Czech, French, Greek, Russian, Polish, Portuguese) and to include Slavic languages. Moreover, it is the first platform to propose an automatic detection of semantic neology. The platform provides textual data that can be used for several purposes. Not only is it an on-line dynamic database that monitors neologisms emergence and lifecycle but also a monitor corpora search engine. The extracted data may also enrich on-line lexical resources, such as embedded reference language dictionaries. The following section describes the Néoveille platform focusing on the formal neologism detection, analysis and monitoring.

2. Stages of neologism analysis on Neoveille

2.1. Automatic detection of neologisms

Monitored Russian corpora are currently composed of around 50 newspapers representing general Russian language in journalistic discourse (https://lenta.ru/rss; NEWSru.com, http://izvestia.ru, among others). The Néoveille web platform enables linguists to manage their corpora (via adding, modifying and suppressing), validate or invalidate the automatically detected formal neologisms, describe them linguistically and then follow their lifecycle on monitor corpora.

Linguistic items as well as meta-data (newspaper title, author, theme and date) are automatically extracted via the newspapers' RSS links on a daily basis, three times a day. A specific program is used to extract the relevant text from html pages (https://pypi.python.org/pypi/jusText).

The neologism detection program follows four steps. First, it performs a morphological analysis to identify unknown words. We use the Treetagger [Schmid, 1994] with the language model designed by Sharoff et al. [2008]. This POS-tagger will mark the unknown words with a specific tag. A second step is performed by Hunspell spell-checker, aiming at checking if unknown words are typographic errors or not. Third, the neologism candidates are compared to a complementary exclusion dictionary, fed by linguistic experts. Finally, the resulting Neologism Candidates (CN) are analyzed by linguistic experts who either confirm their neologism status, or classify them as words belonging to a reference dictionary, a terminological lexical unit or to other categories of words to exclude (e. g. typographic mistakes). This excluded dictionaries enable to considerably improve the automatic detection process, as they are automatically re-used by the automatic detector.

2.2. Manual analysis of candidates for neologisms

The detected and validated database of neologisms for Russian currently contains around 460 items.

Linguists classify each neologism according to a typology designed by Pruvost and Sablayrolles [2016]. At the current stage, automatic detection on Néoveille targets three categories of neologisms in Russian: loanwords/borrowings, morpho-semantic novel words and syntactico-semantic words. The present paper focuses on the first two categories. According to the typology, morpho-semantic novel words include the following sub-categories: affixation (prefixation, suffixation or parasynthesis), inflexion and composition. In the present paper, we will not discuss inflexion and parasynthesis, as these word formation processes are represented by less than 10 occurrences.

3. Neologism Classification

3.1. Loanwords

In line with previous research on Russian neologisms, loanwords represent the largest group among neologisms (49 %). Some loanwords come from Arabic or French, e. g. *дезавуировать* (from French *désavouer*) 'renounce (one's claims)'. English is the major source of borrowing. Overall, loanwords vary in the use of script(s). Detected words are written in either (1) Cyrillic script, or (2) Roman script, or (3) as orthographic blends: (1) *сити-кар* 'city car', *аквафермер* 'aquafarmer'; *тег* 'tag'; *вейпинг* 'vaping, that is, using e-cigarettes'; *суперфуды* 'superfoods' (87 %); 2) *machine-learning; seal-watching* (9 %), *Наблюдение за тюленями — это отдельный вид туризма. Он называется seal-watching. (http://murmansk.mk.ru);* 3) *youtube-канал* 'youtube channel' (4 %).

3.2. Prefixation

Prefixation is a relatively infrequent type of morpho-semantic word formation (15 %). Although foreign prefixes are more frequent in novel word formation than native ones (26 vs. 17 respectively), the latter are more frequent in the context of competition (e. g. *лже-* 'pseudo-' vs. *псевдо-* 'pseudo-'). The most productive foreign prefixes are *экс-* 'ex-' (e. g. *экс-работник* 'former employee') and *анти-* 'anti-' (*антитеррористический* 'counterterrorist'). The most productive native prefixes are *лже-* 'pseudo-' (*лжесайт* 'pseudo website') and *не- 'поп-'* (*недострой* 'unfinished construction site').

3.3. Suffixation

Suffixation is almost twice as productive as prefixation (28 %). A little more than a half of suffixed words are formed with foreign roots, mostly adjectives, e. g. *тюбинговый* <tubing+adjectival suffix -ov> 'related to tubing'. In contrast, words formed with native roots are mostly nouns, e. g. *маршрутчик*'mini-bas driver'.

Results show that formation via foreign suffixes is rare (N=5), e. g. *скутерист* <scooter+ist> 'scooter driver', *зацепер* 'train surfer'.

3.4. Compounds

Compounds represent the largest group of morpho-semantic word creation (31 %). We broadly divided compounds in three groups:

- synthetic <adjective + noun> compounds with gender and number agreement (40%), *инновационная еда* 'innovative (Singular Feminine) food (Singular Feminine)'.
- analytical compounds with no number or gender agreement (35%), the components being either linked with a hyphen (a), or presented as a single word (b), (a) *кафе-кальянная* <café(noun)-hookah> 'hookah bar/lounge'; *директор-распорядитель* 'managing director' (b) *автоподход* <auto(mobile) access> 'space allowing a certain place to be accessed by car'; *электровелосипед* 'e-bike, that is, a bicycle with an integrated electric motor'.
- <noun + noun> combinations denoting new objects or concepts (24%), *автомобили smart особо малого класса* <smart cars of a particularly small class/size>, *либерализация визового режима* 'viza regulation liberalisation'; *технология слежения за глазами* 'eye-tracking'.

4. Conclusions

In this research, we analyzed novel words, automatically detected on the basis of 2016-2017 online newspaper corpora. Half of the neologisms are loanwords. The other half is mainly composed of compounds, formed either of native components only, or a mixture of native and foreign components. Finally, suffixation represents the largest group of word formation via affixation.

References

1. *Galtseva, A.* (2014), Neologizmy XXI veka [Neologisms of the XXI century]. In: Kontsept [Concept], Special Issue 13, pp. 1–8.
2. *Rybushkina, S. V.* (2015), Assimil'atsija inojazytchnyx neologizmov v sovremennom russkom jazyke pod vlijaniem ekstraligvistitcheskix faktorov [Assimilation of foreign neologisms in the modern Russian language under extralinguistic influence]. In: Vestnik Tomskogo gosudarstevennogo universiteta [Tomsk State university Newsletter], no. 392, pp. 34–38.
3. *Zhdanova E.* A., *Raciburskaya L. V.* (2015), Sovremennaya ukrainskaya deistviteľnosť v novoobrazovaniyax rossiyskix massmedia [The actual Ukrain reality in new word-buildings of Russian massmedia]. In: Vesti Nizhegoroskogo universiteta im. Lobatchevskogo 2 [News from the Nizhegorodsky University], pp. 397–401.
4. *Cartier E.* (2016), Neoveille, système de repérage et de suivi des néologismes en sept langues [Neoveille, a system of neologism identification and tracking in seven languages]. In: Neologica 10, pp. 101–131.

5. *Cartier E.* (2017), Neoveille, a Web Platform for Neologism Tracking. In: European Chapter of Association for Computational Linguistics 2017, April 2017.
6. *Humbley J.* (2008), Les dictionnaires de néologismes, leur évolution depuis 1945: une perspective européenne [Dictionaries of neologisms, their evolution since 1945: a European perspectve]. In: Sablayrolles (ed.), Neologie et terminologie dans les dictionnaires, Paris, Honoré Chamion.
7. *Pruvost, J., Sablayrolles, J.-F. (eds.)* (2016), Les néologismes, Paris, Presses Universitaires de France.
8. *Sharoff S., Kopotev M., Erjavec T., Feldman A.* and *Divjak D.* (2008), Designing and evaluating Russian tagsets. In: Proc. LREC 2008, Marrakech.
9. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Tatiana Iakovleva-Vigné
Université Paris Diderot (France)
E-mail: ***tiakovle@eila.univ-paris-diderot.fr***

A. Lavrentiev, S. Heiden, M. Decorde

BUILDING AN OPEN MORPHOLOGICAL LEXICON AND LEMMATIZING OLD FRENCH TEXTS WITH THE TXM PLATFORM[1]

Abstract. This paper presents an experience of lemmatizing Medieval French texts (9th – 15th centuries) with the TXM platform (http://textometrie.org). The project uses available lexical resources to compile an open morphological lexicon of Medieval French (FROLEX), which is used in its turn to perform automatic lemmatization. At the final stage, the lemmas are verified and corrected by a human expert. The methodological solutions proposed and the tools for managing lexicons and applying lemmatization developed for TXM may be used for processing other languages, especially those with high variation in spelling and word segmentation practices.
Keywords. Lemmatization, open morphological lexicon, Old French, TXM platform.

The lemmatization of texts in historical language corpora where word forms vary a lot depending on chronological, dialectal and individual factors has always been a challenging task [Piotrowski 2012: 96; Glessgen 2003]. Even within a single text, the variation in spelling and word segmentation may be considerable. The choice of the authority lemma form may also be a problem, as different reference dictionaries sometimes use different entry forms for the same lexeme. For these reasons the value of quality lemmatization is particularly high for historical corpora.

As far as the French language is concerned there exists a number of digitized and natively digital dictionaries (such as Tobler-Lommatzsch[2], DMF[3], DÉCT[4] or AND[5]), as well as a few lemmatized corpora (NCA[6], DÉCT corpus). Some tools for automatic or computer assisted lemmatization of Medieval French are also available. The NCA corpus comes with a morphological lexicon (AFRLEX) where the word forms are associated with lemmas from several sources. It can be used with TreeTagger software [Schmid 1995] but the lemmas it provides are too complex to be convenient for corpus users, as in the following example:

esjoir|esjöir|jouir_*I|Id|MMd (1)

[1] The results presented in this paper were obtained in the framework of the PaLaFra Research Project (ANR-14-FRAL-0006) financed by the French National Research Agency (ANR) and the German DFG granting agency.

[2] http://www.uni-stuttgart.de/lingrom/stein/tl/index.htm

[3] http://www.atilf.fr/dmf

[4] http://www.atilf.fr/dect/

[5] http://www.anglo-norman.net

[6] http://www.uni-stuttgart.de/lingrom/stein/corpus/

Here, "esjoir" comes from the LFA lexicon (Ottawa University), "esjöir" from the Tobler-Lommatzsch (TL) dictionary, and "jouir" from the verb form list compiled by Robert Martin (at early stages of the DMF project).

The LGeRM tool[7] created for the work on the DMF dictionary offers an online lemmatization service but its output requires heavy human work for disambiguation, as it provides all possible lemmas for a given form regardless of context and of morphological tags. The morphological lexicon of the DMF (DMFLEX) can be downloaded from the LGeRM website.

The PALM platform[8] offers an interface for computer assisted lemmatization but does not allow customizing morphological tagsets and has limited import/export capacities.

Our aim in this project is to lemmatize the texts of the *Base de français médiéval* (BFM)[9] using an open morphological lexicon compiled from the best resources available. For the users' convenience, the lemmas should be connected where possible to the online dictionary entries. The BFM morphological tags verified by human experts in nearly 25 % of the corpus can be used for the primary disambiguation.

BFM includes five lemmatized texts by Chrétien de Troyes provided by the DÉCT project. DÉCT uses TL lemmas where possible and adds some of its own. These texts were morphologically tagged using the BFM language model and verified by experts in Medieval French linguistics. They form therefore the basis of the BFM morphological lexicon (BFMLEX).

The first step was to compare the morphological tagsets of AFRLEX and DMFLEX with that of the BFM [Guillot et al. 2013], and to work out conversion rules in order to merge the lexicons. Different tagsets provide unequal level of detail in morphological description, so the joint tagset has to be less detailed than any of the initial ones. For instance, in the merger of BFM (Cattex 2009) and AFRLEX tagsets we had to erase the information on the verb form classes (finite, participle or infinitive) from the BFM tags, as in AFRLEX the verb tags are not sub-categorized. Even more information is erased in the merger with DMFLEX where the distinction between adjectives, pronouns and determiners is not made systematically. In addition, some DMFLEX lemmas have double or triple morphological tags, as in the following example:

néant subst. masc., adv. et pron. indéf. (2)

[7] http://www.atilf.fr/LGeRM/

[8] http://palm.huma-num.fr/PALM/

[9] http://txm.bfm-corpus.org

Here, the lemma *néant* is associated with three morphological categories: noun, adverb and indefinite pronoun. In these case, separate lemma entries are created in the merger for each association of lemma form with a single morphological category.

As for the form of lemmas, most of AFRLEX lemmas are taken from the TL dictionary entries [Kunstmann et al. 2007], while the DMFLEX uses, where possible, the modern French forms. The choices of lemma forms (from the variety of spellings found of the Old French texts) are not entirely homogeneous throughout the TL dictionary, due to the long history of its compilation (over 80 years) and to the absence of "standard" spelling in the Old French. The choice of modern lemma forms ensures the compatibility with modern lexicons (such as that of the TLF dictionary[10]), which is convenient for compiling large diachronical corpora. However, the words that disappeared (or became extremely rare) in the history of the French language are problematic: the DMF uses either the modernized forms that look artificial (such as *cuider* for the Old French *cuidier*, 'to think') or keeps old forms (such as *estovoir* that should have given **étovoir* if the word existed in modern French), which introduces a certain kind of heterogeneity to the lemma list. The rules of creating DMF entry forms are presented in [Martin 1998: 970–973].

In the merger of BFMLEX, AFRLEX and DMFLEX into FROLEX the lemma form of the DMFLEX was preferred, and a "lemma_src" column was created to record the information on the lemma source. A separate table was created to provide correspondences between lemma forms from different sources.

The second step was to develop an extension for the TXM platform [Heiden et al. 2010][11] for working with morphological lexicons. This extension includes commands for importing lexicons in TSV format, for querying different columns using regular expressions, sorting entries, recoding morphological tags, merging lexicons and exporting the compiled lexicon in TreeTagger format. It also includes a set of commands for operating TreeTagger from the TXM interface: train, apply, project lemmas and remove properties. This extension is already available for public beta-testing from an update site dedicated to the PALAFRA project[12].

[10] *Trésor de la langue française*, http://www.atilf.fr/tlfi

[11] http://textometrie.org

[12] http://textometrie.ens-lyon.fr/dist/palafra

The third step, which is currently under way, consists in developing a concordance based user interface for verifying and correcting automatically tagged lemmas of a TXM corpus. While this work is in progress, the verification of lemmas can be done in a spreadsheet software (Libre Office Calc or Miscrosoft Word) thanks to annotation concordances export and import macros.

The first version of the open Medieval French lexicon FROLEX has been published on the GitHub platform under an open-source license so that the NLP community can use it and share its enrichment[13].

References

1. *Glessgen M.* (2003), La lemmatisation de textes d'ancien français: méthodes et recherches [Lemmatization of Old French texts: methods and research]. In: Kunstmann P. et al. (ed.), Ancien et moyen français sur le Web. Enjeux méthodologiques et analyse du discours [Old and Middle French on the Web. Methodological issues and discourse analyses], Ottawa, Les Éditions David, pp. 55–75.
2. *Guillot C., Prévost S., Lavrentiev A.* (2013), Manuel de référence du jeu Cattex09 [Reference Manual for Cattex09 tagset]. Lyon, Équipe BFM. Available at: bfm.ens-lyon.fr/spip.php?article323
3. *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. Available at: halshs.archives-ouvertes.fr/halshs-00549764.
4. *Kunstmann P., Stein A.* (2007), Le nouveau corpus d'Amsterdam [The new Amsterdam corpus]. In: Kunstmann P., Stein A. (ed.), Le nouveau corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23–26 février 2006 [The New Amsterdam corpus. Proceeding of Lauterbad workshop, 23–26 February 2006], Stuttgart, Steiner, pp. 9–27.
5. *Martin R.* (1998), Le Dictionnaire du moyen français (DMF) [The Dictionary of Middle French]. In: Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres [Reports of Assemblies of the Academie des Inscriptions et Belles-Lettres], vol. 142(4), pp. 961–982. Available at: www.persee.fr.
6. *Piotrowski M.* (2012), Natural language processing for historical texts, Morgan & Claypool Publishers.
7. *Schmid H.* (1995), Improvements in part-of-speech tagging with an application to German. Proceedings of the ACL SIGDAT-Workshop, Dublin, pp. 47–50.

Lavrentiev Alexei
Centre national de la recherche scientifique (France)
E-mail: alexei.lavrentev@ens-lyon.fr

[13] https://github.com/sheiden/Medieval-French-Language-Toolkit

Serge Heiden
École normale supérieure de Lyon (France)
E-mail: slh@ens-lyon.fr

Matthieu Decorde
École normale supérieure de Lyon (France)
E-mail: matthieu.decorde@ens-lyon.fr

Bill Louw

ENTAILING PROXIMITY

Abstract. The notion of possible words was always linked in philosophy with the proposition, and propositions were deemed to have a link with sentences in natural language. Propositions and sentences are syntactic. Any outreach between contiguous sentences of the same text looked improbable until Firth [1957] declared that collocation is ‚abstracted at the level of syntax'. This paper demonstrates that subtext uncovers collocates that are germane to and linking of texts as worlds. The empiricism of proximity is therefore entailed as science and philosophy.
Keywords. Entailment, abstracted, variables, collocation, subtext, reference corpus, induction, probability, necessary meaning, worlds.

1. Introduction

Within philosophy, it is common cause that although any sentence of natural language may be deemed to constitute a «possible world», no provision can be made to allow the contiguous sentences of a text to constitute an expanded possible world (personal communication, Siobhan Chapman).

For such a situation to alter within philosophy and, thereafter, for philosophy to hand the resultant instrumentation over to science (as was the case within nuclear philosophy after fission became a reality), the prevalence of predictable empiricism across sentence boundaries would need to be incorporated into the structure of propositions.

This paper will make the case both for corpus proofs and for the science involved in and across sentence boundaries that will allow for the incorporation of *collocation* that bridges the putative gap between natural language *logic* as grammar strings and vocabulary as *metaphysics*. Note that $S \rightarrow L{+}M$ and no longer as $S \rightarrow NP{+}VP$. This divide is as old as philosophy itself and was well understood as early as 450 BC.

The Cobuild dictionary, which was the brainchild of the late John McH. Sinclair, defines entailment as follows: "if one thing entails another, it involves it or *causes* it" [COBUILD 1995: 551] (my emphasis).

This definition is inadequate for dealing with the probabilistic aspect of proximity, especially the proof that is now required in order to deal with the most up-to-date and state of the art nature of *inductive reasoning* and its corpus-based findings.

Computers have finally put to rest the notion that induction suffers from the fatal flaw of circularity. They have accomplished this by providing, outside of real time, the justification of induction as a probabilistic phenomenon. We

no longer need to wait for experience to strike us personally by, for example, supplying us with a black swan in order to dispel our naive notion that *all* As are B to the effect that all swans are white.

The constants that were provided by mathematics in order to reveal the relevant variables are now provided both by the concordancers of modern computing and the genius of Turing through random access.

Resetting time across disciplines is necessary for the reason that language study has lagged badly behind advances in other areas of science, such as physics. The reasons for this lag are varied, but intuitive opacity plays a major part in the matter as does interested opportunism that seeks to delay the advent of a true science of language and uses opacity in order to accomplish its objective. The deficit in temporal terms is fairly shocking and amounts to 117 years if we take for our starting point the discovery by Max Planck of the quantum in 1900. Hence, our starting date for language as its own instrumentation.

The chronoclasm of 117 years must be brushed aside if we consider the benefits of science as these operate especially within the discipline of induction as this is understood across all disciplines. In the same way that we see Whewell [1860] predict the *consilience of induction* that allows Kepler to correct to the orbits of Galileo, we take courage from the quantum as it battles with the major complications of massive empiricism, no computing and the intuitive opacity that comes with dealing with infinitesmally small particles and atoms with *no* instrumentation such as colliders like CERN anywhere in sight.

Suffice it to say that our definition of entailment needs to be adjusted to cope with the major requirements of the quantum era as set out by Hans Reichenbach. Our definition for entailment relies (1) almost entirely upon *causality* rather than probability and (2) makes no provision for dealing with *invisible* elements. The definition must be changed in order to assist *consilience* as we begin to consider the quantum as it affects the discipline of corpus studies. In the UK attempts were apparently made to cap our discipline by preventing the use of reference corpora inductively. To its credit, the Russian Federation has made no such move. Apart from anything else, at the present time the notion of «deep vetting» is likely to remain shallow unless induction and corpus-derived subtext are used as its instrumentation [see Louw in press].

Reichenbach makes the following two stipulations for the future of quantum mechanics and both now need to be added to our earlier definition. Note that he keeps his discourse focused upon philosophy in the service of

science. He says: "The philosophical problems of quantum mechanics are centered around two main issues. The first concerns the transition from *causal* laws to *probability* laws; the second concerns the *interpretation* of *unobserved objects*" [(1944) 1998: 1] (my emphasis).

2. Stages in the Development of the Instrumentation

If we compare the battle in quantum mechanics with the battle in corpus studies, we are immediately struck by the opportunism that is resorted to in corpus studies as opposed to the rigid and disciplined science in quantum mechanics. In the latter case all detail is directed to the creation of better instrumentation. But in the case of semantic prosody corpus findings are departed from in order to champion mentalist models of the imagination. Dominic Stewart [2010] sets out a whole chapter on lexical priming in which intuition asserts that vocabulary primes grammar and logic. This, of course, can never be the case because logic is prior! The science of the quantum wins through while the corpus practitioners operate in a mentalist world on their own making.

Some of the issues listed below can be followed up by readers in order to see why the science of the quantum wins through and corpus studied are hobbled by intuition: Schrödinger's cat; Einstein's hidden variables; determinate values that are arrived at after two separate measurements; measurement that alters the state of what is being measured, hence the EPR Paradox; the Copenhagen interpretation etc. We now need to show how semantic prosody (SP), although probabilistic, is accessible to intuition and part of the laws of causality. And we need to demonstrate that where the nascent quantum is involved, it is opaque to intuition.

This paper will now deal with the following three areas and their acceptability within an enlarged notion of possible worlds. They are all obtainable using corpus-derived subtext. Where examples demonstrate an intention on the part of the author to induce ambiguity, we see that our instrumentation copes readily. However, where meaning is downstepped, it reverts to the laws of causality as these are demanded by the context of situation.

For example, *That is no country for old men* (*that is no * for **) from Yeats's poem «Sailing to Byzantium» purports to criticise Byzantium and its QPVs (most frequent lexis found within it in the reference corpus) are *reason* and *excuse* and devote themselves to the debate as to whether to travel there or not.

However, there is a constant insinuation that Ireland is unfit for habitation and in need of revolution. And this shared meaning is skillfully handled by the poet and by the collocates. The QPVs remain hidden as do the variables in the quantum and they represent hidden objectives. We all remember the test for phonemes and their power but we little expect the hidden QPVs to change the world of the sentence from travel to revolution if we say *this* is no country for old men. The new QPV prompted by *this* is *time*. It is a semantic prosody of haste. Almost a slogan for revolution now!

```
this is no * for *
7 lashed the horse on. </p> <p>  This is no hour for conversation. I do not wish t
8 my mother put on a brave face.  This is no time for crying, she told me firmly. V
9 y to steer him away from them.  This is no time for delicacy and posing, it is ti
10 h, Hubert, McGann said softly. This is no time for divisive talk. We've got a fi
11 hould help and defend the RUC. This is no time for equivocation. </p> Historic d
12 at back on his heels. </p> <p> This is no time for false modesty, Paige. If we d
13 lves quietly and with dignity. This is no occasion for jubilation, certainly not
14 question in a broader context. This is no comfort for a patient who has to wait
15 he work by economic output but this is no argument for lack of careful managemen
16 in some specialities. However, this is no excuse for inaction, for a number of s
17 an seem quite bewildering. Yet this is no reason for ignoring what, taken as a w
18 emporary, says scientists, but this is no reason for complacency about ozone-dep
19 rough its indicators. However, this is no reason for despair since it is by exam
20 enerally increased every year. This is no victory for women, even though it is o
   (Source: The British National Corpus)
```

Some collocates are shared with the line starting with *that* but the causality dominates as readiness to take action. The shared collocates between the sentence with *that* and the sentence with *this* are enough to change the possible world but enough also to *shar*e two worlds. This sharing is catered for by the research of Carnap as «recollection of similarity», but particularly by de Finetti [1980]. Downstepping is not intuitive, but solidly empirical, and *entails* both sets of collocates simultaneously.

Notice how the change to *this* identifies Ireland whereas *that* idealizes what purports to be Byzantium, except of course for old men, but old poets obtain accolades there as part of a pagan ritual and convention.

Another example is *being what it* is. This line is all subtext and no vocabulary. But the hidden objects are all part of the reprehensible conduct of people. The hidden variables can be obtained by opening the individual lines of the concordance in which *human nature* is the main QPV. They are hidden by the phrase because of their monstrously evil nature.

```
35 ential"; but the art of leaking being what it is in Washington, it might has fe
36 o you to do. Live in life. Life being what it is, you'll encounter standards or
37     <F01> Right  <M01> and life being what it is nobody's going to fork on thei
38 nts to hear again </h> <p> Life being what it is, we aren't perfect, said   tha
```

39 inspired move and, Tapie's **luck** being what it is, Alen Boksic came good of s
40 aid: With the automative **market** being what it is at the moment we are wi
41 or not. And with the mass **media** being what it is today, no one can say, ` Amba
42 esirable or not. And the **media**, being what it is, will try to draw a <p> Cadbury
43 h us. I felt that, human nature being what it is, they would be an to liv
44 eeple. `Regrettably, owl nature being what it is, some do conform to the be
45 ecause otherwise, `human nature being what it is", we would have total with attr
46 towards it. Human nature being what it is, this more profound and fash
47 to live. However, human nature being what it is, we do not always live but
48 but 50 and above. Human nature being what it is, if we don't get this week. A
49 e spokesman said: `Human nature being what it is, they want to keep what little
50 ou, my perverse Scottish nature being what it is, the chances of my can he deli

And so we see that *being what it is* is a portmanteau form. It carries collocates that the proposition cannot object to and these spill over into the adjacent sentences of the texts on which they occur and the proposition accommodates them probabilistically and as an enlarged possible world.

Finally, the following example comes closest to a truly quantum-based example. It is drawn from the closing chapter of a book about the stylistic work of John Sinclair [Louw 2016]. Sinclair has great difficulty offering a stylistic interpretation of a poem by a Welsh poet. The poem's title is "Legs". The poem asserts that our legs get us into difficulties of a sexually promiscuous kind.

However, the instrumentation that Sinclair evolves for writing the Cobuild dictionary transcends the difficulties if and only if it is used subtextually. To the surprise of the investigator, the subtext of the poem comes up with only two examples of contexts of situation that are identical out of 44.5 million words of a newspaper corpus [the Times for the year 1995]. One context involves the celibate clergy and the other a chaste relationship between a man and a woman in love that baffles all onlookers. These are the two contexts yielded for the search line *there+was+this+*+and:*

1 Then there was this interesting and pathetic handling of my friendship with Limahl. It received an extraordinary amount of attention for a relationship that was not sexual.
2 The story began as a sort of bad yuletide joke there was this monk and this nun, see, and they got lost in a storm when suddenly lo, they came across a rude hovel replete with inviting hay and fragrant clover.

3. Conclusion

Corpus studies have at last broached the frontier of the quantum. The fact that the subtext of a poet can be obtained will also have profound implications for negotiating and for "deep vetting" as part of the international migrant crisis [Louw in press]. Other forensic uses begin to point to fractured worlds such as those involving modus operandi within the law, as the concordance

below for natural justice suggests. It contains only lines where in practice natural justice is breached. But there are no collocates saying that it can ever be obtained. As Einstein suggested of the quantum, the corpus may now be in a position to identify an absence of ontological status in the case of a modus operandi! What chance is there in such cases that a litigant could interpret the term *natural justice* safely using intuition alone? Our instrumentation may now have come of age.

```
37         claiming a breach of natural  justice. <p> They also claim the
38  back constitute a breach of natural  justice." The letter said Idemitsui
39              was in breach of natural  justice. <h> HOMELESS RICHARDS </h> The
40  considers to be breaches in natural  justice in the official response to the
41 <p> But in a bizarre case of natural  justice, when he ret urned some time later
42 <p> But in a bizarre case of natural  justice, when returned some time later to
43         and, consistent with natural  justice, no further announcement will be
44  both unfair and contrary to natural  justice. They asked for the entire group
45        of acting contrary to natural  justice. <p> The Labour leadership,
```

References

1. *Cobui1d* (1995), The Collins Cobuild English language dictionary, (Ed.) J McH Sinclair et al, Collins, London and Glasgow.
2. *Finetti, B. De* (1980), On the Condition of Partial Exchangeability. In: Jeffrey, R. C. (Ed.) Studies in Inductive Logic and Probability. Vol. 2. University of California Press. Berkeley Ca, USA
3. *Firth J. R.* (1957), Papers in Linguistics: 1934–1951. Oxford, OUP.
4. *Louw W. E.* (2016), Chapter 15. Coda: Unlearning the intuitive analogue as Sinclairan digital proofs transcend stylistics. In: Zyngier S. (Ed.), Language, Discourse, Style: Selected works of John McHardy Sinclair. Amsterdam, John Benjamins.
5. *Louw W. E.* (in press), Negotiators' Language as Code and Cipher. In: Kryachkov D. (Ed.) Proceedings of the International Conference "The Magic of Innovation", MGIMO, 24–25 March 2017.
6. *Reichenbach H.* [(1944) 1998]: The Philosophy of Space and Time. New York, Dover Books.
7. *Stewart D.* (2010), Semantic Prosody: A Critical Evaluation. London, Routlegde.
8. *Whewell, W.* (1860), On the Philosophy of Discovery. Part III of the third edition of The Philosophy of the Inductive Sciences. London.

Bill Louw
Coventry University (UK)
E-mail: godelkreiss@gmail.com

THE RECONSTRUCTION OF THE CONTEXT OF CULTURE THROUGH CORPUS STYLISTICS

Abstract. The paper presents a corpus stylistic view of the context of culture. If cultural contexts are viewed as sets of stereotypes reflected in language use, and the reference corpus of a language may be taken as consisting of frequent and less frequent lexico-grammatical combinations, then a particular set of cultural stereotypes may be fathomed through investigating the linguistic patterns of the given culture. To illustrate the point, the node *south* is investigated in the corpus of Faulkner's *Absalom, Absalom*. In particular, the corpus-derived subtext of one of the key strings containing it is discussed.

Keywords. Corpus stylistics, context of culture, semantic prosody, corpus-derived subtext, Contextual Prosodic Theory, Louw, collocation.

1. Introduction

1.1. The context of culture and the reference corpus

If culture is a set of stereotypes [Kozhin 2007], and a reference corpus is a balanced one and representative of a language, then the most frequent collocations in the reference corpus, whether lexical or lexico-grammatical, will reflect the cultural stereotypes of the language under discussion. It follows from this that if we are to study the context of culture of a particular epoch or social milieu, we may do so by comparing the most frequent collocational patterns of its texts (viewed as language stereotypes) with the most frequent patterns in the reference corpus.

1.2. The context of culture and the grammar of its lexis

The approach presented here breaks no new theoretical ground when it comes to lexical collocation. It stands to reason that frequent lexical collocates of particular nodes in the studied cultural context will differ from their most frequent counterparts in the reference corpus of a particular language. But the socio-emotional specificities [Kozhin 2007] of their use will be more fully understood if we study the grammar strings in which key lexis is embedded. The reason for this is that not only lexis [Louw 1993] but grammar also has its semantic auras, also accessible through their most frequent lexical collocates.

1.3. How do we determine the semantic auras of grammar?

The most frequent lexical collocates of a grammar string are referred to by Louw as quasi-propositional variables (QPVs), and the list of these

for any given string, according to Louw, determine the string's corpus-derived subtext [Louw, Milojkovic 2016]. However, in practice it has been shown that grammatical strings tend to appear in particular contexts, often attidudinally charged, and therefore their semantic auras are best studied at both the level of their QPVs and the level of their semantic prosody (SP). Thus, this paper assumes that if a certain grammar string is used in a text, then its frequent lexical collocates in the reference corpus and its overall semantic prosody will determine its semantic aura and shed additional light on the meaning of the textual segment in which it appears, together with the lexis used therein.

1.4. Why do we need corpus-derived subtext when studying a cultural context?

Sinclair [2006] talks about the lexical collocates of *when she was*. They are either positive (e. g. *approached*) or negative (e. g. *raped*). Stefanowitsch [in press: 271–273] explores a similar approach. Both take a reference corpus as the given and study the most frequent collocations of its grammar (*she* or *he* in this case). The approach presented here focuses on the key lexis in a particular corpus of texts and proposes to use the reference corpus to study the grammar in which the key lexis is embedded in order to investigate semantic auras in a particular cultural setting.

2. Method

William Faulkner's major novel *Absalom, Absalom* is one of the many he wrote describing the history of the American South. The novel starts with the description of a 'long still hot weary dead September afternoon' in the life of a southern gentlewoman Miss Coldfield and young Quentin Compson and finishes with Quentin claiming vehemently to his Northern companion Shreve that he does not hate the South. The node *south*, therefore, is legitimate for research. Below are its contexts found in the novel, in chronological order.

1 son preparing for Harvard in the **South**, the **deep South dead** since 1865 and **peopled** w
2 r Harvard in the **South**, the **deep South dead** since 1865 and **peopled** with **garrulous ou**
3 he was **born and bred** in the **deep South** the same as she was-the two **separate** Quentins
4 that there is **little left** in the **South** for **a young man**. So maybe you will enter the
5 mpson said. „Years ago we in the **South** made our **women** into **ladie**s. Then the **War** came
6 outh, since what **creature** in the **South** since 1861, **man woman nigger or mule**, had had
7 face what the **future** held for the **South** but his **bare hands** and the **sword** which he at
8 parently come into town from the **south — a man** of about twenty-five as the town lear
9 . Yes, **fatality and curse** on the **South** and on our **family** as though because some **ance**
10 l forage wagon from Charleston, **South** Carolina and set above the faint grassy depre

11 you now hold the best of the **old South** which is **dead**, and the words you read were wr
12 g them that if **every man** in the South would do as he himself was doing, would see t
13 xas or California or maybe even **South** America, **daughter doomed to spinsterhood** to l
14 or, when the very **future** of the **South** as a **place bearable** for our **women and childre**
15 as all he was after. **Jesus**, the **South** is **fine**, isn't it. It's better than the theat
16 be present on that day when the South would realise that it was now **paying the pric**
17 ot curiously at all) facing the **South** where further on in the **darkness** the pickets
18 he pickets who, watching to the **South**, could see the **flicker and gleam of the Feder**
19 sure **hate to have come from** the **South**. **Maybe I wouldn't come from the South anyway**,
20 **Maybe I wouldn't come from** the South anyway, even if I could stay there. Wait. Lis
21 blowing of the **fireflies**. „The **South**," Shreve said. „The South. **Jesus**. **No wonder y**
22 „The South," Shreve said. „The **South**. **Jesus**. No wonder you **folks all outlive yours**
23 thing more. Why do you **hate** the **South**? "I don't **hate** it," Quentin said, quickly, at

South collocates with *deep* (3), *dead* (3), *old* (1), *fine* (1), *hate* (3). There are three references to upbringing (*born and bred in the deep South; hate to have come; wouldn't come*), to the present state (*little left, paying the price; you folk all outlive yourselves*), and *future* collocates with *bare hands and the sword* and *place bearable* to the right of the node. There are two references to ill fate (*fatality and curse* and *daughter doomed*. *Family*, *ancestors*, *daughter* are mentioned once. There are several references to people generally: *peopled* (2), *man* (4), *woman* (1), *women* (2). There is a reference to *darkness* and there are two mentions of light of a kind that breaks up darkness (*flicker and gleam* and *fireflies*). Overall, the picture is that of little happiness or hope for the generations of southerners.

These are the contexts containing *Southern*:

1 literary profession as so many Southern **gentlemen and gentlewomen** too are doing now
2 heless the first of the odes to Southern **soldiers** in that portfolio which when your
3 he natural thing for her or any Southern **woman**, **gentlewoman**. She would not have need
4 it to be. Because that's what a Southern **lady** is. Not the fact that, penniless and w
5 n the made-over dress which all Southern **women** now wore, in the carriage still but d
6 der enough since I was not only Southern **gentlewoman** but the very modest character o
7 ll, that there was actually one Southern Bayard or **Guinevere** who was no kin to y ou?

Interestingly, there are two references to men (*gentlemen, soldiers*) and six references to women (*gentlewomen, woman, gentlewoman* (2), *lady, women, Guinevere*). Moreover, the woman is *not* of higher birth only in one case out of the six. This fact connects this concordance with Context 5 of the concordance of *south,* and only this one:

„Ah," Mr Compson said. „Years ago we in the **South** made our women into ladies. Then the War came and made the ladies into ghosts. So what else can we do, being gentlemen, but listen to them being ghosts?"

Thus we arrive at an example of an investigation of a grammar string containing a key lexical item. The string *we in the * made* yielded the following three contexts in the COCA (Davies 2008-),

1 Maybe it's time ***we in the West made*** a similar bow to reality by admitting Taiwan -- the world's 13th largest trading nation (and the sixth largest trader with the U.S.) -as a full-fledged member of GATT.

2 Instead of proceeding without Aideed, the conference was held up for days until the general was finally cajoled into returning. […] To be fair, ***we in the media made*** -- and are still making -- our own contribution to the Aideed persona. Typically, when Aideed arrived 30 minutes late to make a dramatic solo entrance into the final signing ceremony, we swallowed our disgust and ran after him for a quote.

3 No doubt we paid too little attention to potential public reaction as we ***in the industry made*** our program more international, urged on by our trading partners in the developing world. Perhaps we should have seen that the same technology that brought instantaneous trading around the world would foster instantaneous „ anti „ communication and organizing.

These three examples, notwithstanding the differences, all describe a situation in which someone in an inferior position should be/should not have been/was admitted to a higher status than they previously enjoyed. In two cases out of the three poorer countries join richer ones, and in one an undeserving person is elevated to the status of fame and influence by the press. As for the QPVs, *West*, *media* and *industry*, these are all groups in posession of some influence that is exercised.

In the attempt to discover what contexts a verb in the past tense would call up in the given grammatical string, I searched *we in the * *d* in the COCA.

1	irregular reality threads (as	***we in the trade called***	them) to the omnivorous informa
2	What have	***we in the West learned***	that must be shared with the Ea
3	at he meant, at least the way	***we in the town understood***	it, was that some people had
4	d the poor, but in whose name	***we in the West carried***	out the Crusades and imperialis
5	failure to sustain attention.	***We in the UK wanted***	our troops back home almost as soo
6	to execute all three of them.	We in the theatre had	a sense of purpose. It wasn't ab
7	o about the fact that I think	***we in the press missed***	the voter engagement. Do you th
8	try. Then the guests left and	***we in the family visited***	for a while, and then to bed.
9	il. You're talking about what	***we in the industry called***	convergence. The convergence
10	Nor have	***we in the humanities helped***	matters; we can not even a
11	about twenty minutes of one.	***We in the congregation waited***	for him to bring up what
12	, and -- absolutely not true.	***We in the board approved***	the creation of the partnersh

In this concordance, seven contexts out of twelve are critical; five contexts (lines 4, 5, 7, 10, 12) criticise the action described by the verb in the **d* slot. To summarise, the contexts of situation above can be described as "we, a particular group of people bound together by roots or occupation and hence sharing the same customs or expertise made a certain decision that in 41.66 % of the cases we had reason later to regret".

In the Google Books — US corpus (Davies 2011-) the only expression yielded by the search line *we in the * *d* was *we in the West had*, *West* and *had* being the QPVs of the string in the given reference corpus.

1 Nevertheless, ***we in the West had*** to go a long way until we were able to implement a comprehensive policy towards the East.
2 But it was all that ***we in the West had*** going for us.
3 ***We in the West had*** the masses of church members, money, education and status, and quite naturally we imposed our missional priorities and models on the rest of the world.
4 ***We in the West had*** equally unrealistic expectations about the speed of change in Russia. We tried to assist the transition, but failed to appreciate how deeply humilitating, painful and destabilising the 1990s were for the Russian people.
5 as though ***we in the West had*** nothing better to do than spy on our spiritually impoverished neighbour.
6 ***We in the West had*** to turn "nepotism" and "corruption" from tribal virtues into criminal offenses, and we struggle with it.
7 ***we in the West had*** an exaggerated notion of the repressive nature of the Soviet system.
8 we had forgotten the responsibility that ***we (in the West) had*** towards them.
9 It was bad enough that ***we in the West had*** such knowledge, but then the Soviets got it too

In six contexts out of nine (lines 1, 3, 4, 5, 7, 8) the West is criticised for not having shown enough understanding of the Other. This is in agreement with the previous concordance as far as the QVP *West* is concerned, but not when it comes to *had*. The QPVs *West* and *South* are semantically similar as they represent a geographical area opposed to another such area elsewhere, but this does not apply to Context 5 of the concordance of *south* from Faulkner's novel in that «the Other» in this context is in fact the women as opposed to the men, and they all lived in the American South.

In the studied example, the node *south* was studied in the corpus of Faulkner's novel *Absalom, Absalom*, because not only this major novel but a number of his works are about the American South, and because the novel opens and closes with references to it. The concordance of *south* shows a dark picture of no hope, whether for the past, the present or the future generations of Southerners. Additionally, the concordance of Southern yielded 7 contexts from the novel, which highlight that the Southern gentlewoman is one of the novel's themes. For this reason I chose to study the grammar string containing the node south in Context 5 (*we in the South made our women into ladies*), having already expressed the intention of studying a grammar string in the concordance of a key lexical item. The searchline *we in the * made* in the COCA pulled up the contexts of accepting an inferior outsider, and the searchline *we in the * *d* in the same reference corpus showed that the first lexical slot was always occupied by a noun signifying a group of people sharing the same origins or expertise as opposed to possible other such groups. In addition, in Google Books — US that lexical slot was shown to be occupied by *West*, the West being a culturally specific geographical area as opposed to the Other. To conclude, the grammar string underlying

the studied contexts betrays the feeling of the speaker that Southern women were actually (though deservedly) elevated to the status of men, and that they are viewed as the Other. This semantic aura has a bearing on our understanding of the cultural context of Faulkner's novel, and could not have been fully understood without recourse to reference corpora[1].

3. Conclusion

The consistency of corpus-derived subtext and its interpretive power has been proven in existing publications [Louw and Milojkovic 2016], but its role as the determinant of a particular cultural context has not been emphasized. When it comes to cultural studies in corpus linguistics, lexis, and not grammar, has been in the focus of Russian research available to date [Masevich, Zakharov 2016], although they notice certain semantic variation depending on the grammatical case employed.

The innovative quality of this research is that it (a) starts from particular texts and looks for how these deviate from the reference corpus linguistic norm, (b) pronounces the deviation from the linguistic norm to be a secondary, and the linguistic norm to be a primary cultural scenario, and (c) looks at the most frequent lexical collocates of grammar strings as the norm, and the given lexical collocates in particular texts as the deviation from the norm that conceals the cultural norm of a particular age/milieu/text. The approach presented here does not advocate the study of grammar to the exclusion of lexical collocation and co-selection; rather, it highlights (a) the possible hierarchy of meaning construal processes, and (b) how such processes may be culturally pre-conditioned, when it comes to native speakers of a language. Although cultural stereotypes may differ from text to text, the language norm of a particular language remains the frame of reference for all communication. Thus, a cultural set of stereotypes may be revealed through the grammar it employs.

References

1. *Kozhin P.M.* (2007), Etnokul'turnye kontakty naseleniya Evrazii v eneolite — rannem zheleznom veke (paleokul'turologiya i kolesnyj transport) [Ethnocultural Contacts of the Eurasian Population During the Eneolithic Period — Early Iron Age (Paleo-Culturology and Wheeled Transport)]. Vladivostok, Dal'nauka.

[1] In this case, the findings from COHA (Davies 2010-) were sparse and did not significantly change the picture yielded by COCA and Google Books US.

2. *Masevich A. C., Zakharov V. P.* (2016), Metody korpusnoj lingvistiki v istoriheskih i kul'torologicheskih issledovaniyah [Corpus Linguistics Methods in Historical and Cultural Studies]. Available at: https://www.academia.edu/29209174 (accessed 19 November 2016)
3. *Davies, M.* (2008–), The Corpus of Contemporary American English: 520 million words, 1990-present. Available at: http://corpus.byu.edu/coca/
4. *Davies, M.* (2010–), The Corpus of Historical American English (COHA), 400 million words, 1810–2009. Available at: http://corpus.byu.edu/coha/.
5. *Davies, M.* (2011–), Google Books Corpus. (Based on Google Books n-grams). Available at: http://googlebooks.byu.edu/
6. *Louw, W. E.* (1993), Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In: Text and Technology: In Honour of John Sinclair, M. Baker, G. Francis and E. Tognini-Bonelli (eds), 152–176. Amsterdam, John Benjamins.
7. *Louw, B., Milojkovic, M.* (2016), Corpus Stylistics as Contextual Prosodic Theory and Subtext. Amsterdam, John Benjamins.
8. *Sinclair, J. M.* (2006), Phrasebite. Pescia, TWC Sinclair, J.
9. *Stefanowitsch, A.* (in press), Corpus Linguistics: A Guide to the Method. Language Science Press.

Marija Milojkovic
Faculty of Philology, University of Belgrade (Serbia)
E-mail: marija.milojkovic@fil.bg.ac.rs

INFORMATION DENSITY IN A CORPUS OF UNIVERSITY STUDENT WRITING

Abstract. This paper provides an overview of the British Academic Written English (BAWE) corpus and reports on two multidimensional analyses of the corpus, focusing on the findings relating to informational production and density. The most highly informational texts were found in the Social Sciences, at Master's level and in the final year of undergraduate study.
Keywords. Multidimensional Analysis, university student writing, academic writing, BAWE corpus, genre, register.

1. The British Academic Written English (BAWE) corpus

The British Academic Written English (BAWE) corpus contains about 6.5 million words of proficient university student writing collected at several British universities in the first decade of the 21st century. BAWE represents assessed university writing across four years of study (first, second and final year undergraduate, and taught Masters level) in four disciplinary domains (Arts and Humanities, Life Sciences, Physical Sciences and Social Sciences). More than 30 disciplines are represented. All the corpus texts were originally prod-uced as coursework for Bachelor's or Master's degrees, and all have been assessed and awarded high grades by relevant subject lecturers. Research into the corpus contents has been enhanced by qualitative analysis of course documentation, and interviews with students and academics in the contributing departments.

BAWE was developed as part of a project funded by the UK Economic and Social Science Research Council (RES-000-23-0800) with the primary aim of discovering and characterising the types of writing British university students produce. For this purpose all the texts in the corpus were analysed to determine their educational and social purpose, ultimately leading to their categorisation into 13 'genre families' [Nesi & Gardner 2012]. These include not only the more traditional university genres such as Essays and Methodology Recounts (e.g. lab reports), but also genres that are more discipline-specific, such as Case Studies and Problem Questions, and genres that are often given the broad general label of 'reports' but which deserve to be distinguished from each other because they are characterised by rather different linguistic features. Access information, the corpus manual, a spreadsheet of the corpus contents, and references to associated research can all be found at www.coventry.ac.uk/bawe.

2. Multidimensional analysis (MDA)

2.1. MDA and the BAWE corpus

BAWE has now been named as a major data source in more than 80 publications, and has been examined from many perspectives. For example, corpus linguistic techniques have been used to identify and analyse stance [Henderson & Barr 2010] and cohesive devices [Zhao 2014]. The techniques of MDA complement studies of individual aspects, because they permit multiple aspects to be examined simultaneously and enable us to map their distribution across different datasets. In the case of BAWE, we can use MDA to compare texts across levels of study, disciplines and genre families. BAWE has been subject to two MDA analyses. The first was made with reference to the dimensions identified in [Biber 1988], enabling us to compare sub-sets of BAWE with those analysed in similar studies. The second identified new BAWE-specific dimensions, and is a more delicate characterisation of the writing produced by British university students.

2.2. MDA procedures

MDA involves tagging a corpus for linguistic and functional features and then determining their co-occurrence patterns by means of factor analysis. Texts are placed along dimensions depending on the extent to which they contain particular clusters of co-occurring features. There is often complementarity in feature distribution, so that groups of texts containing features belonging to one particular cluster do not contain many features belonging to another particular cluster, and vice versa. Groups of texts contrasting in this way will be placed towards opposite poles of a given dimension. Multiple dimensions cover the range of variation; registers are often similar to each other in some respects, but very different from each other in other respects.

3. Informational density in the BAWE corpus

3.1. The BAWE 1988 and BAWE2016 dimensions

The initial MDA analysis of BAWE worked with the five dimensions that emerged from Biber's study of the Lancaster-Oslo-Bergen (LOB) and the London-Lund corpora [Biber 1988]:

1. Involved versus informational production.
2. Narrative versus non-narrative concerns.
3. Elaborated versus situation-dependent reference.

4. Overt expression of persuasion.
5. Abstract versus non-abstract style.

Dimension 1 contrasts verbal and nominal styles: more 'inform-ational' texts have negative scores on this dimension, with lower fre-quencies of present tense verbs, private verbs, pro-verb DO, contract-ions, and 1st and 2nd person pronouns, and more longer words, nouns, attributive adjectives and prepositions. Dimension 1 scores are entirely negative across all subsets of BAWE, and become increasingly so in the more advanced levels of study, reaching equivalence with academic prose in the LOB corpus. This might indicate progression towards a more 'academic' writing style.

A second MDA analysis (BAWE2016) identified four dimensions unique to BAWE:

1. Compressed Procedural Information versus Stance towards the Work of Others.
2. Personal Stance.
3. Possible Events versus Completed Events.
4. Informational Density.

Texts with high scores on the BAWE2016 Dimension 4 are characterised as containing more noun groups and fewer verb groups, more nominalisations of verbs and adjectives, and a greater number of abstract nouns and long words. BAWE2016 makes finer distinctions between texts displaying different degrees of informational density, with scores ranging from 19.31 to -12.89. Again, texts at higher levels of study tend to cluster at the informational end of the dimension.

3.2. Informational density across disciplines and genres

In both analyses, informational density was most strongly associated with the Social Sciences, in contrast to Arts and Humanities discplines at the other end of the scale. Table 1 shows individual disciplines with high scores on BAWE2016 Dimension 4.

Both analyses also identified Literature Surveys and Proposals as the most informational genre families (see Table 2), although differences in the clustering of features slightly changed their rank order. (Critiques, for example, had the same score as Research Reports in BAWE2016 Dimension 4, but displayed slightly more 'involved' characteristics on BAWE1988 Dimension 1). Literature Surveys contain summaries of key works, and this may help to explain their high informational load: student writers may, consciously or un-consciously, adopt the style of the professional writers they are reviewing.

Case Studies, Proposals and Research Reports tend to be comparable to professional/expert academic genres, written to meet the specifications of industry or a research community.

Table 1. **The highest scoring disciplines on BAWE2016 Dimension 4**

Discipline	Score
Politics (*n* = 110)	2.75
Economics (*n* = 96)	2.33
Business (*n* = 146)	1.92
Medicine (*n* = 80)	1.88
Architecture (*n* = 9)	1.70
Hospitality etc. (*n* = 92)	1.68
Law (*n* = 134)	1.47
Sociology (*n* = 110)	1.22

Table 2. **Genre families most associated with informational texts**

Genre Family	BAWE1988	BAWE2016
Literature Survey (*n* = 35)	–17.91	1.62
Proposal (*n* = 71)	–16.42	1.69
Case Study (*n* = 189)	–16.40	1.10
Research Reports (*n* = 61)	–16.19	0.78

3.3. Characteristic features of informationally dense texts

Informationally dense texts are associated with the written language rather than spontaneous speech, because they require more pre-planning on the part of the writer, and more attentive decoding on the part of the reader. The information load tends to be heavier in written texts because they are permanent and relatively context free; they can be edited and revised, and they do not have to be immediately understood at the time of production. Traditionally, 'density' has been calculated in terms of the ratio between lexical words and grammatical words [Stubbs 1986]. Grammatical words are lost if verbs and adjectives are nominalised and long noun groups replace clauses. When compared to the BAWE corpus as a whole, the 20 highest

scoring texts on BAWE2016 Dimension 4 contain more than three times as many cases of head nouns preceded by at least four nouns or adjectives (4.9 as opposed to 1.5 per 10,000 words), such as:

- abnormal earnings equity value estimates;
- business corporate social responsibility performance;
- femtosecond laser-driven photodissociation reaction output;
- high cost safety-critical system development.

This subset also contains far more words of 14 letters or longer (66.4 as opposed to 34.4 per 10,000 words), such as:

- disintermediation;
- intercommunication;
- photodissociation;
- self-actualisation;
- stereoselectivity.

Language items such as these are highly discipline-specific, and serve to demonstrate the students' socialisation into their fields of study.

4. Discussion

Although informational density may be equated with academic maturity, the register is not suitable for all the purposes of university student writing. For example, texts ostensibly written for a non-expert readership, and/or describing human actions rather than abstractions, are likely to be more successful if their information load is lighter. Narrative Recounts (including reflective writing), Empathy Writing (including public engagement) and Problem Questions (typically offering legal advice on everyday issues) have high scores on BAWE 1988 Dimension 1, and low scores on BAWE2016 Dimension 4.

Perhaps most students are capable of adjusting the information load to suit the genre they are producting. There is, however, a danger of informationally dense texts becoming incomprehensible even to expert readers, if they are compressed to the extent that there are insufficient grammatical words to signal the meaning relations between lexical words. [Billig 2013] criticizes the over-use of nouns, nominalisations, and jargon in the social sciences, arguing that university education socialises students into adopting this disciplinary style despite its communicative failures. This is something we should be wary of; there may sometimes be a fine line between succinct, discipline-specific, information-rich text, and wilful obfustication.

References

1. *Biber D.* (1988), Variation across Speech and Writing. Cambridge University Press.
2. *Billig M.* (2013), Learn to Write Badly: How to succeed in the social sciences. Cambridge University Press.
3. *Henderson A., Barr R.* (2010), Comparing indicators of authorial stance in psychology. students' writing and published research articles. In: Journal of Writing Research 2 (2), 245–2.
4. *Nesi, H., Gardner, S.* (2012), Genres across the Disciplines: Student writing in Higher Education. Cambridge, Cambridge University Press.
5. *Stubbs M.* (1986), Lexical Density: A technique and some findings. In: R. M. Coulthard (ed.) Talking about Text, Discourse Analysis Monographs 13, English Language Research, University of Birmingham, pp. 27–42.
6. *Zhao C.* (2014), Lexical Cohesion of Sino-British College Students' EAP Writing. In: Theory and Practice in Language Studies, 4 (10). pp. 2123–2128.

Hilary Nesi
Coventry University (UK)
E-mail: h.nesi@coventry.ac.uk

O. Scrivner, I. Trapido, J. Lee

TEXT MINING TOOLKIT FOR DIGITAL CORPORA

Abstract. Access to large digitized collections present a new opportunity to digital humanities researchers. Novel visualization methods help discover new insights and hidden patterns that are often hard to detect for a human eye. Researchers have also access to a variety of software to perform these tasks. However, most programs are built for different purposes and often require programming skills. Our project addresses these issues by using a dynamic web framework Shiny for developing a user-friendly multipurpose web application.
Keywords. Text mining, visualization, corpora.

1. Introduction

Exploratory analysis has been widely used in quantitative research since the 1970s [Tukey 1977]. This type of analysis is often *open-ended* and provides a general impression of data, such as interaction between variables, preliminary trends, and data summary [Russell et al. 1993]. Furthermore, recent advances in text mining techniques and visualization made it possible to extend these methods to raw data, e. g., written documents and literary collections. This type of data has been traditionally analyzed by using *close reading* approach and was limited to word frequencies, concordances, and key-word-in-contexts. As a result, exploratory analysis has started gaining attention in the digital humanities field. Several studies have shown that these techniques can be used to analyze language data at the sentence level (e. g., grammatical patterns), as well as at the document level [Muralidharan et al. 2012, Jockers 2013]. Many available tools, however, are built for different purposes and often require programming skills (even for software installation). In addition to these technical issues, researchers do not have an interactive control over available GUI tools. That is, the user is limited to a predefined set of stopwords, preprocessing steps, and model parameters. Finally, there is a growing demand for text mining literacy in the digital humanities fields. To our knowledge, the existing tools are built to perform data analysis, providing only *factual* knowledge. To improve text mining literacy, the learner should have access to *conceptual* and *metacognitive* knowledge. In this project, we propose to address these issues by developing a dynamic user-friendly web application that provides the user with full control over preprocessing steps, such as stopwords removal, stemming, data cleaning, as well as selection of parameters for a variety of text mining tasks, such as clustering and topic modeling.

The remainder of this paper is organized as follows. In section 2 we will describe existing text mining methods and tools for digital humanities and

their limitations. Section 3 will introduce Shiny, a dynamic framework that is used to build our application. We will describe implemented features and functionalities of our tool in section 4, followed by presentation of use cases in section 5. Finally, we will provide our conclusion in section 6.

2. Text Mining Tools for Digital Humanities

As the volume of digital collections continues to grow, traditional digital humanities methods become increasingly ineffective. To facilitate large scale data analysis, various visual methods have been put forward, such as word clouds, heatmaps, trends, and network graphs [Janicke et al. 2015: 7–9]. For example, word clouds were applied to the stylistic analysis of *The Making of America* [Clement et al. 2009] and Federal Budget Speech of Australia [Dann 2008], whereas heatmaps and network graphs were used to look at the distribution and relationship of literary characters in novels [Oelke et al. 2012]. While many software packages are available, there exist only a few web-based tools, e. g. *Voyant*[1] and *TAPoRware*[2], that do not require installation and programming skills.

With recent advances in computing, more sophisticated models have become available to researchers, such as machine learning algorithms for topic modeling and cluster analysis. However, the use of these techniques in mainstream humanities research remains infrequent, as there exists a gap between text mining approaches and humanities methods of inquiry. Some scholars point to the need for developing interactive user interfaces and more interactive user control [Eisenstein et al. 2013, Blei 2012]. *Paper Machine* is one of the recent tools providing such an interactive visualization. *Paper Machine* is a plug-in for the bibliographic management software Zotero.[3] Built with Django web framework, this software allows for querying from Zotero library by date, document title, or location. The graphical output represents an interactive layout of documents, which are organized spatially and chronologically. In addition, the user can add or remove topics or documents, thus refining their research. Another recent tool is *TopicViz* and its successor *TOME* that aims to offer visual topic analysis for digital humanities collections. However, this tool is still under development and unavailable for users.

[1] http://voyant-tools.org/

[2] http://taporware.ualberta.ca/

[3] http://papermachines.org/

Despite the growing interest in these methods for digital humanities, their use remains infrequent in the mainstream research due to several factors. First, the available tools are limited and restricted to a specific data type. Second, the user has a very limited control over data preprocessing and model selections. In the next section, we will demonstrate how our solution helps overcome these limitations.

3. Shiny Framework and ITMS Application

The ITMS application is built with R as a back-end and Shiny app as a front-end. R is an open source programming language that is connected via Shiny app with HTML, CSS, and JavaScript. That is, the ITMS application takes user inputs through a web browser, processes the data on the server, and returns the result to the user. Furthermore, Shiny app uses Bootstrap framework, which is a commonly used framework in modern web application development, popular for its responsive UI, user friendly interfaces, and easy development. Finally, the web framework format enables users to access Shiny application from any browser and device (e. g., PC, OS, Linux, and mobile device). We believe that providing this accessibility and platform independence is one of the key strengths of our toolkit. The second strength of this application is its flexibility for future development. Based on users' needs and R packages availability, new functions can be easily added and deployed to the server. Our next section will introduce *Interactive Text Mining Suite*[4] (henceforth, ITMS), a successor of the pilot project [Scrivner et al. 2016].

4. ITMS Features and Functionality

At the heart of the nascent field of digital humanities lies the idea of applying computational methods to the analysis of large collections of digitally available textual data. Such textual data come in a variety of formats (PDF, TXT, CSV, XML, JSON, HTML, etc.), and can range in scale from small collections of articles stored on the researcher's personal computer or in the cloud to massive digital text collections, available either from commercial organizations, such as ProQuest and Google, or nonprofit groups, such as Project Gutenberg, the Internet Archive, and HathiTrust. Given this diversity, one of the goals of this project was to extend ITMS to allow users to import data in a variety of formats. Our text mining

[4] https://interactivetextminingsuite.com

toolkit now supports not only imports in PDF and text formats, but also in other formats that are commonly used for data interchange, such as CSV, JSON, and XML. ITMS can handle such structured data in two ways. First, it can be imported as metadata to accompany the main text files and can be subsequently used for topic modeling (Figure 1 — left panel). Second, it can be imported as a source of data for further processing and analysis. The system can currently process XML/MODS formats (Metadata Object Description Schema), which are used in library applications, and Google Books JSON schema. In addition to supporting direct uploads of JSON data files, the system was further extended to serve as a client for Google Books API: it allows the user to enter a query, encodes it, passes it to the API and parses the returned JSON data (see Figure 1 — right panel).

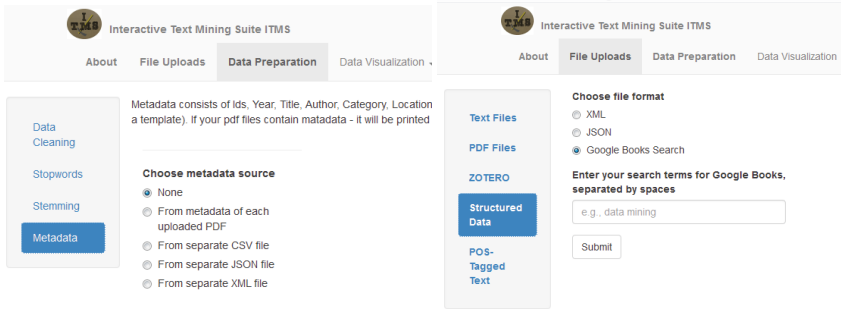

Fig. 1. **Data Import and Preprocessing**

The pilot version included some of the core text-processing tools. In this round of development, the user interface was redesigned to present this functionality in a much more logical and user-friendly fashion and new functionality was added (for example, stemming in multiple languages). Thus, a user can now perform a variety of text filtering tasks, including conversion of text to lower case, removal of punctuation and html tags, stemming, and stop word removal. The system's dynamic interactive interface allows maximum flexibility and control over each of the preprocessing steps: for example, the user has the option to use the default stop word list, upload his own, or manually select words to be removed from the corpus. Once the basic preprocessing is done, the user can move on to higher-level analysis, such as topic modeling and clustering. The current version implements two topic modeling methods: Latent Dirichlet Allocation [Blei 2012] and Structural Topic Models [Roberts et al. 2016]. Finally, cluster analysis allows users to select the agglomeration method and the distance measure and

explore how individual texts within a collection are grouped using a cluster dendrogram.

5. Implementation

The overarching goal of the current work on ITMS was to expand its user base and make the system a one-stop shop for both basic text processing tasks and advanced text analysis. This was achieved in two ways: by offering maximum flexibility of data imports and by simplifying the user interface, making it easy to use even for novice users. In its current version, the tool can support a variety of tasks. First, ITMS enables exploratory analysis of literary texts and it can be performed both at micro and at macro levels. For example, it is possible to look at word usage patterns, punctuation frequency, word and sentence length, and other stylistic signals, as well as word usage in context, topics coverage in the corpus, or texts grouping. Second, users can perform the analysis of scholarly articles, tweets, and blogs and gain insight into major themes covered in the collection, their changes over time, or partition into groups. Third, ITMS offers the ability to analyze bibliographic metadata. Jockers suggests that library metadata "has been largely untapped as a means of exploring literary history" and could "reveal useful information about literary trends" [Jockers 2013]. We believe that ITMS can facilitate such metadata analysis not only by providing text analysis tools, but also by helping users obtain bibliographic data (via Google API). Finally, ITMS can be used as a teaching tool. The system provides an interactive way to illustrate the core NLP tasks.

6. Conclusion

In recent years, we have seen a growing interest in the use of exploratory tools for digital corpora. However, many of the existing tools are unable to integrate the "synthesis of computational and humanistic modes of inquiry" [Eisenstein et al. 2013]. To address these needs, we have developed a user-friendly application for exploratory analysis providing a wide range of analytical and graphical tools, as well as the ability to control the preprocessing steps. In addition, the accessibility of our web application promotes the use of advanced text mining techniques, as researchers are not constrained by programming skills, memory limitation, or platform dependency.

References

1. *Blei D.* (2012), Probabilistic Topic Models. In: Communications of the ACM, 55(4), pp. 77–84.
2. *Clement T., Plaisant C., Vuillemot R.* (2009), The Story of One. In: Humanity Scholarship with Visualization and Text Analysis. Relation, 10, pp. 84–85.
3. *Dann S.* (2008), Analysis of the 2008 Federal Budget Speech: Policy, Politicking and Marketing Messages.
4. *Eisenstein J., Klein L.* (2013), Reading Thomas Jefferson with TopicViz. In: Towards a Thematic Method for Exploring Large Cultural Archives. Scholarly and Research Communication, 4/3.
5. *Jänicke S., Franzini G., Cheema M. F., Scheuermann G.* (2015), On Close and Distant Reading in Digital Humanities. In: A Survey and Future Challenges. Eurographics Conference on Visualization (EuroVis), pp. 1–21.
6. *Jockers M. L.* (2013), Topics in the Digital Humanities: Macroanalysis. In: Digital Methods and Literary History. University of Illinois Press, Urbana, IL, USA.
7. *Muralidharan A., Hearst M. A.* (2012), Supporting Exploratory Text Analysis in Literature Study. Literary and Linguistic Computing, 27/4.
8. *Oelke D., Kokkinakis D., Malm M.* (2012), Advanced Visual Analytics Methods for Literature Analysis. In: Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 35–44.
9. *Russell D. M., Stefik M. J., Pirolli P., Card S. K.* (1993), The Cost Structure of Sensemaking. In: INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems, pp. 269–276.
10. *Scrivner O., Davis J.* (2016), Topic Modeling of Scholarly Articles. In: Interactive Text Mining Suite. Proceedings of Dialogue 2016.
11. *Tukey J. W.* (1977), Exploratory Data Analysis. Addison-Wesley.

Olga Scrivner
Indiana University (USA)
E-mail: obscrivn@indiana.edu

Irina Trapido
University of Illinois Urbana-Champaign (USA)
E-mail: trapido2@illinois.edu

Jay Lee
University of Illinois Urbana-Champaign (USA)
E-mail: lee9@illinois.edu

TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING: "TAIGA" SYNTAX TREE CORPUS AND PARSER

Abstract. The "Taiga" project unites the corpus and the syntactic parser, being created in a new field of the corpus linguistics: the material obtained primarily meets the needs of machine learning, rather than linguistic search. The authors consider in detail the methodology for constructing the corpus, balance, volume and composition of its' segments, format and quality of tagging – which meets the current requirements for the development of tools for processing Russian language. Within the framework of the project, the creation of a large and open-source syntactic corpus in the Universal dependencies format is planned.
Keywords. Corpus construction, web corpus, syntax parsing, machine learning, corpus representativity, parsers for Russian.

1. Introduction

Modern corpus linguistics relies heavily on the concept of web as corpus [Kilgarriff 2001], which regards the web as an inexhaustible source of linguistic and extralinguistic data in a vast number of languages. Text data on the Internet is already presented in electronic form and is available for downloading and searching through search engines. Under such favorable conditions, almost everyone is able to assemble a web corpus for their work, and the number of such big corpora (billions of words) grows every year: this is the WaC family [Baroni et al., 2009], Aranea [Benko 2014], the TenTen corpora [Jakubicek et al., 2013], and for the Russian language — the General Internet Corpus of Russian language [Belikov et al., 2013a]. At the same time, as noted in the work [Sharoff 2006], the newest web-corpora are often opposed to the national ones, which are expected to be more representative, or, according to [Sichinava 2002], "culturally representative". Texts for national corpora are selected by semi-automatic methods, with manual correction, and therefore, such data collections are expensive to build and still they exist only for a small number of languages. Thus, against the background of the growing volume of data, the volume of useful and accurately collected data remains small and expensive to receive. Therefore, there is a shortage of linguistic resources available for downloading and changing, which is more felt by developers rather than by lexicographers and linguists. But big, and at the same time, pure data in open-source is a guarantee of creating good NLP-tools for the language.

2. Actual corpus production principles

Annotated corpora today are needed not only for linguists, but for machine learners as well. For Russian language there already exists a number of corpus resources (the subcorpus with resolved homonymy of RNC, Open corpora of Russian, the subcorpus with the resolved homonymy of GICR, Syntagrus, etc.), but each of them has some hidden disadvantages. The main criteria that are relevant from the point of view of the engineering approach to linguistic development are the following:

1) open source or Creative Commons licence — the ability to use the material at one's discretion, to modify it, to add new data to it, and publish the re-ranked data is very important in the development process.
2) sufficient volume of the data (more than 10 million words) — thus it is possible to collect implicit information, for example about rare words and their compatibility, syntactic behavior of individual words and different meanings of homonyms.
3) minimum share of errors embedded in the data itself — this includes both the sufficient quality of linguistic markup, and the completeness of meta-text information, the ability to uniquely separate one segment or genre from another, to find out their balance.
4) representativeness — for each development task the data should be sufficient and it should represent all possible variability in unbiased proportions.
5) solvability in a given metric — adequacy of data composition and its' features to the task.

Today developers still have to collect the data meeting these criteria themselves, doing the same chain of downloading and tagging the texts from the Internet, having different degrees of understanding of linguistic data preprocessing.

3. Modern web-corpora for machine learning

As noted, the cheapest way to collect a voluminous training text collection is to resort to Internet resources. For Russian language, large collections of web corpora are assembled — projects like RuTenTen and Aranea Russicum, which are not available for downloading, unlike resources based on Common Crawl, but all these corpora are crawled from an unbalanced set of links and represent the "black box" statistically. The situation is slightly

better with GICR [Lagutin et al. 2016], however, only 2 million words are available for download.

Within our project, we offer a new corpus that meets the aforementioned development requirements. Taiga Corpus is available for download and modification under the Creative Commons license (CC BY 4.0)[1]. The corpus is assembled from open sources, which are selected for covering the main branches of NLP development: training of syntactic and morphological parsers, extraction of named entities, studying of readability, automatic text attribution, genre definition, thematic modeling, chat-bots training, sentiment analysis, etc. The volume of the corpus in beta-version is 50 million words, in the full version — 500 million words. This size for a balanced corpus has all the potential to show results equival to those for a larger but less accurate one (as showed, for instance, in [Kutuzov, Andreev 2015]). The texts are enriched with metatextual information, which is important for specific development tasks (more in Section 5). Annotation is fully automatic and is achieved using the combination of parsers: by now we have concerned MALT-Parser, SyntaxNet, UDPipe and some others, having examined their quality on literary data. Therefore, the corpus will be even more convenient for machine learning needs as being already applied to some relevant algorithms and accordingly cleaned.

4. Format

The corpus is stored in xml format in UTF-8 encoding with all the relevant metainformation tags. For each text, indent and paragraph structure is kept as in source. All the texts from each source separately have gone deduplication by URL, and are also filtered for non-UTF symbols, html-tags, non-breaking space, etc. by the BeautifulSoup[2] Python package.

5. Segments and features

By now, our corpus contains data from 8 resources, all documenting normative and modern Russian language (we considered "modern Russian" as a language of speakers younger than 60 years old). For training, we have chosen 3 main segments, all differing stylistically and by word frequency: literary texts, news and "other" — various resources for specific needs of NLP. Literary texts come from Russian Magazine Hall, and have such metatextual features as date, title, author, URL; news are taken from Lenta.ru, Interfax

[1] https://creativecommons.org/licenses/by/4.0/

[2] https://en.wikipedia.org/wiki/Beautiful_Soup_(HTML_parser)

and Komsomolskaya Pravda, and have as metaproperties date, title, rubrics, thematic tags, named entity tags, URL, sometimes author name; "other" is extracted from Stihi.ru and Proza.ru (only recommended authors), NPlus1, TV Subs. As literary texts and news are good for parser training, news and Stihi.ru, Proza.ru are also convenient for thematic modelling, news have also main named entity tags to train extraction, NPlus1 has expert-annotated text "difficulty", which can be used for readability studies, and TV Subs contain a lot of dialogues from films and are good for chat-bot training. Table 1 represents the distribution of main text parameters in our corpora:

Table 1

resourse	words	texts	authors	mean text length	rubrics
Stihi Ru	750 217	4167	107	177	34
Proza Ru	20 513 805	7527	82	2729	36
NPlus1	1 692 326	7696	34	221	26
Interfax	6 579 301	48 107	0	137	8
Koms.Pravda	5 000 341	45 503	652	109	986
Lenta Ru	7 001 491	34 399	0	198	38
TV Subs	28 403 842	3965	0	7163	0
Magazine Hall	216 763 813	47 629	0	4551	346

By now the corpora size is about 290 millions of words. The main ratios of the segments are presented in Diagram 1:

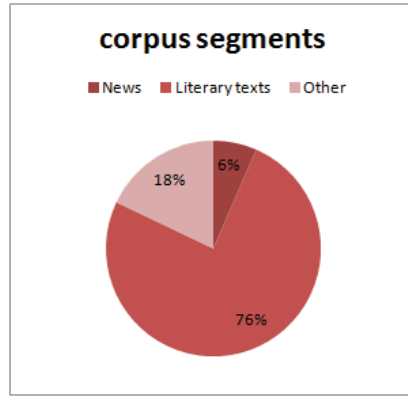

Diagram 1

Some of the data is distributed not quite normally, as you can see on Diagram 2 — yet we are going to fix such metaattribute problems when collecting the smaller version of 50 mln words.

Diagram 2 shows the distribution of texts by their difficulty in NPlus1:

Diagram 2

6. Future work

As we believe, morphological and syntactic standard for a big corpora for machine learning should be 1) concise 2) widely-adopted and compatible with international formats 3) suitable for rapid, consistent annotation by a human annotator 4) suitable for computer parsing with high accuracy 5) must be easily comprehended and used by a non-linguist (the last three points are Manning's Laws [Nivre 2016])). A new standard for multilingual morphological and syntactic tagging — Universal Dependencies[3] (UD) meets all the mentioned requirements. UD initiative has already developed treebanks for 40+ languages with cross-linguistically consistent annotation and recoverability of the original raw texts and now seeks to become the main annotation paradigm for many languages and the main evaluation tracks using it[4].

[3] http://universaldependencies.org/

[4] http://universaldependencies.org/conll17/

Our goal is to develop an open-source dependency parser (Taiga parser) and to obtain our corpus data tagged morphologically and syntactically in UD 2.0 format. We hope that our work will be useful for Russian natural language processing and will help developing new tools and projects.

7. Acknowledgments

The authors are sincerely grateful to Olga Lyashevskaya, Danil Skorinkin and Anastasia Bonch-Osmolovskaya, who expressed their deep insight and helpful advice at every stage of our work.

References

1. *Kilgarriff, A*. (2001), The Web as corpus. Proceedings of Corpus Linguistics.
2. *Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E.* (2009), The WaCky wide Web: A collection of very large linguistically processed Web-crawled corpora. In: Language Resources and Evaluation, 43, pp. 209–226.
3. *Benko, Vladimir* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: Petr Sojka, Ales Horak, Ivan Kopecek and Karel Pala (Eds.), Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655.Springer International Publishing Switzerland, 2014. pp. 257–264. ISBN: 978–3–319–10815–5 (Print), 978–3–319–10816–2 (Online).
4. *Jakubicek, M., A. Kilgarriff, V. Kovar, P. Rychly, and V. Suchomel* (2013), The TenTen corpus family. Lancaster: In: Proc. Int. Conf. on Corpus Linguistics.
5. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In: Web as Corpus Workshop (WAC-8).
6. *Sharoff, S.* (2006), Creating general-purpose corpora using automated search engine queries. In: Baroni and Bernardini (2006), pp. 63–98.
7. *Sichinava D.* (2002), K zadache sozdaniya korpusov russkogo yazyka [To the task of creation of Russian corpora]. In: NTI, cep. 2, 2002, no. 12.
8. *Benko V., Zakharov V. P.* (2016), Very Large Russian Corpora: New Opportunities and New Challenges. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4, 2016.
9. *Kutuzov A, Andreev I.* (2015), Texts in, meaning out: neural language models in semantic similarity task for Russian. In: Proceedeings of the international conference "Dialogue 2015" Moscow, May 27–30, 2015.
10. *Лагутин М. Б., Куратов Ю., Копылов Н.* (2016), Статистическая обработка результатов поиска в дифференциальных корпусах. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4, 2016.
11. *Granovsky D., Bocharov V., Bichineva S.* (2010), Otkrytyi korpus: principy raboty i perspektivy [OpenCorpora: principles of work and perspectives].

12. *Nivre J.* (2016), Reflections on Universal Dependencies. Uppsala University. Department of Linguistics and Philology.

Tatiana Shavrina
NRU HSE, GICR
E-mail: rybolos@gmail.com

Olga Shapovalova
NRU HSE
E-mail: olya_shapovalova@bk.ru

A. Stefanowitsch

A LOT OF DATA: TEXTUALLY DISTINCTIVE COLLEXEMES IN A CORPUS OF SCIENTIFIC ENGLISHES

Abstract. Associations between words and grammatical patterns have been studied under various labels Such studies have consistently shown that grammatical structures are typically associated with an above-chance frequency with sets of lexical items that are often functionally or semantically motivated. The stability of such associations across text types is less clear: since vocabulary differs quite strongly depending on text type, the same would be expected of lexicon-grammar associations. In this paper, I show that such variation exists and can be used to investigate domain-specific functions of grammatical patterns as well as the functional relationship between text types.
Keywords. Collocational frameworks, collostructional analysis, text types, Scientific English, quantitative corpus linguistics.

1. Introduction

In this paper, I combine the logic of keyword analysis, a method for uncovering associations between words and text types, and collostructional analysis — specifically, distinctive collexeme analysis, a collocational method for investigating associations between words and alternating grammatical constructions. I will apply this combination to a well-studied collocational framework, [*a*(*n*) N *of*] in a corpus of Scientific English and a general corpus of (American) English, in order to determine the extent and quality of variation in lexicon-grammar associations across text types.

2. Descriptive and methodological background

By *text type* I mean here varieties defined externally by situation and topic area — roughly, what is referred to in applied contexts as «language for specific purposes», such as Business English, Academic English etc. Such text types have been investigated through keyword analysis (cf. Scott 1997 and the work building on it). They have specific vocabulary associated with them, which is unsurprising in the case of content words. However, function words also show such associations, pointing to grammatical differences between text types. That such grammatical differences exist is, of course, also known, it has been demonstrated impressively, for example, in the research tradition started in Biber (1985), where bundles of lexico-grammatical features are used to identify and categorize text types.

Distinctive collexeme analysis is one of a family of collocational methods that focus on statistical associations between words and grammatical structures (collocational frameworks, grammar patterns, constructional

idioms, constructions etc., cf. Stefanowitsch and Gries 2009 for an overview). Specifically, distinctive collexeme analysis compares association of lexical items to a functionally equivalent slot of two related constructions (for example, verbs in the ditransitive and the prepositional dative). Words that are statistically significantly associated with one of the constructions are referred to as distinctive collexemes of that construction.

In this paper, I combine this procedure with the idea of keyword analysis such that I compare the associations of lexical items to a slot in a single given construction in two (or more) text types. Specifically, I investigate the nouns associated with the collocational framework [*a*(*n*) N *of*] in Scientific English as compared to general usage; since there is growing evidence that Scientific English itself is not a monolithic text type (e.g. Biber and Gray 2016), I also investigate the nouns associated with this framework in different scientific domains. Words that are associated with a construction in one text type as opposed to another are referred to as textually distinctive collexemes of that construction in that text type.

The patterns [*a* N *of*] and [*an* N *of*], treated here as a single pattern, are two examples of sequences of two function words interrupted by variable slot for a content word called *collocational frameworks* by Renouf and Sinclair (1991). The content-words collocates (or, in terms of collostructional analysis, collexemes) of these frameworks typically come from a small number of semantic fields, suggesting that many such frameworks are (parts of) functionally motivated linguistic units. Specifically, Renouf and Sinclair find that the words in the pattern(s) [*a*(*n*) N *of*] tend to be measurements or partitives (although other possessive relations are also found) — one of the central functions of the framework seems to be quantification.

The corpora used in this paper are COCA, a 400-million-word corpus of spoken and written general American English, and FUSE-F, a 100-million+ word corpus of open access scientific research papers under development at the Freie Universität Berlin.

3. Case studies

3.1. [a(n) N of] in Scientific English

The collocational framework [*a* N *of*] (without the variant *an*) is one of three patterns investigated in Marco (2000) with respect to their occurrence in a small proprietary corpus of medical research papers. Using relative frequency as an association measure, he finds domain-specific associations that differ in their specifics from those in general usage, but that partially

conform to it in that they fall in the domain of measurements (*dose*, *group*, *measure*); in addition, he finds words that express quantifiable properties (*specificity*, *sensitivity*, *accuracy*).

Before zooming in on specific domains like medicine, I will attempt a broader and statistically more stringent replication of his study. Based on a comparison of the collocates in the framework [*a*(*n*) N *of*] in the FUSE-F and the COCA, I identified textually distinctive Scientific English and those more strongly associated with general usage. Table 1 shows the textually distinctive collexemes of the framework in Scientific English.

Table 1. **Textually distinctive collexemes of [*a*(*n*) N *of*] in Scientific vs. General English**

Collexeme	FUSE (O:E)	COCA (O:E)	Coll. Str.
function	(18008:5403)	(3062:15667)	33921.30
subset	(6063:1683)	(500:4880)	13330.10
number	(25815:15133)	(33202:43884)	9462.88
total	(10015:4219)	(6437:12233)	9168.40
consequence	(5383:1757)	(1468:5094)	8448.28
variety	(16314:8863)	(18251:25702)	7607.35
range	(7971:3342)	(5062:9691)	7357.52
set	(9533:4316)	(7298:12515)	7332.22
effect	(2527:709)	(236:2055)	5417.00
measure	(5075:2033)	(2855:5897)	5173.96
role	(2133:569)	(87:1651)	5131.59
reduction	(2428:712)	(349:2065)	4725.85
increase	(2963:983)	(870:2850)	4488.15
combination	(6518:3101)	(5575:8992)	4394.52
decrease	(1524:423)	(127:1228)	3332.13
marker	(1497:422)	(148:1223)	3171.24
overview	(2435:873)	(968:2530)	3145.88
model	(2992:1192)	(1655:3455)	3083.83
result	(9673:5867)	(13208:17014)	3034.56
inhibitor	(1173:305)	(17:885)	3027.24

Interestingly, the textually most distinctive collexemes of the framework are not overwhelmingly quantifying expressions. There are a few cases (*number*, *total,* and arguably *variety* and *range,* though these stress diversity rather than pure quantity); however, most collexemes are best characterized as relatively abstract possessive uses encoding causality (*function*, *consequence*, *effect*, *reduction*, *increase*, *decrease*, *result*) or categorization (*subset, set*). In addition, there are individual items relating to the scientific process in general (*measure*, *model, overview*) or specific scientific concepts (*marker, inhibitor*).

In contrast, as Table 2 shows, the textually distinctive collexemes of the pattern in general usage are mainly the kind of quantifying and/or partitive

Table 2. **Textually distinctive collexemes of [*a*(*n*) N *of*] in General vs. Scientific English**

Collexeme	COCA (O:E)	FUSE (O:E)	Coll. Str.
lot	(143005:107256)	(1237:36986)	78626.65
couple	(35816:27058)	(572:9331)	17163.17
kind	(16085:12592)	(849:4342)	5147.64
bit	(9466:7097)	(78:2447)	4935.59
bunch	(8179:6103)	(28:2104)	4563.59
sense	(15776:12657)	(1245:4365)	3861.93
piece	(11171:8693)	(520:2998)	3802.28
matter	(16621:13802)	(1940:4759)	2724.68
professor	(4111:3069)	(16:1058)	2273.66
friend	(3856:2878)	(15:993)	2132.47
sort	(6004:4707)	(326:1623)	1881.99
handful	(7726:6207)	(622:2141)	1852.13
way	(7100:5676)	(533:1957)	1799.63
man	(3093:2304)	(5:794)	1774.24
bottle	(3148:2366)	(34:816)	1583.73
cup	(3441:2607)	(65:899)	1571.09
glass	(3094:2342)	(56:808)	1425.40
part	(12987:11193)	(2066:3860)	1289.59
act	(3023:2307)	(79:795)	1272.20
pile	(2376:1782)	(20:614)	1232.30

expressions (*lot*, *couple*, *bit*, *bunch*, *piece*, *handful*, *bottle*, *cup, glass*, *part*, *pile*) that Renouf and Sinclair (1991) found; additionally, there are type expressions (*kind*, *sort*) and various possessive constructions from the social domain (*professor*, *friend*, *man*, *act*) — the latter being completely absent from the textually distinctive collexemes of Scientific English.

Thus, while the pattern is used for quantification in Scientific English, it is used in this way much less frequently than in general usage. This result, which may appear somewhat surprising at first glance, given that quantification plays a crucial role in scientific discourse, makes sense once we take into account the *kind* of quantification that the pattern is used for: it is used for relatively imprecise quantities like *lot*, *couple*, *bunch*, etc., which are unlikely to be used in reporting scientific results.

In sum, while the pattern serves the same range of functions both in Scientific English and in general usage, Scientific English places a greater emphasis on the relational exploits the pattern in different ways. One crucial difference to Marco's (2000) results is that the collocates identified are less domain-specific, but this is due to the fact that our corpus includes text from a broader range of disciplines, so that collexemes have a higher chance of becoming textually distinctive if they are used *across* these disciplines — they really are typical of Scientific English in general rather than any particular discipline-specific English.

3.2. [a(n) N of] across Scientific Englishes

Let us turn to a more direct (if still quantitatively more rigorous) replication of Marco's (2000) and similar studies, focusing on individual disciplines. The subcorpora for these disciplines were constructed by grouping the journals in the FUSE-F corpus into five broad categories — medicine, neurosciences, life sciences (biology and biochemistry), physical sciences (physics, chemistry, engineering) and psychology. Each subcorpus was individually compared against the COCA. Table 3 lists the top 5 textually distinctive collexemes of each discipline (this limit is due to length restrictions, see the section Data and Software below for a link to more extensive supplementary materials).

The direct comparison of individual discipline-specific Englishes with the general usage represented by COCA shows clear differences between these text types. In small part, this is due to domain-specific terminology becoming textually distinctive, as in the case of *inhibitor* for Medicine or *solution* (in the sense of «liquid mixture of a substance and a solvent») for the Physical Sciences. However, most of the textually most distinctive

Table 3. **Textually distinctive collexemes of [*a(n)* N *of*] in five Scientific Englishes vs. General English**

Collexeme	Sci. Engl. (O:E)	COCA (O:E)	Coll. Str.
Medicine			
subset	(1090:73)	(500:1517)	4811.33
variety	(3567:996)	(18210:20781)	4421.44
number	(4785:1736)	(33161:36210)	4057.96
consequence	(908:109)	(1464:2264)	2594.35
inhibitor	(422:20)	(17:419)	2464.21
Neurosciences			
function	(9008:1454)	(3001:10556)	25741.66
subset	(2703:388)	(500:2815)	8809.51
set	(4263:1397)	(7281:10147)	4738.66
consequence	(1999:419)	(1464:3044)	4120.90
total	(3599:1214)	(6434:8819)	3805.53
Life Sciences			
subset	(1534:160)	(500:1874)	5638.74
total	(3325:767)	(6434:8992)	5522.84
function	(2188:408)	(3001:4781)	4593.05
number	(7256:3176)	(33161:37241)	4491.47
consequence	(1577:239)	(1464:2802)	4071.37
Physical Sciences			
function	(720:23)	(3001:3698)	3765.48
factor	(100:6)	(814:908)	395.90
solution	(55:1)	(170:224)	310.86
decrease	(46:1)	(127:172)	268.68
increase	(76:6)	(869:939)	254.61
Psychology			
function	(5479:532)	(3001:7948)	20066.17
effect	(1073:82)	(236:1227)	4751.77
measure	(1641:282)	(2849:4208)	3588.19
set	(2263:599)	(7281:8945)	3060.81
total	(1935:525)	(6434:7844)	2527.53

collexemes are still from the semantic field Science in general, the disciplines differing in the importance that these collexemes play (for example, *subset* plays a very important role in Medicine, the Life Sciences and Neuroscience, but not the Physical Sciences or Psychology, and *decrease/increase* play a very important role in the Physical Sciences but not the other disciplines). When more than just the top five collexemes are included, the overlap of course becomes greater, but the differences in importance remain and could provide interesting insights into the relative role of particular scientific concepts in different disciplines.

To get at the domain-specific vocabulary, a more direct comparison of the texts from the different scientific disciplines *amongst each other* rather than to general usage is necessary. There are different ways in which such multiple comparisons can be achieved, in the collostructional literature, no single method has become the standard. Here, I use a method proposed by Oakes and Farrow (2007), who simply create a large two-dimensional contingency table of all lexical items and their frequencies in all corpora involved and calculate the contribution of each cell to the overall chi-square value. These chi-square components are then used as association measures. Table 4 lists for each variety the five *attracted* collexemes with the largest chi-square component (i. e. the ones significantly more frequent than expected) and the five *repelled* collexemes with the largest chi-square components (i. e. the ones significantly less frequent than expected). This tells us not only what vocabulary is preferred in each discipline as opposed to the others, but also what vocabulary is avoided.

Using this method yields an abundance of domain-specific terminology, such as *panel, dose* and *GOR* (*grade of recommendation*) for Medicine, *network* and *threshold* for Neuroscience, *homolog* and *MOI (multiplicity of infection*) for the Life Sciences, *LOD (limit of detection*), *MAAT (mean annual air temperature*) and *solution* for the Physical Sciences. Interestingly, Psychology does not have such domain-specific vocabulary among the very strongest collexemes, suggesting that it favors a more broadly accessible style of scientific writing. Of course, if we include more data, there will be domain-specific vocabulary for all fields, e. g. *illusion* and *representation* for Psychology (ranked 18th and 19th). Even the direct comparison of different Scientific Englishes against each other, however, shows that general scientific vocabulary is associated with different disciplines to different degrees. For example, the word *function* plays a very important role in Neuroscience, Physical Sciences and Psychology, but not in the other two disciplines.

Table 4. **Textually distinctive collexemes of [*a(n)* N *of*] in five Scientific English as a subtypes of Scientific English**

Attracted	Coll. Str.	Repelled	Coll. Str.
Medicine			
variety	740.01	*function*	1481.65
panel	670.87	*set*	476.91
dose	464.48	*measure*	245.84
GOR	435.94	*sequence*	157.51
inhibitor	409.05	*pair*	136.35
Neuroscience			
function	458.38	*member*	225.70
history	174.99	*variety*	156.91
network	173.80	*panel*	125.66
train	156.39	*source*	107.85
threshold	115.77	*homolog*	98.14
Life Sciences			
member	494.07	*function*	1152.75
homolog	450.91	*sense*	261.65
MOI	313.76	*measure*	249.64
total	289.24	*effect*	249.43
suite	271.31	*sequence*	178.28
Physical Sciences			
LOD	581.81	*total*	83.36
function	470.26	*subset*	72.39
MAAT	317.35	*group*	50.37
factor	286.53	*history*	31.10
solution	271.12	*role*	29.51
Psychology			
sense	1222.95	*variety*	264.26
function	1126.17	*member*	228.51
effect	688.34	*number*	228.34
sample	453.34	*inhibitor*	216.26
measure	437.67	*concentration*	189.92

Among the repelled textual collexemes in the different disciplines, we find, unsurprisingly, domain-specific vocabulary from other disciplines, for example, *history* in the Physical Sciences and *inhibitor* in Psychology. Again, however, we also find general scientific vocabulary that is avoided in particular disciplines, such as *measure* in Medicine and Life Sciences and *total* in Physics.

Interestingly, the prominent function of quantification, which was already weakly represented in Scientific English as a whole (cf. Table 1 above), is almost completely absent from the domain-specific collocates in Table 4, the only exceptions being *variety* and *dose* in Medicine. The obvious and most likely explanation is that this function is evenly distributed across disciplines, but as a consequence, the domain-specific phraseological patterns of the framework [*a*(*n*) N *of*] in Scientific Englishes are radically different from general usage not just with respect to domain-specific vocabulary, but also with respect to the dominant meaning(s) of the pattern.

3.3. Collexemes of *[a(n) N of]* as indicators of text type

To get a more general idea as to how the function of the framework [*a*(*n*) *N of*] differs across general usage and various Scientific Englishes, we can cluster text types by the distribution of collexemes within this framework in the spirit of Biber's research mentioned above. Here, I selected 1000 collexemes on an *n*-th line basis from each text type represented in COCA and each discipline in FUSE-F that had at least 1000 occurrences. These were used as a basis for a distance matrix that was submitted to a hierarchical cluster analysis.

The results are surprisingly consistent: the first main difference is between spoken English and all written varieties, pointing to differences that are not unexpected but that have not, to my knowledge, been investigated. The next split is between the non-academic text types in COCA and all Scientific Englishes, including those represented in COCA as «academic». Among the Scientific Englishes, there are various well-motivated clusters of sub-disciplines from medicine, biology and chemistry: for example, pediatrics and public health cluster together, as do neurology and psychiatry, as do immunology, oncology, endocrinology and pharmacology, which are joined by chemistry one level up. The only unexpected cluster is the one containing Physics and Psychology, which may simply be due to the fact that these two disciplines are relatively distant from the others, which form a sort of continuum from chemistry over biology to neuroscience.

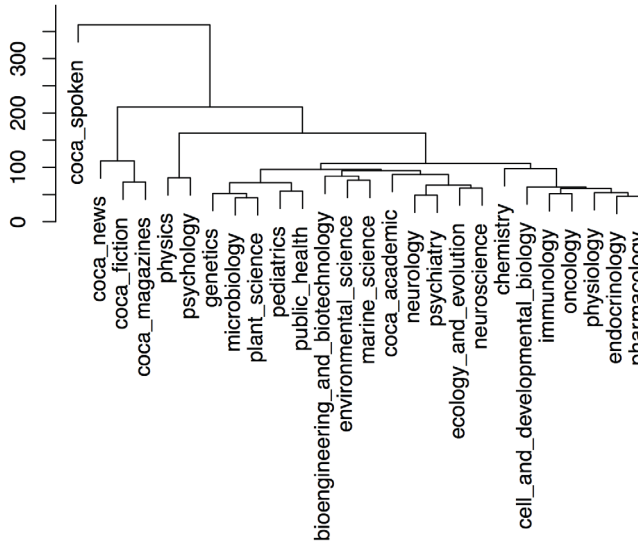

Fig. 1. **Text types in COCA and FUSE-F clustered by collexemes in the collocational framework [*a(n)* N *of*]**

4. Conclusion

The case studies in this paper have show that even highly entrenched collocational frameworks like [*a*(*n*) N *of*] may vary across text types in two ways. First, in their specific lexical associations, which differ due to domain-specific vocabulary and due to domain-specific preferences for general vocabulary. Specifically, [*a*(*n*) N *of*] is used for quantification in general usage but serves a wider range of functions in Scientific English. The studies also show that while there is good reason to assume a broad category of Scientific English that differs clearly from non-academic varieties, there are considerable differences between scientific disciplines, so that Scientific English is best thought of as a cluster of varieties that share a general scientific vocabulary but are differentiated by their specific terminology and that these differences interact with grammatical patterns systematically.

Supplementary materials

The data sets for the case studies reported here may be downloaded from www.stefanowitsch.de/data/2017alod.zip

Data and Software

1. *Davies M.* (2008-) The Corpus of Contemporary American English (COCA), 2016 ed. (commercial version). Provo (Utah), 2016.
2. *Flach S.* (2017), {collostructions}. An R implementation for the family of collostructional methods, v 0.0.10., www.bit.ly/sflach
3. *R Development Core Team* (2017), R: A language and environment for statistical computing, v. 3.3.3. www.R-project.org.
4. *Stefanowitsch A, Flach S.* (2017), The Frontiers Free University Scientific English corpus (FUSE-F), Beta. Berlin, 2016.

References

1. *Biber D.* (1985), Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. In: Linguistics 23, pp. 337–360.
2. *Biber D., Gray B.* (2016). *Grammatical complexity in Academic English: Linguistic change in writing.* Cambridge, 2016.
3. *Marco M. J. L.* (2000), Collocational frameworks in medical research papers: a genre-based study. In: English for Specifc Purposes 19 (1), pp. 63–86.
4. *Oakes M. P., Farrow M.* (2007), Use of the Chi-Squared test to examine vocabulary differences in English language corpora representing seven different countries. In: Literary and Linguistic Computing 22 (1), pp. 85–99.
5. *Renouf A., Sinclair J.* (1991), Collocational frameworks in English. In: Aijmer K., Altenberg B. (eds.), English corpus linguistics. London, 1991, pp. 128–143.
6. *Scott M.* (1997), PC analysis of key words — And key key words. In: System 25 (2), pp. 233–245.
7. *Stefanowitsch A., Gries St. Th.* (2009), Corpora and grammar. In: Lüdeling A., Kytö M. (eds.), Corpus linguistics: An international handbook, vol. 2. Berlin, New York, 2009, pp. 933–952.

Anatol Stefanowitsch
Freie Universität Berlin (Germany)
E-mail: anatol.stefanowitsch@fu-berlin.de

A. Vishenkova

VERBLESS *KAKOJ*-EXCLAMATIVES IN RUSSIAN: CORPUS STUDY[1]

Abstract. The paper presents a corpus study of verbless exclamatives with *kakoj* in Russian. Four types of exclamatives were distinguished. Each of them has its own morphosyntactic features and restrictions and conveys different functions, such as expressive amazement, admiration, disgust and objection. Among other particularities indicating the exclamative construction is intonation emphasizing. In the paper we attempted to distribute all possible construction types into several intonation contours according to Bryzgunova (1980)'s classification of intonation contours (IC). The research is based on the data from the Russian National Corpus (RNC).
Keywords. Exclamatives, verbless exclamatives, exclamative construction, corpus study.

1. Introduction

Exclamation is an effective tool to convey the speaker's attitude towards some state of affairs in the real world, especially when it violates the speaker's expectations. Recently, there has been an increase of interest in studying exclamatives and exclamative constructions [Michaelis & Lambrecht 1996; Koenig and Siemund 2007, 2013 a.o.,], particularly in studying verbless (or reduced) exclamatives which are claimed to be a non-prototypical clause type [Siemund 2015].

Despite the fact, that in linguistics different theoretical approaches to study constructions exist, there has not been proposed any unified description of exclamation in Russian, which would take into account the experience of the available theoretical materials. Since there is no comprehensive study on verbless exclamatives in Russian, consequently we devote the present research to this missing issue.

1.2. Research question

The paper investigates verbless *kakoj*-exclamatives in Russian language. The goal of the paper is to reveal morphosyntactic, lexical and prosodic features of exclamative *kakoj* 'what' in reduced clauses e. g., *Kakoj zamečatel'nij večer!* in the framework of the theory of Construction Grammar (CG) relying upon the data from Russian National Corpus (hereinafter RNC) (www.ruscorpora.ru).

[1] This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE).

2. Analysis of verbless exclamative constructions

After the data analysis following types of constructions were distinguished (all in all, more than 4600 exclamative sentences were analyzed):

1. *Kakoj* NP! (*Kakoj zamečatel'nij večer!*)
2. *Kakoj* NP AP! (*Kakoj večer zamečatel'nij!*)
3. *Kakoj* particle NP! (*Kakoj že večer!*)
4. *Kakoe* (particle) NP/VP/AP/AdvP/NumP! (in the context of objection and disagreement)

2.1. Kakoj NP! and Kakoj NP AP!

Kakoj plus NP construction is a 'basic' and tone-neutral construction type, because of ellipsis or word order change it is possible to get other constructions. *Kakoj* NP AP! is derived from the basic construction via movement of an AP from an NP to the right periphery of a clause. In Russian word order with an adjective preceding a noun is more desirable and the most frequent in RNC (1518 out of 1589 instances).

Due to ellipsis, it is also possible to distinguish constructions *Kakoj* with a single noun or a single adjective. Ellipsis of the noun in the construction is possible only under the condition that the referent was mentioned in the context, therefore, it can be restored (1). From a pragmatic point of view, this means that the referent is active in the consciousness of the Speaker and the interlocutor.

(1) Kakoj krasivyj dom! Kakoj bol'šoj!

However, in these constructions there are some morphosyntactic limitations. Firstly, ellipsis of abstract nouns is absolutely ungrammatical (2). Secondly, noun ellipsis in verbless exclamatives with two adjacent adjectives is infelicitous (3). Thirdly, position of a noun after *kakoj* in constructions with two adjacent adjectives is absolutely impossible (4). Adjectives can not occur in comparative and superlative forms and it is ineligible to use relative and possessive adjectives. If so, we deal with exclamatives in the context of objection.

(2) Kakaja bol'šaja l'ubov'! → Kakaja l'ubov'! *Kakaja bol'šaja!

(3) Kakoj voshititel'nij töplij večer! → *Kakoj voshit'it'el'nij! *Kakoj töplij!

(4) *Kakoj večer voshititel'nij töplij!

From morphosyntactic point of view, Nominative is the most frequent case for an NP in all construction types: e.g., for clauses with a bare NP,

Nominative exclamatives comprise 1635 out of 1696 instances. Genitive and Instrumentalis are used in idiomatic expressions (e. g., *Kakogo čerta!* or *Kakimi sud'bami!*), and other cases are quite rare.

Lexically, affective nouns are the most frequent. In exclamative constructions with a bare NP such nouns comprise 1151 out of 1696 instances. Furthermore, the number of negative tone nouns[2] prevails (784 instances out of 1151).

2.2. Kakoj particle NP!

This type of construction conveys a negative tone attitude of the speaker towards some state of affairs and contains particles *že*, *už*, *tam*, or *tut* or their combinations. All morphosyntactic features and limitations are the same as for *kakoj* plus NP (AP) exclamatives.

2.3. Exclamatives in the context of objection

Exclamatives with objection or negation contain indeclinable neutral-gender *kakoj* or *kakoe* which shows cross-categorial compatibility and allows disagreement of a constituent in person, number and case with the interrogative pronoun. The disagreement can occur if the degree of describing quality is higher on the scale (5) or if this degree is on the opposite side of the scale (6) (from Speaker's point of view). Interestingly, in objections Speaker almost always repeats precisely the world to which he or she reacts (5–6).

(5) — *Eto očen' krasivij dom. — Kakoj krasivij! On prosto užasen!*

(6) — *Ona simpatičnaja? — Kakoe simpatičnaja! Ona neverojatno krasivaja ženšina!*

These constructions are compatible with NPs, VPs, AdvPs, NumPs, non-gradable APs and with comparatives/superlatives of gradable APs.

3. Prosodic Features and Functions

In the paper all exclamative construction types were classified according to Bryzgunova (1980)'s intonation contours (IC) classification. Moreover, relying on the contexts where exclamatives can occur several functions of exclamative clause were distinguished: admiration (*Kakoj talant!)*, expressive amazement *(Kakoj bol'šoj dom!),* disgust (*Kakaja gadost'!*) and

[2] Examples of negative tone nouns: *užas, gadost', pozor, nesčast'e*, etc.

objection (see § 2.3). These functions are intonationally marked: IC-3, IC-5, IC-6 mark the contexts of admiration, amazement and disgust and these are for all construction types, excluding exclamatives with negation, while IC-7 is used only for the context of objection.

4. Conclusion

In this research we analyzed verbless exclamatives which start with *kakoj* in Russian language. It has been acknowledged that *Kakoj* plus NP is a 'basic' construction type. Nominative case turned out to be the most frequent (if not the solely possible) case. Based on corpus data, new functions of exclamatives were identified, such as admiration, expressive amazement, disgust and objection. Particular intonation contour depends on above-mentioned functions and inseparable from them. Exclamatives with negation can be considered as a distinct clause type, because they show cross-categorial compatibility.

Present study has a research contribution to exclamatives, particularly to verbless exclamative constructions in Russian.

References

1. *Bryzgunova E. A.* (1980), *Russkaya Grammatika* [Russian Grammar]. Moscow.
2. *Koenig E., Siemund P.* (2007). In: Language typology and syntactic description, vol. 1, pp. 276–324. Cambridge.
3. *Koenig E., Siemund P.* (2013). Satztypen des Deutschen. Berlin, pp. 846–873.
4. *Michaelis L.* (2001). Exclamative constructions. In: M. Haspelmath, E. Koenig, W. Oesterreicher, W. Raible (eds). Language typology and language universals. An international handbook. Berlin, New York.
5. *Michaelis L., Lambrecht K.* (1996). Toward a Construction-Based Theory of Language Function: The Case of Nominal Extraposition. In: Language 72:2, pp. 215–247.
6. *Siemund P.* (2015). Exclamative clauses in English and their relevance for theories of clause types. In: Studies in Language, 39:3, pp. 698–728.

Anna Vishenkova
National Research University Higher School of Economics (Russia)
E-mail: annaramon59@gmail.com

И. В. Азарова, Е. Л. Алексеева, Д. М. Миронова
I. V. Azarova, E. L. Alexeeva, D. M. Mironova

КОРПУСНОЕ ИССЛЕДОВАНИЕ РУКОПИСНОЙ ТРАДИЦИИ СЛАВЯНСКОГО ЕВАНГЕЛИЯ

CORPUS STUDY OF SLAVONIC GOSPELS MANUSCRIPTS

Аннотация. В статье рассматривается использование корпуса славянских рукописей, созданного на базе издания Евангелия от Матфея по 28 рукописям, для текстологического исследования. Нормализованная орфография и выделенные в тексте узлы разночтений позволяют для любого фрагмента текста строить матрицу разночтений и проводить кластерный анализ для выявления динамики разбиения рукописей на текстовые группы на протяжении текста евангелия.
Ключевые слова. Корпус, славянское Евангелие, кластерный анализ.

Abstract. The critical edition of 28 Slavic Gospels was used as a basis for creating a tagged corpus. Normalized spelling and division of the mss texts into variation units makes possible the automatic construction of variation matrices for any fragment of the text for the further cluster analysis of the mss, thus revealing possible changes of mss groupings throughout the text of the Gospels.
Keywords. Corpus, Slavic Gospels, cluster analysis.

1. Корпус славянских вариантов Евангелия

Над Славянским проектом, ставящим своей целью научное изучение рукописной традиции славянского Евангелия, с 1993 г. работает коллектив ученых СПбГУ и ИРЯ РАН под руководством А. А. Алексеева [Азарова и др. 2015].

Изучение славянских рукописей евангелий насчитывает более 200 лет, и за это время стало понятно, что построить их генеалогическую классификацию крайне трудно: во-первых, от древнего периода (X–XI вв.) дошло всего 5 рукописей, во-вторых, существовала так называемая контролируемая текстологическая традиция, когда переписчики в качестве источника использовали 2–3 текста антиграфа. Для классификации рукописей по типу текста было решено использовать кластерный анализ [Алексеева и др. 2014].

Из рукописных хранилищ 14 стран было привлечено более 1000 рукописей Евангелия X–XVI вв. Для проведения кластерного анализа были выбраны два фрагмента — из Евангелия от Матфея и Евангелия от Иоанна — соответственно, из первой и второй половины евангельских кодексов. Поскольку большое количество поздних рукописей содержало практически идентичный текст, анализу было подвергнуто 650 рукописей Евангелия от Иоанна и 525 — Евангелия от Матфея.

На материале Евангелия от Иоанна было выявлено 7 типов текста, на материале Евангелия от Матфея — только 5, так как между некоторыми типами текста граница оказалась размытой. Кластерный анализ позволил отобрать около 30 рукописей, характеризирующих разные типы текста, для включения в критическое издание этих евангелий [Алексеев 1998, Алексеев 2005]. Далее были подготовлены к печати и ждут издания Евангелия от Марка и Луки.

В основу издания положен текст Мариинского евангелия начала XI в. (РГБ, фонд 87 (Собрание В. И. Григоровича), 6 (М., 1689)), в научном аппарате последовательно приводятся в нормализованной старославянской орфографии все разночтения по 27–28 рукописям. Такая организация издания сделала возможным его преобразование в структурированный корпус, содержащий нормализованные тексты включенных в издание рукописей [Азарова и др. 2015: 81–84].

2. Выявление группировок рукописей в разных частях текста евангелия

Наличие корпуса позволяет ставить новые исследовательские задачи.

Как было отмечено выше, славянские книжники при переписывании текста евангелия пользовались не одним источником, что препятствует установлению генеалогических связей между рукописями. Этот же фактор осложняет и типологическую классификацию: как известно из наблюдений над рукописями, начало рукописи демонстрирует намного более активную редактуру и правку, чем конец, и именно поэтому при реализации Славянского проекта кластерный анализ был проведен на двух фрагментах из разных частей рукописей.

Первоначально сравнение рукописей и формирование матрицы узлов разночтений для последующего кластерного анализа делалось вручную. Теперь, когда около 30 текстологически значимых рукописей доступны в электронном виде в единой орфографии с размеченными разночтениями, становится возможным провести более тонкий кластерный анализ. На этом этапе анализа задается окно в 300–400 узлов разночтений, сдвигая которое по всей длине текста, становится возможным отследить динамику группировок рукописей в тексте евангелия в целом. При этом исходные матрицы разночтений формируются автоматически.

Литература

1. *Азарова И. В., Алексеева Е. Л.* (2015), Использование аппарата критического издания Четвероевангелия для создания корпуса славянских переводов Евангелия. Структурная и прикладная лингвистика, Вып. 11. СПб, с. 75–85.
2. *Алексеев А. А.* (ред.) (1998), Евангелие от Иоанна в славянской традиции. СПб.
3. *Алексеев А. А.* (ред.) (2005), Евангелие от Матфея в славянской традиции СПб.
4. *Алексеева Е. Л., Азарова И. В., Миронова Д. М.* (2014), Кластеризация рукописей на базе совпадения разночтений как основа публикации славянской традиции. Материалы XLIII Международной филологической конференции. Секция прикладной и математической лингвистики, СПб., с. 10–22.

References

1. *Azarova I. V., Alekseeva E. L.* (2015), Ispol'zovanie apparata kriticheskogo izdanija Chetveroevangelija dlja sozdanija korpusa slavjanskih perevodov Evangelija. [Transforming the Critical Edition of the Slavic Gospels into a Corpus of Slavic versions of the Gospels]. In: Strukturnaja i prikladnaja lingvistika, Vyp. 11. [Structural and Applied Linguistics, Vol. 11] SPb., pp. 75–85.
2. *Alexeev A. A.* (ed.) (1998), Evangelie ot Ioanna v slavjanskoj tradicii [Gospel according to John in Slavic Tradition]. St Petersburg.
3. *Alexeev A. A.* (ed.) (2005), Evangelie ot Matfeja v slavjanskoj tradicii [Gospel according to Matthew in Slavic Tradition]. St Petersburg.
4. *Alexeeva E. L., Azarova I. V., Mironova D. M.* (2014), Klasterizacija rukopisej na baze sovpadenija raznochtenij kak osnova publikacii slavjanskoj tradicii [MSS Clusterization according to Common Readings as a Basis for the Critical Edition of Slavic Tradition]. In: Materialy XLIII Mezhdunarodnoj filologicheskoj konferencii. Sekcija prikladnoj i matematicheskoj lingvistiki [Proceedings of the 43rd International Philological Conference. Section of Applied and Mathematical Linguistics]. St Petersburg, pp. 10–22.

Азарова Ирина Владимировна
Azarova Irina
E-mail: ivazarova @gmail.com

Алексеева Елена Леонидовна
Alexeeva Elena
E-mail: el.alexeeva @gmail.com

Миронова Дина Марковна
Mironova Dina
E-mail: 235dina @gmail.com

Санкт-Петербургский государственный университет (Россия)
Saint-Petersburg State University (Russia)

И. В. Азарова, Е. Л. Алексеева, Е. А. Рогозина
I. V. Azarova, E. L. Alexeeva, E. A. Rogozina

ДИСКУРСИВНАЯ «НОРМАЛИЗАЦИЯ» ТЕКСТОВ СЕВЕРНОРУССКИХ ЖИТИЙ

DISCURSIVE “NORMALIZATION” OF NORTH RUSSIAN HAGIOGRAPHIC TEXTS

Аннотация. В докладе рассматриваются варианты стереотипного построения житийных текстов севернорусских святых, которые составляют основную часть Санкт-Петербургского корпуса агиографических текстов. Предлагается, что существование прототипической схемы жития, отвлеченной от конкретных бытовых особенностей, приводило к полному или частичному воспроизведению других житийных текстов. Выявление подобных цитат в текстах в многообразии их грамматического и лексического варьирования позволит применить к исследованию житийных текстов методы текстологического анализа.
Ключевые слова. Санкт-Петербургский корпус агиографических текстов СКАТ, древнерусская агиография, текстологический анализ.

Abstract. The paper deals with text patterns of biographies of North Russian saints comprising a major part of St Petersburg Corpus of Hagiographic Texts. It is suggested that adhering to a conventional hagiographic structure devoid of specific details of everyday life resulted in recurrent repetitions of certain fragments from one text to another. Cataloguing these repeated text fragments in their varying grammatical and lexical form will make it possible to study text structures using methods of textual criticism.
Keywords. SCAT, St Petersburg Corpus of Hagiographic Texts, Old Russion hagiography, text analysis.

1. Представление текстов житий в корпусе СКАТ

Основной целью Санкт-Петербургского корпуса агиографических текстов [СКАТ] является представление электронных изданий севернорусских житийных текстов XV–XVII вв., подготовленных к изданию и таким образом прошедших стадию смысловой интерпретации и сопоставления с другими списками (вариантами) повествования.

Для проведения работ по автоматической обработке текстов корпуса они записаны в базовый xml-формат, в котором зафиксировано написание слов рукописи в оригинальной и упрощенной орфографии. В виде идентифицируемых единиц текста выступают слова, цифровые обозначения и знаки препинания.

Базовое представление текстов корпуса используется в качестве «полигона» для постановки и последующего решения задач аннотации корпуса в отношении морфологической формы [Алексеев и др. 2011], синтаксических связей [Алексеева 2014; Азарова и др. 2012] и дискурсивной структуры житийных текстов [Азарова и др. 2016].

2. Дискурсивная структура житийного текста

В отношении построения текста житий существует довольно много интересных литературоведческих и философских концепций, которые мы стремимся использовать в системе структурирования фрагментов текста. Отмеченные нами ранее элементы определенного сходства [Рогозина 2015] и даже заимствования фрагментов северно-русских житий М. Л. Тузов предлагает рассматривать как проявление средневекового мышления в виде «репрезентации бытия по подобию», т. е. реализацию осмысленных подражаний Иисусу Христу [Тузов 2011: 149].

Применительно к содержанию агиографических текстов существовало мнение, что в них будет представлена некая картина быта как энциклопедия средневековой жизни, однако довольно скоро выяснилось, что житийные тексты тяготеют к определенной схеме, отвлеченной от конкретных бытовых особенностей, что приводит к значительному числу заимствований, «скрытых, зашифрованных цитат» [Кузьмина 2004: 65], при этом «цитирование облекается в конспиративные формы», чтобы скрыть его от глаз непосвященных [Кузьмина 2004: 61]. По мнению О. Н. Бахтиной, «агиографический стиль… наделялся идеальной природой, … формировалась "норма" христианского поведения, вырабатывались определенные приемы представления нормы жизни настоящего христианина — "умилительная чувствительность", "цветистая, патетическая фразеология", панегиризм и лиризм» [Бахтина 2009: 48].

Первоочередная задача изучения данного явления заключается в выявлении, систематизации и описании методов варьирования подобных цитат.

Сопоставляя житийные тексты в корпусе мы отметили, что существует некий «прототипический» вариант структуры жития и способов выражения компонентов этой структуры, что выражается в практически дословном повторении фрагментов. В Таблице приводится фрагмент такого построения на примере двух рукописей из нашего корпуса в сопоставлении с житиями Феодосия Печерского и Сергия Радонежского.

Применение автоматических методов поиска таких повторов затруднено, как видно, вариантами орфографического написания слов, заменой грамматических форм и синтаксическими трансформами. Применяя метод выявления разночтений, необходимо выбрать базовый текст и привести тексты к морфологически и орфографически

Таблица. **Сходство текстов четырех житий**

Житие Феодосия Печерского	Житие Сергия Радонежского	Житие Корнилия Комельского	Житие Александра Свирского
егда же ли пакы кого слышааше бесѣдующа дъва ли или трие съшедъшеся въкупѣ.	ег(д)а же ли кого слышаша бесѣдующа, два или трiе съше(д) ше(с) вкупѣ, или смѣхы ткуща, w се(м) убо негодоваше, и sѣло не терпя таковыа вещи, рукою своею ударяше	аще ли кыя обрѣташе бесѣдующихъ и г̀лющи(х) не на ползу д̀ши. и праз(д) нословящи(х). о си(х) негодоваше.	и ег(д)а же ли пакы слыша(ше) ко(г) бесѣдующа. два или три, съше(д)ши(х)ся вкупе.
то же ту ударивъ своею рукою въ двьри	въ двери, или въ wконцо потолкавъ, w(т)хожаше,	и ударяа пръсты въ оконце.	тои(ж)де ту ударяше въ двери
ти тако w(т) хожааше. назнаменавъ тѣмь свои приходъ.	симъ wбразо(м) назнаменавъ тѣ(м) свое к ни(м) прихоженiе и посѣщенiе	симъ назнаменуя свои прихо(д).	и тако w(т)хож(да) ше. назнамена(в) тѣ(м) свои прихо(д).

нормализованному виду. Поэтому мы предполагаем использовать автоматическую лемматизацию текста (в настоящее время реализовано для существительных) с морфологической спецификацией (тегсетом). В качестве базового в нашем корпусе будет Житие Сергия Радонежского, на материале которого будут сформированы идентифицированные узлы разночтений. Критический аппарат будет показывать преобразования базового текста в конкретной рукописи, что позволит в дальнейшем применить метод кластерного анализа для выявления групп текстов.

Литература

1. *Азарова И. В., Алексеева Е. Л.* (2012), Морфо-синтаксическая разметка текстов корпуса СКАТ. Информационные технологии и письменное наследие: материалы IV международной научной конференции. Петрозаводск; Ижевск, с. 6–7.
2. *Азарова И. В., Алексеева Е. Л., Рогозина Е. Л.* (2016), Дискурсивный компонент номенклатуры синтаксических групп для разметки корпуса СКАТ. El'Manuscript-2016. Rašytinis palikimas ir informacinės technologijos: VI tarptautinė mokslinė konferencija. Vilnius, c. 81–85.

3. *Алексеев В. А., Алексеева Е. Л., Касьяненко С. Е.* (2011), Грамматическая разметка в корпусе СКАТ. Труды международной конференции «Корпусная лингвистика — 2011». СПб, с. 69–73.
4. *Алексеева Е. Л.* (2014), Синтаксическая разметка корпуса древнерусских агиографических текстов СКАТ. Структурная и прикладная лингвистика. Вып. 10. СПб, с. 345–351.
5. *Бахтина О. Н.* (2009), Проблемы анализа житийных текстов русской литературы (культурно-историческая традиция и код культуры). Вестник Томского государственного университета. Филология. № 4 (8), с. 47–61.
6. *Кузьмина М. К.* (2004), О смысле скрытых цитат в древнерусской агиографии. Мир русского слова. № 2, с. 61–65.
7. *Рогозина Е. А.* (2015), Уточнение и XML-разметка сюжетной схемы житий в корпусе агиографических текстов СКАТ. Структурная и прикладная лингвистика. Вып. 11. СПб., с. 168–173.
8. СКАТ — Санкт-Петербургский корпус агиографических текстов. Санкт-Петербургский гос. университет. URL: http://project.phil.spbu.ru/scat/page.php?page=project.
9. *Тузов М. Л.* (2011), Древнерусская агиография в философском аспекте. Ученые записки Казанского университета: Гуманитарные науки. Том 153, кн. 1, с. 149–158.

References

1. *Azarova I. V., Alekseeva E. L.* (2012), Morfo-sintaksicheskaja razmetka tekstov korpusa SKAT [Morpho-Syntactic Tagging in SCAT]. In: Informacionnye tehnologii i pis'mennoe nasledie: materialy IV mezhdunarodnoj nauchnoj konferencii [Information Technologies and Textual Heritage. Proceedings of the Fourth International Conference]. Petrozavodsk; Izhevsk, pp. 6–7.
2. *Azarova I. V., Alekseeva E. L., Rogozina E. L.* (2016), Diskursivnyj komponent nomenklatury sintaksicheskih grupp dlja razmetki korpusa SKAT [Discursive Component of the Syntactic Taxonomy for the Annotation of Saint-Petersburg Corpus of Hagiographic texts (SCAT)]. In: El'Manuscript-2016. Rašytinis palikimas ir informacinės technologijos: VI tarptautinė mokslinė konferencija [Textual Heritage and Information Technologies. Proceedings of the Sixth International Conference]. Vilnius, pp. 81–85.
3. *Alekseev V. A., Alekseeva E. L., Kas'janenko S. E.* (2011), Grammaticheskaja razmetka v korpuse SKAT [Grammatical Tagging in SCAT]. In: Trudy mezhdunarodnoj konferencii "Korpusnaja lingvistika — 2011" [Proceedings of the International Conference "Corpus Linguistics — 2015"]. SPb., pp. 69–73.
4. *Alekseeva E. L.* (2014), Sintaksicheskaja razmetka korpusa drevnerusskih agiograficheskih tekstov SKAT [Syntactic Tagging in SCAT — the Corpus of Hagiographic Texts]. In: Strukturnaja i prikladnaja lingvistika. Vyp. 10 [Structural and Applied Linguistics, Vol. 10]. SPb., pp. 345–351.
5. *Bahtina O. N.* (2009), Problemy analiza zhitijnyh tekstov russkoj literatury (kul'turno-istoricheskaja tradicija i kod kul'tury) [Problems of Analysis of Hagiographic Texts in

Russian Literature (Cultural and Historic Tradition and Culture Code)]. In: Vestnik Tomskogo gosudarstvennogo universiteta. Filologija. № 4 (8) [Tomsk State University Journal of Philology. No 4 (8)], pp. 47–61.
6. *Kuz'mina M. K.* (2004), O smysle skrytyh citat v drevnerusskoj agiografii [On Meaning of Hidden Quotations in Old Russian Hagiography]. In: Mir russkogo slova. № 2 [The World of Russian Word. No 2], pp. 61–65.
7. *Rogozina E. A.* (2015), Utochnenie i XML-razmetka sjuzhetnoj shemy zhitij v korpuse agiograficheskih tekstov SKAT [Elaboration of Hagiographic Text Content Structure and Its XML-markup in SCAT]. In: Strukturnaja i prikladnaja lingvistika. Vyp. 11 [Structural and Applied Linguistics, Vol. 11]. SPb., pp. 168–173.
8. SKAT — Sankt-Peterburgskij korpus agiograficheskih tekstov [SCAT — St Petersburg Corpus of Hagiographic Texts]. Sankt-Peterburgskij gos. universitet [Saint-Petersburg State University]. Available at: http://project.phil.spbu.ru/scat/page.php?page=project.
9. *Tuzov M. L.* (2011), Drevnerusskaja agiografija v filosofskom aspekte [Old Russian Hagiography in Terms of Philosophy]. In: Uchenye zapiski Kazanskogo universiteta: Gumanitarnye nauki. Tom 153, kn. 1 [Proceedings of Kazan University: The Humanities. Vol 153, Part 1], pp. 149–158.

Азарова Ирина Владимировна
Azarova Irina
E-mail: ivazarova @gmail.com

Алексеева Елена Леонидовна
Alexeeva Elena
E-mail: el.alexeeva @gmail.com

Рогозина Елена Андреевна
Rogozina Elena
E-mail: renehorn.r@gmail.com

Санкт-Петербургский государственный университет (Россия)
Saint Petersburg State University (Russia)

Е. Г. Андреева
E. G. Andreeva

ДВУЯЗЫЧНЫЙ КОРПУС В СОПОСТАВИТЕЛЬНОМ АНАЛИЗЕ ГЛАГОЛЬНЫХ ФОРМ

PARALLEL CORPUS AS A TOOL IN CONTRASTIVE ANALISYS OF VERBAL FORMS

Аннотация. Анализ примеров английских соответствий русской аналитической формы (АФ), почерпнутых из параллельных корпусов текстов, продемонстрировал, что такая видовая форма, как простое будущее английского глагола оказывается эквивалентом русской АФ будущего времени лишь в 40 % рассмотренных примеров, тогда как в оставшихся 60 % случаев русская АФ переводится либо другими видо-временными формами английского глагола, либо модальными глаголами, либо иными способами передачи футуральности.
Ключевые слова. Параллельный корпус, будущее время, аналитическая форма, соответствие, переводческое решение.

Abstract. The data from parallel corpora aims to show correspondences between the Russian analytical form (AF) of the Future Imperfect tense and various forms English verbs. An easily predictable translation by the Future Simple tense covers mere 40% of all the forms in the sample, while the rest 60% of examples analysed demonstrate a variety of translator's decisions, ranging from present tenses to modal verbs and other ways of expressing the Future.
Keywords. Parallel corpus, the future tense, analytical form, correspondence, translator's decision.

Аналитическая форма (АФ) будущего времени в русском языке традиционно ассоциируется с простым будущим временем английского языка. Правомерность их сближения объясняется тем, что: 1) им присуще структурное сходство и 2) их объединяет функциональный критерий, выражающийся в том, что данные формы являются регулярным средством отнесения действия в будущее [Логунов 2007: 4]. В реальности, однако, оказывается, что далеко не всегда русская АФ соответствует английскому простому будущему, и преподаватель-практик может задаться вопросом, насколько верны все «привычные» утверждения грамматик и учебников по переводу.

Собственно этим и продиктована цель настоящего сообщения — продемонстрировать возможность использования параллельного корпуса в качестве источника информации о грамматической категории и о способах ее передачи при переходе от русского языка к английскому, с одной стороны, и в качестве фактической основы при обучении английскому языку, переводу и также в практике перевода, с другой.

Источником примеров для анализа стал двуязычный корпус художественных текстов, объем которого составляет около 3 млн. слово-

форм, а также более 400 примеров с ресурса «Национальный корпус русского языка». Двуязычный корпус создавался рабочей группой при Лаборатории речевого моделирования СПбГУ в 2005–2007 годах. Используемая здесь часть корпуса представляет собой собрание рассказов известных британских и американских писателей, равно как и рассказов или глав из произведений русских авторов (в частности, А. Чехова, М. Булгакова, и других), сопровождаемых профессионально выполненными переводами на русский и английский языки. Тексты выровнены по предложениям. При отборе примеров использовался метод случайной выборки, при этом общее количество примеров составило более 1200 единиц. Из каждой (русско-английской и англо-русской) части корпуса отбирались те фразы, в которых русская глагольная форма является аналитической формой будущего времени (далее — АФ), независимо от того, является ли данная форма «оригиналом» или «переводом» с английского. Таким образом, анализируемый корпус примеров представлен как русско-английской, так и англо-русской частью, причем в первом случае англоязычный переводчик подбирал наиболее подходящий с его (ее) точки зрения вариант перевода, тогда как во втором — переводным эквивалентом является русский глагол, так что АФ продиктована преференциями русскоязычного переводчика.

Все примеры были классифицированы по двум параметрам: по семантике русской аналитической формы будущего времени (как в оригиналах, так и в переводах) и по формальным характеристикам ее английских эквивалентов (оригинальным и переводным).

Согласно академической Грамматике русского языка важнейшим отличием значений русской АФ является то, что действие полностью отделено от момента речи [Русская грамматика 1980: 634–636]. Благодаря этому в семантике русской АФ будущего времени выделяются такие значения, как намерение- решимость, предположение-предсказание, сообщение о плане, предостережение, обещание, побуждение. Разделение некоторых из этих оттенков может показаться спорным. В частности, иногда трудно провести четкую грань между намерением и обещанием; не всегда очевидно различие между предостережением и побуждением. Эти же значения характеризуют разные видовые формы английского будущего времени (будущего простого, длительного, перфектного и перфектно-длительного) и некоторых форм настоящего (в частности, простого и длительного), хотя очевидно, что иногда одно и тоже значение может передаваться разными видо-временными

формами в силу специфических прагматических особенностей этих форм [Quirk et al. 1985; Петрова 2011].

Общая картина распределения оттенков значений представлена на графике 1. На АФ с семантикой намерения и решимости приходится больше всего примеров (38%), причем чаще всего, как и предполагалось, получает в качестве переводной формы Future Simple (65%), например: (1) *Впрочем, все это я* ***буду делать*** *сам. — But I* ***will do*** *all that myself.*

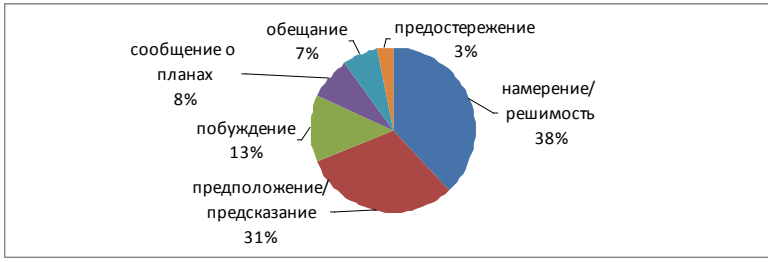

График 1. **Частотность оттенков значения русской АФ**

Неожиданным оказалось то, что около 14% случаев в английском языке получили перевод с помощью конструкции to be going + infinitive, которая традиционно считается способов передачи запланированного действия: (2) ***Теперь*** *всегда* ***буду*** *с малышками* ***дружить****. — From now on I'****m going to play*** *with girls. С* вышеприведенным вариантом перевода смыкается конструкция to be about to которая также является малопредсказуемым способом перевода русской АФ: (3) *Я* ***буду говорить*** *о небольшом инциденте, происшедшем недавно. Пожалуйста, слушайте внимательно.* ***—*** *I* ***am about to speak*** *of a small incident that occurred recently. Please listen carefully.* Еще более неожиданным оказалось такое переводческое решение: (4) *Я* ***буду здесь работать****. — "I* ***intend to work*** *here". Или, скажем, такое: (5) Итак, я этого* ***делать не буду.*** ***—*** *So you see I* ***can't do*** *it.*

АФ русского будущего времени передает значение предположения-предсказания в 31% проанализированных примеров, более 70% которых переводятся на английский язык формой простого будущего времени: (6) *Возможно, что завтра я и* ***не буду верить*** *и самому себе — вот этой своей записи. — Perhaps, tomorrow I* ***will not believe*** *even myself — even these notes.* Единичными примерами представлены варианты передачи русской АФ английскими формами простого настоящего,

длительного настоящего и будущего, *to be going to*. Около 10 % примеров составили случаи соответствия русской АФ английским модальным глаголам, причем все отобранные примеры отражают переход от английского языка к русскому: (7) *Others* ***may find*** *a better way. — Найдутся другие, они* ***будут действовать*** *искуснее.*

Семантика сообщения о планах в АФ русского будущего встречается редко (8 % случаев), и переводческие решения ограничиваются тремя формами, это — *be going to*, настоящее длительное и простое будущее. Если две первые формы вполне предсказуемы, то последняя, поскольку в ее семантике «плановость» обычно не усматривается, кажется неожиданной тем более, что этот вариант перевода составил 40 % всех таких случаев: (8) ***Служить будет*** *в депо. — He****'s going*** *to stay here and* ***work*** *at the railway yards;* (9) *Она* ***будет лечить*** *вас. — She****'ll make you well****.*

Русское будущее время в АФ со значением предостережения встретилось меньше всего и составило всего лишь 3 % всех отобранных примеров. Английские эквиваленты также не отличаются разнообразием, кроме единичных случаев это простое будущее время: (10) *Предупреждаю: ни я, ни доктор Борменталь* ***не будем*** ***с*** *тобой* ***возиться****, когда у тебя живот схватит… — I warn you that neither I nor Doctor Bormenthal* ***will lift*** *a finger for you when your stomach finally gives out . . .* Иной вариант перевода представлен формой простого настоящего: (11) *Стой!* ***Стрелять буду****! — Halt or I* ***shoot****!*

Семантика обещания в русской АФ будущего времени встретилась в 7 % и снова по большей части английским эквивалентом ее оказалась форма простого будущего времени: (12) *Честное слово, я* ***буду стараться****! Honour bright, I* ***will do my best****!* Помимо простого будущего времени английским аналогом русской АФ оказались модальные глаголы: (13) *Завтра вы* ***будете лакать*** *шампанское в неограниченном количестве. — Tomorrow you* ***can lap up*** *champagne in unlimited quantities.*

Наконец, такое значение русской АФ будущего времени, как побуждение, составило 13 % всех проанализированных примеров. Следует признать, что заметного преобладания какого-либо из вариантов перевода (односоставные глагольные структуры, модальные глаголы со значением долженствования, конструкции *to be going to*, *to be to*) среди проанализированных примеров не отмечено: (14) ***Будем****, Александр Давидыч,* ***продолжать*** *наш разговор, — сказал он. — "****Let us go on*** *with our talk, Alexandr Daviditch," he said.*

Таким образом, проведенный анализ позволил установить, что наиболее частотным эквивалентом АФ оказалась конструкция Future Simple (40 %). Также достаточно широко используются оборот *to be going to*, конструкции с различными модальными глаголами и неличными формами глагола. Настоящее длительное используется сравнительно редко, поскольку в тех случаях, когда его семантика совпадает с семантикой АФ, зачастую реализуются альтернативные формы. Немногочисленны также такие соответствия, как обороты *to be about*, *to be up*, причину этого можно усмотреть в свойственной им семантике жесткого планирования и нацеленности на ближайшее будущее. Подтвердилось предположение о том, что свойства несовершенного вида русских глаголов АФ отразятся в отсутствии среди их английских эквивалентов перфектных форм, равно как и предельных глаголов. Редкие исключения из этого правила обусловлены переводческими трансформациями. И наконец, различные оттенки значений, специфические для АФ и несвойственные английским футуральным конструкциям, могут передаваться дополнительными лексическими средствами. «Удельный вес» выбранного переводческого решения представлен на графике 2.

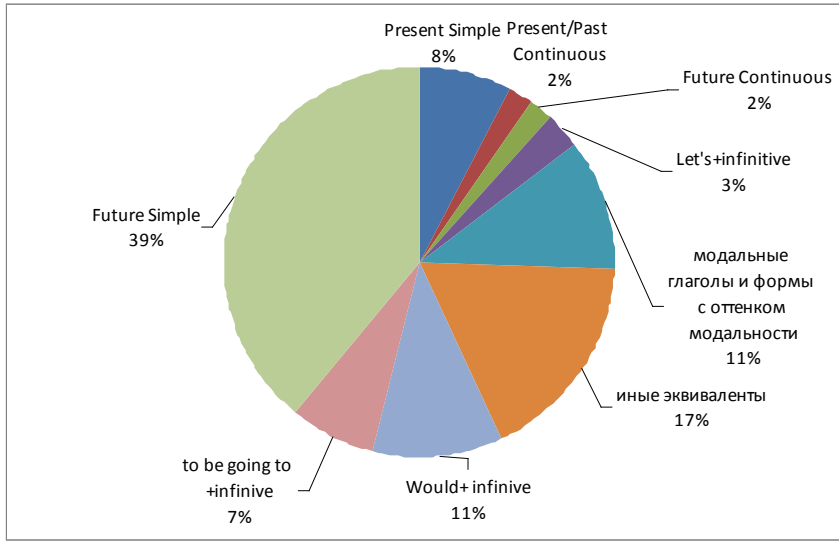

График 2. **Частотность английских эквивалентов**

Литература

1. *Логунов Т. А.* (2007), Аналитические формы будущего времени как лингвистический феномен (на материале английского и русского языков): автореф. дисс. канд. филол. наук. Кемерово.
2. Национальный корпус русского языка. URL: www.ruscorpora.ru.
3. *Петрова Е. С.* (2011), Сопоставительная типология английского и русского языков. Грамматика. Филологический факультет СПбГУ, СПб.
4. Русская грамматика (1980), Том I. Фонетика и морфология / под ред. Н. Ю. Шведовой. М.: Наука.
5. *Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.* (1985), Comprehensive Grammar of the English Language. London.

References

1. *Logunov T. A.* (2007), Analiticheskiye formy buduschego vremeni kak lingvisticheskiy fenomen (na materiale angliyskogo I russkogo yazykov) [Analytical Forms of the Future Tenses as a Linguistic Phenomenon (Russian-English)]. Avtoreferat dissertacii kandidata filologicheskih nauk. Kemerovo.
2. *Nacionalny korpus russkogo yazyka* [Russian National Corpus].
3. *Petrova E. S.* (2011), Sopostavitelnaya tipologiya angliyskogo i russkogo yazykov. Grammatika [Contrastive Typology of English and Russian. Grammar]. St Petersburg.
4. *Russkaya grammatika* (1980) (ed. Shvedova) [Russian grammar. Phonetics and Morphology].
5. *Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.* (*1985*), Comprehensive Grammar of the English Language. London.

Андреева Екатерина Георгиевна
Санкт-Петербургский Государственный университет (Россия)
Andreeva Ekaterina
St. Petersburg State University (Russia)
E-mail: e.andreeva@yahoo.com

А. Н. Баранов, М. М. Вознесенская,
Д. О. Добровольский, К. Л. Киселева, А. Д. Козеренко

A. N. Baranov, M. M. Voznesenskaja,
D. O. Dobrovol'skij, K. L. Kiseleva, A. D. Kozerenko

СТАТИСТИЧЕСКИЕ ИССЛЕДОВАНИЯ ВО ФРАЗЕОЛОГИИ: ПРОБЛЕМА ФРАЗЕОЛОГИЧНОСТИ КОРПУСОВ ТЕКСТОВ[1]

STATISTICAL STUDIES IN PHRASEOLOGY: ADVANTAGES AND SHORTCOMINGS OF CORPUS-BASED METHODS

Аннотация. В статье рассматриваются особенности фразеологии, затрудняющие ее корпусный анализ. Ставится вопрос о возможности создания фразеологически ориентированного корпуса художественной прозы, учитывающего фразеологичность авторов. Показано, что в силу варьирования частоты употребления идиом у каждого автора в рамках как различных произведений, так и различных частей одного произведения, эта задача оказывается практически нереализуемой.
Ключевые слова. Фразеология, идиоматика, частотный словарь, корпус текстов, корпусный анализ.

Abstract. The paper deals with specific properties of phraseology that complicate corpus analysis. The question in focus is how to compile an idiom-oriented corpus representing literary works that would account for the level of phraseological activity of individual writers. It is shown that phraseological activity of an author varies greatly from work to work as well as within the scope of different parts of one work, which makes the task extremely difficult.
Keywords. Phraseology, idioms, frequency dictionary, text corpora, corpus analysis.

Работая над созданием частотного словаря русских идиом [Баранов и др. 2012; Baranov et al. in print], мы столкнулись с рядом особенностей фразеологии, затрудняющих ее корпусной анализ (см. также об этом [Парина 2008; Кротова 2011]). Эти особенности можно объединить в четыре группы.

1. Употребление фразеологизмов существенно зависит от типа дискурса. Это касается прежде всего частоты употребления: наиболее насыщенными с этом отношении являются публицистические тексты, где регулярно встречаются многочисленные штампы, ср.: *ловить момент, не за горами, задавать тон, прекрасная половина, блюстители порядка, на автопилоте, белая смерть*. С точки зрения разнообразия используемых единиц выделяются устная речь и наиболее близкий к ней язык драматургии, где представлены выражения всех стилистических регистров. Кроме того, для каждого типа дискурса характерны

[1] Работа выполнена при поддержке РФФИ, грант №17-04-00420.

идиомы с определенной стилистической окраской: для публицистики — журнализмы, советизмы, элементы языка советской идеологии (*со школьной скамьи, поднимать голову, время «Ч», горячая точка, сердце нашей родины, утечка мозгов, смена караула, верхушка айберга, враг народа, важнейшее из искусств, почтовый ящик*); для разговорной речи и драматургии — разговорные идиомы, которые в Тезаурусе [Баранов и др. 2007] не имеют помет (*на коне, держать дистанцию, на птичьих правах, как миленький*), сниженные выражения (*Вася Пупкин, рога пообломать, взять за жабры*) и просторечие (*идти на попятный, хошь не хошь, чин чинарем, руки в боки*), а также жаргонизмы. Для художественной литературы более характерны книжные идиомы, а также некоторые высокие и устаревшие выражения (*смирение паче гордости, взять грех на душу, тяжелый крест, проливать кровь, предавать огню, душой и телом*).

2. Формирование полного словника идиом, отражающего современный узус, затруднено в силу ряда обстоятельств. В большинстве текстов идиомы либо встречаются относительно редко, либо круг этих выражений ограничен. Современные фразеологические словари не успевают за изменениями, которые происходят в языке: многие выражения, упомянутые в словарях, уже вышли из употребления, а, скажем, фразеологизмы в интернет-общении описаны недостаточно (*яростно плюсую, капитан очевидность, запастись попкорном, словить лулзов*).

3. Яркой особенностью фразеологии является наличие вариантов у многих словарных единиц, что делает практически невозможным автоматический поиск в тексте и подсчет вхождений многих выражений, например:

- *под крылом / крылышком*, с *какого бока / боку*; *навешивать ярлык / ярлыки* (морфологические варианты);
- *воротить нос, воротить носом; оставить мокрое место, мокрого места не оставить* (синтаксические варианты);
- *в упор не видеть / не замечать*; *спустить собак, спустить полкана* (лексические варианты);
- *иметь / затаить зуб* (на кого-л.); *есть зуб* (у кого-л. на кого-л.); *вырос зуб* (у кого-л. на кого-л.); *по разные **с**тороны баррикад* [*быть / находиться*]; *другая сторона баррикад*; *по другую сторону баррикад* [*быть / находиться*] (лексическо-синтаксические варианты).

Широкое варьирование компонентного состава и отдельных компонентов идиом затрудняет определение словарной формы. В ряде случаев неочевидно, следует ли два или три выражения считать вариантами одной леммы или разными фразеологизмами:

- *какого лешего? за каким лешим?*
- *отпустить душу / душеньку* [*на покаяние*]*; отпустить на покаяние* (кого-л.);
- *товарищ по несчастью*; *собрат по несчастью;*
- *держать руку на пульсе*; *держать руку на рычаге / рычагах;*
- *удержа / удержу не знать; удержу нет / не стало* (на кого-л.); *без удержу.*

4. Наконец, последняя группа затруднений, имеющая самое непосредственное отношение к определению частотности, создается значимыми индивидуальными различиями в употреблении фразеологизмов у разных авторов. Эти различия касаются как набора частотных выражений идиолекта, так и общей (не)склонности конкретного автора использовать идиомы.

Изначально создание фразеологических словарей в Отделе экспериментальной лексикографии Института русского языка им. В. В. Виноградова РАН опиралось на четыре жанрово-специфических корпуса текстов, которые отражают различные сферы использования современного русского языка (подробнее об этом см. [Козеренко и др. 2013]),

- Русская проза (35 млн. словоупотреблений).
- Русская драматургия (около 23 млн).
- Русская публицистика (29 млн).
- Русский детектив (около 16 млн).

Однако эти корпусы не дают полного представления о частотности идиом в языке в целом: некоторые идиомы, особенно стилистически сниженные, не встречаются вообще, а часть авторов «нефразеологичны», т. е. употребляют фразеологизмы крайне редко. Так, в Базе данных по современной идиоматике, включающей более 50 000 контекстов употребления идиом, произведения некоторых писателей представлены большим количеством контекстов, а других — весьма незначительным (подробнее об этом см. [Баранов и др. 2012]). В связи с этим возникла идея создать **фразеологически ориентированные корпусы**, исключив из рассмотрения авторов, не употребляющих идиомы. Действительно, бессмысленно исследовать какой-либо признак

(функцию) на множестве, где этот признак не определен. Для решения этой задачи некоторые авторы, чьи тексты представлены в Корпусе русской прозы, были проанализированы с точки зрения частотности употребления фразеологии. На основе полученных данных для каждого автора была выведена средняя относительная частота употребления идиом по выбранным произведениям. Результаты представлены в табл. 1.

Таблица 1. **Идиоматичность авторов (на 1 тыс. словоупотреблений)**

Авторы	Средняя частота употребления идиом
Е. Гинзбург	14
Юз Алешковский	11,1
В. Шендерович	7,6
Ю. Трифонов	5,3
Абрам Терц	5,2
Э. Рязанов	5,2
А. и Б. Стругацкие	5,1
В. Войнович	5
Саша Соколов	2,9
В. Астафьев	2,4
А. Приставкин	1,9
В. Богомолов	1
В. Распутин	0,6

В табл. 1 присутствуют как высокочастотные авторы, так и низкочастотные. Представляется целесообразным считать «3» пороговым значением частоты использования идиом, позволяющим включить автора во фразеологически ориентированный корпус. Для некоторых писателей, которые по этому критерию должны быть отнесены к числу «неидиоматичных», такая характеристика представляется сомнительной (например, для Саши Соколова). Действительно, если сопоставить частоту употребления фразеологизмов в разных произведениях одного автора, мы обнаружим достаточно большое варьирование:

Таблица 2. **Индивидуальное варьирование по частоте**

Юз Алешковский	
Кенгуру	8,3
Маскировка	11,7
Николай Николаевич	16,6
Синенький скромный платочек	14
Среднее	11,1
Саша Соколов	
Между собакой и волком	6,5
Палисандрия	1,75
Школа для дураков	2,3
Среднее	2,9
Валентин Распутин	
Деньги для Марии	0,4
Живи и помни	0,5
Пожар	0,36
Последний срок	0,46
Прощание с Матёрой	0,9
Среднее	0,6

Из табл. 2 видно, что варьирование по произведениям тем выше, чем более «фразеологичен» автор, и наоборот — варьирование по произведениям тем ниже, чем менее «фразеологичен» автор.

Кроме внешнего варьирования (различной частоты использования идиом в разных произведениях одного автора), тексты произведений часто демонстрирует и внутреннее варьирование (различную частоту употребления идиом внутри одного произведения). Так, внутреннее варьирование в повести А. и Б. Стругацких «Понедельник начинается в субботу» выглядит следующим образом: 9+10+17+9+23+8+9+32+20, где цифры обозначают абсолютную частоту из расчета на каждый последовательный фрагмент в 40 тыс. знаков.

Приведенный анализ показывает, что методика оценки идиоматичности авторов по произвольно выбранному фрагменту размером 40 тыс. знаков не работает. Необходимо исследовать частоту употре-

бления идиом во всех произведениях автора, что по очевидным причинам нереалистично. Кроме того, создание фразеологически ориентированных корпусов публицистики практически невозможно в силу того, что в состав этих корпусов входят тексты, написанные сотнями, если не тысячами авторов, каждый из которых обладает индивидуальными особенностями употребления идиом. Вопрос о создании фразеологически ориентированных корпусов русской прозы, детективов и драматургии остается открытым. Вероятно, более адекватную оценку идиоматичности авторов можно получить, используя в качестве «мерила» базовый словник из 300–400 идиом, достаточно частотных для литературного языка и отражающих основные семантические поля русской фразеологии, ср. [Баранов и др. 2007]. Таким образом, создание фразеологически ориентированных корпусов превратилось из вспомогательной процедуры, предваряющей составление Частотного словаря, в самостоятельную нетривиальную задачу, которая, возможно, будет решена в процессе создания этого словаря.

Литература

1. *Баранов А. Н., Вознесенская М. М., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* (2012), Проект частотного словаря русских идиом. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2012», с. 28–37.
2. *Баранов А. Н., Добровольский Д. О.* (ред.) (2007), Словарь-тезаурус современной русской идиоматики. М.
3. *Козеренко А. Д., Баранов А. Н., Вознесенская М. М., Добровольский Д. О., Киселева К. Л.* (2013), Квантитативные характеристики идиомы как показатель ее стилистических свойств. Корпусная лингвистика: Труды международной научной конференции, с. 309–317.
4. *Кротова Е. Б.* (2011), Корпусный подход к фразеографии. Вестн. Моск. ун-та. Сер. 19. Лингвистика и межкультурная коммуникация. № 4, с. 153–159.
5. *Парина И. С.* (2008), Корпусный анализ в исследовании фразеологии: достоинства и недостатки. Вестник МГУ. Сер. 19. Лингвистика и межкультурная коммуникация. № 1, с. 83–89.
6. *Baranov A. N., Dobrovol'skij D. O., Kiseleva K. L., Kozerenko A. D., Voznesenskaja M. M.* (in print), Frequency and style: evidences from Russian phraseology.

References

1. *Baranov A. N., Dobrovol'skij D. O.* (ed.) (2007), Slovar'-tezaurus sovremennoj russkoj idiomatiki [Thesaurus of modern Russian idioms]. Moscow.
2. *Baranov A. N., Dobrovol'skij D. O., Kiseleva K. L., Kozerenko A. D., Voznesenskaja M. M.* (in print), Frequency and style: evidences from Russian phraseology.

3. *Baranov A. N., Voznesenskaja M. M., Dobrovol'skij D. O., Kiseleva K. L., Kozerenko A. D.* (2012), Proekt chastotnogo slovarja russkih idiom [Towards a frequency dictionary of Russian idioms]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii "Dialog-2012" [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog-2012"], pp. 28–37.
4. *Kozerenko A. D., Baranov A. N., Voznesenskaja M. M., Dobrovol'skij D. O., Kiseleva K. L.* (2013), Kvantitativnye harakteristiki idiomy kak pokazatel' ee stilisticheskih svojstv [Correlation between idiom frequency and its stylistic properties]. In: Korpusnaja lingvistika: Trudy mezhdunarodnoj nauchnoj konferencii [Proceedings of the International Conference «Corpus Linguistics — 2013»], pp. 309–317.
5. *Krotova E. B.* (2011), Korpusnyj podhod k frazeografii [Corpus based approach to phraseography]. In: Vestn. Mosk. un-ta. Ser. 19. Lingvistika i mezhkul'turnaja kommunikacija [Bulletin of Moscow University. Series 19. Linguistics and Cross-cultural Communication]. no. 4, pp. 153–159.
6. *Parina I. S.* (2008), Korpusnyj analiz v issledovanii frazeologii: dostoinstva i nedostatki [Corpus based analysis in phraseology studies: values and shortcomings]. In: Vestnik MGU. Ser. 19. Lingvistika i mezhkul'turnaja kommunikacija [Bulletin of Moscow University. Series 19. Linguistics and Cross-cultural Communication]. no. 1, pp. 83–89.

Баранов Анатолий Николаевич
Baranov Anatoly
E-mail: baranov_anatoly@hotmail.com

Вознесенская Мария Марковна
Voznesenskaja Marija
E-mail: voznes-masha@yandex.com

Добровольский Дмитрий Олегович
Dobrovol'skij Dmitrij
E-mail: dm-dbrv@yandex.ru

Киселева Ксения Львовна
Kiseleva Ksenija
E-mail: xenkis@mail.ru

Козеренко Анастасия Дмитриевна
Kozerenko Anastasija
E-mail: akozerenko@mail.ru

Институт русского языка им. В. В. Виноградова РАН (Россия)
Vinogradov Russian Language Institute of the Russian Academy of Sciences (Russia)

Л. Н. Беляева
L. N. Beliaeva

ПАРАЛЛЕЛЬНЫЙ КОРПУС ТЕКСТОВ НА ОСНОВЕ ДВУЯЗЫЧНЫХ ГЛОССАРИЕВ: ПРОБЛЕМЫ СЛИЯНИЯ И КОНВЕРТАЦИИ

PARRALEL TEXT CORPORA ON THE BASE OF BILINGUAL GLOSSARIES: MERGING AND CONVERTING PROBLEMS

Аннотация. В статье рассматриваются проблемы и процедуры создания переводного словаря на основе слияния и конвертации специализированных словарей и глоссариев. Параллельный корпус, построенный на их основе, послужил основой для создания базового англо-русского словаря по сейсмозащите и для выявления дополнительных терминов, конвертация словаря в русско-английский вариант показала применимость предложенной процедуры.
Ключевые слова. Параллельный корпус текстов, двуязычный терминологический словарь, слияние и конвертация.

Abstract. The paper considers problems and procedures for translation dictionary creation to be done as the result of merging and converting specialized dictionaries and glossaries. A parallel corpus built on these glossaries was the base for English-Russian seismic construction dictionary creation and additional terms acquisition, converting this dictionary into Russian-English version had shown applicability of the procedure discussed.
Ключевые слова. Parallel text corpus, bilingual terminological dictionary, merging and converting.

Постоянное изменение научно-технической ситуации, рост суммы научных и технических знаний и технологий, возникновение новых областей знаний и модификация уже общепризнанных и известных приводит сегодня к драматическому отставанию специализированных лингвистических ресурсов, в частности, терминологических и переводных словарей. Эти ресурсы абсолютно необходимы для поддержания работы и тех, кто работает в области лингвистики и перевода (в современной терминологии — *language workers*), и специалистов в различных областях знаний.

Унификация терминов и терминосистем является одним из основных направлений прикладного терминоведения, основной задачей которого является стандартизация, упорядочение и гармонизация терминологий на различных уровнях описания и фиксации. Результатом такой унификации является создание терминологических переводных словарей, ориентированных на узкие предметные области и языки для

специальных целей. Такие словари, как правило, являются бинарными, хотя существуют и многоязычные, для нашей страны основную массу переводных словарей составляют те, в которых исходным языком является английский, словари с русским языком в качестве исходного создаются реже.

Теоретически двуязычные переводные словари необратимы, поскольку лексические системы языков, рассматриваемые как множества лексических единиц (ЛЕ), несимметричны. Отношение симметричности работает только на множествах номенов. В случае языков для специальных целей вариативность переводов ЛЕ резко уменьшается и можно говорить о введении локальной симметричности отношения между множествами терминов двух языков. Эта локальная симметричность позволяет разрабатывать методы преобразования словарей. Особое значение при этом приобретают процедуры, позволяющие «перевернуть» словарь, в принятой сегодня терминологии — конвертировать [ср. Egorova 2015]. Рассмотрим возможности гармонизации терминологии при создании исходного словаря на основе отраслевых глоссариев, а также процедуры и средства его конвертации.

Различие терминосистем разных языков, в частности, исходного языка и языка перевода определяет необходимость установления и изучения именно пар терминов вида исходный термин — переводной эквивалент, что позволяет выявить расхождения в терминополе и терминосистемах соответствующей предметной области. Установление таких расхождений и определяет дальнейшее упорядочивание, стандартизацию и унификацию терминологий разных языков, гармонизацию терминосистем этих языков, что и обеспечивает решение проблем перевода терминов и эффективность межъязыковой коммуникации [Беляева 2016].

Лексикографическое исследование в области создания переводных терминологических словарей предполагает проведение предварительной работы для отбора и описания терминов разных языков, гармонизацию этих описаний и согласование терминологических систем разных языков. Сегодня характерной особенностью технической коммуникации в активно развивающихся областях знаний является оперативное создание и публикация бумажных и/или электронных двуязычных глоссариев, отражающих представления переводчиков, осуществляющих информационную поддержку различных проектов, о двуязычной терминологии конкретной области.

Как правило, эти глоссарии фиксируют результаты перевода конкретных материалов, необходимых для поддержки разрабатываемых проектов, и даже в узких предметных областях оказываются плохо совместимыми, поскольку «привязаны» к переводу конкретных материалов особой важности, международных стандартов в том числе. Гармонизация терминологии важна для терминологии областей повышенного риска, к которым, например, относятся предметные области сейсмозащиты и сейсмостойкого строительства. В области сейсмобезопасности задача создания переводного словаря решается в соответствии с требованиями стандарта, называемого Еврокод8 (Eurocode8: Design of structures for earthquake resistance), европейского стандарта, разработанного для проектирования конструкций в сейсмической зоне. Стандарт был утвержден Европейским комитетом по стандартизации (CEN) 23 апреля 2004 года и включает основные положения и требования предыдущих стандартов.

При решении задачи гармонизации этой терминологии рядом ведущих исследовательских и проектных институтов нашей страны были разработаны и изданы следующие словари, глоссарии и правила:

- Проектирование и строительство. Понятийно-терминологический словарь к Еврокодам EN 1992 ÷ EN 1996, EN 1998, EN 1999. Рекомендации национального объединения строителей. М.: Р НОСТРОЙ, 2014. 102 с.
- Англо-русский словарь по проектированию строительных конструкций (в порядке возрастания номеров Еврокодов). М.: Межгосударственная научно-техническая комиссия по стандартизации, техническому нормированию и оценке соответствия в строительстве (МНТКС), 2011. 35 с.
- Англо-русский словарь по проектированию строительных конструкций (в алфавитном порядке). М.: Межгосударственная научно-техническая комиссия по стандартизации, техническому нормированию и оценке соответствия в строительстве (МНТКС), 2011. 29 с.
- Свод правил. Здания сейсмостойкие и сейсмоизолированные. Правила проектирования. Издание официальное. М.: Министерство регионального развития Российской Федерации, 2014. 51 с.
- Свод правил. Строительство в сейсмических районах. Правила проектирования. Издание официальное. М.: Министерство регионального развития Российской Федерации, 2014. 85 с.

- Терминологический словарь для национальных нормативных документов, реализующих Еврокоды. М.: ЗАО «ЦНИИПСК им. Мельникова», 2014. — 199 с.

Рассматриваемые материалы представляют собой базу для создания исследовательского корпуса текстов, в котором выравнивание осуществляется по ЛЕ, рассматриваемым авторами перечисленных изданий как термины, а также по толкованиям терминов и их переводам. На основе такого выровненного корпуса могут извлекаться и исследоваться кандидаты в термины, которые при последующем анализе могут вводиться в действительно гармонизированный переводной словарь.

В составе рассмотренных двуязычных глоссариев можно выделить пары типа «термин — переводной эквивалент» (*anchorage — анкерное крепление*) и пары типа «толкование новых и/или многозначных терминов на исходном языке — перевод толкования, сопровождаемый комментариями специалистов или переводчиков» (*everything that is constructed or results from construction operations — законченные строительством здания и другие строительные сооружения, а также их комплексы*). Как правило, часть переводов включают расширения, уточняющие значения терминов в конкретном тексте.

Большая часть терминов, зафиксированных в глоссариях, представляет собой составные термины, удобные для перевода конкретного документа, но не специальных текстов предметной области (*first order linear-elastic analysis without redistribution — линейно-упругий расчет первого порядка по недеформированной схеме*).

Глоссарии представляют собой особый вид текста, будучи объединены в единую систему, они могут формировать параллельный корпус и служить базой для извлечения терминов из составных конструкций и из толкований, установления терминов, имеющих несколько переводов, а также для формирования переводного двуязычного словаря и его конвертации. На основе глоссариев, разработанных в предметной области *Сейсмозащита,* создан параллельный англо-русский корпус, на основе которого разработана следующая процедура построения англо-русского переводного терминологического словаря и конвертации его в русско-английский вариант:

1) Построение параллельного корпуса на материале объединения глоссариев и терминологических словарей. В основе такого объединения лежат базовые словари, разработанные Межгосудар-

ственной научно-технической комиссией по стандартизации, техническому нормированию и оценке соответствия в строительстве (МНТКС), эти словари были дополнены информацией из всех остальных словарей.

2) Составление единого словарного массива (индекса) на основе сопоставления английских и русских лексических единиц, выявление кандидатов в английские термины на основе использования программы выделения простых именных групп в английской части и извлечения их эквивалентов.
3) Получение русско-английского словаря на основе конвертации англо-русского словаря, дополненного кандидатами в термины, пополнение словаря кандидатами в термины — простыми именными группами.
4) Работа с экспертами и уточнение списка исходных терминов и их переводов, фиксирование отношений синонимии и нормализация исходных терминов и переводов.

Особую сложность при гармонизации терминов и их переводов вызывают случаи, когда в перевод включается экстралингвистическая информация, известная специалисту в конкретной области, но отсутствующая в структуре исходного термина. Так, например, словосочетание *clearance height from roadway surfacing* переводится как *габаритная высота от поверхности дорожного полотна до нижней кромки конструкции моста*. При этом перевод термина *roadway surfacing* дополняется уточнением *до нижней кромки конструкции моста*, что связано с особенностями строительства именно мостовых и/или арочных строений. Для английского словосочетания это является имплицитной информацией, не требующей экспликации. Подобная словарная экспликация терминов, взятых вне контекста, может оказаться некорректной. Так, например, перевод терминологического словосочетания *coefficient of earth pressure at rest*, уточненный как *коэффициент бокового давления грунта в состоянии покоя*, верен только для случая описания объекта, не полностью погруженного в грунт, во всех иных случаях это давление не будут боковым и, следовательно, перевод термина не будет адекватным.

При переводе английских именных групп сложную задачу представляют конструкции типа определитель (Adjective, Participle I или Participle II) и несколько существительных подряд. К сожалению, при анализе сводного словаря приходится констатировать отсутствие единообразия в их переводах. Так, например, терминологическое слово-

сочетание *effective stiffness centre* зафиксировано с двумя переводами *центр эффективной жесткости* и *эффективный центр жесткости*. При этом термин *effective stiffness* отдельно не зафиксирован, но в словосочетаниях типа *effective stiffness of* он встречается с переводом *эффективная жесткость*, что в идеале должно приводить только к одному переводу рассматриваемого термина: *центр эффективной жесткости*.

Количественно подобные термины составляют не менее 10 % от объема словаря, что характеризует недостаточную лингвистическую компетентность их составителей. В то же время подобные глоссарии представляют собой важный источник для разработки соответствующего англо-русского переводного словаря, поскольку фиксируют не только современную терминологию, но и экспертное знание. Поскольку в исследуемых глоссариях в основном зафиксированы именные группы разного уровня сложности, то применение элементарной программы выделения простых именных групп позволяет дополнить словарь программно выделяемыми универбами и составными терминами.

Процесс редактирования состоял в изменении, удалении или добавлении заглавных единиц, соотнесении и переупорядочении их связей с английскими эквивалентами. Исходный вариант русско-английского словаря автоматически генерировался из дополненного и отредактированного англо-русского словаря с использованием основных средств операционных систем, в частности, средств работы с таблицами. Полученный словарь (индекс) представляет собой таблицу русских терминов и их переводов, отсортированных в алфавитном порядке. Процесс редактирования состоял в изменении, удалении или добавлении заглавных единиц, соотнесении и переупорядочении их связей с английскими эквивалентами. При этом модифицировались конструкции, представляющие собой дефиниции, толкования, описания ситуации, и устанавливались входящие в них термины, весь комплекс их переводов, синонимия переводов или многозначность исходного термина.

Реализация предложенной процедуры позволяет оперативно создавать переводные словари на основе имеющихся глоссариев, словари должны тщательно редактироваться, но объем этой работы не сопоставим с трудоемкостью создания переводного словаря вручную. При этом следует учитывать, что даже самая изощренная система извлечения терминов не дает окончательного варианта переводного словаря,

а предоставляет лишь удобно организованный и оперативно получаемый ресурс для работы терминолога или лексикографа.

Литература

1. *Беляева Л. Н.* (2016), Лингвистические технологии в современном сетевом пространстве: *language worker* в индустрии локализации. СПб.: Книжный дом. 134 с.
2. *Egorova K.* (2015), Editing an automatically-generated index with K Index Editing Tool // Proc. of the fourth biennial conference on electronic lexicography, eLex 2015: Linking lexical data in the digital age. Sussex, United Kingdom, from 11–13 August 2015, pp. 268–280.

References

1. *Beliaeva, L. N.* (2016), Lingvisticheskie tekhnologii v sovremennom setevom prostranstve: language worker v industrii lokalizacii [Linguistic technologies in modern network space: language worker in localization industry]. Sankt Peterburg: Knizhnyj dom.
2. *Egorova K.* (2015), Editing an automatically-generated index with K Index Editing Tool. In: Proc. of the fourth biennial conference on electronic lexicography, eLex 2015: Linking lexical data in the digital age. Sussex, United Kingdom, from 11–13 August 2015, pp. 268–280.

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

О. В. Блинова
O. V. Blinova

ПОБУДИТЕЛЬНАЯ РЕПЛИКА В ДИАЛОГИЧЕСКОМ ОКРУЖЕНИИ: ПАРЫ РЕПЛИК ТИПА «ИМПЕРАТИВ + ВЕРБАЛЬНАЯ РЕАКЦИЯ» И СПОСОБЫ ИХ РАЗМЕТКИ В РЕЧЕВОМ КОРПУСЕ

DIRECTIVE UTTERANCE IN THE DIALOGUE CONTEXT: PAIRS OF UTTERANCES OF THE TYPE «IMPERATIVE + VERBAL RESPONCE» AND THEIR TAGGING IN A SPEECH CORPUS

Аннотация. В докладе обсуждается формат разметки для пар реплик типа «императив + вербальная реакция». Предлагается учитывать: (1) информацию о внутренних свойствах и функциях реплик, (2) информацию о дискурсивном контексте, (3) информацию о сходстве между стимульными и реактивными репликами. Первые два блока информации рассматриваются подробно.
Ключевые слова. Моделирование диалога, речевой корпус, русский устный диалог, разметка, побудительные реплики, императив, пары реплик, речевой акт, диалогический акт, ручное аннотирование.

Abstract. The paper discusses the annotation format for the pairs of utterances of the type «Imp. + verbal response» in Russian naturally occurring spoken dialogue. The following kinds of information are under consideration: (1) information on internal features of the utterances, (2) information on dialogue history, (3) information on the congruence between an initial utterance and subsequent response. The first two blocks of information are discussed in detail.
Keywords. Dialogue modeling, speech corpus, Russian spoken dialogue, tags, directive utterances, imperative, pairs of utterances, speech act, dialogue act, manual annotation.

1. Побудительная реплика в диалоге

В ходе протекания диалога собеседники трактуют высказывания друг друга. Речевые реакции являются результатом интерпретации стимульных реплик в ситуации говорения. Сказанное справедливо для высказываний с разными иллокутивными функциями, в том числе — побуждений. Соответственно, семантические и прагматические свойства побудительных (в частности, императивных) реплик имеет смысл анализировать, основываясь в том числе на наличие и характер речевых реакций, вызванных этими репликами.

Корпус «Один речевой день» (ОРД) является прекрасным «полигоном» для тестирования предлагаемого подхода: этот корпус содержит преимущественно диалоги, он достаточно большой (1 млн. словоупотреблений в расшифровках) и звуковой, см. [Bogdanova-Beglarian et al. 2016].

Необходимо выработать формат разметки пар реплик типа «императивная реплика + речевая реакция» (или «речевой стимул + императивная реплика»), пригодный для применения ко всем лично-числовым формам и семантическим подтипам императива.

В настоящей статье предлагается подход к аннотации, выработанный с учётом богатого опыта по разметке диалогических актов (речевых актов в контексте диалога), накопленного в компьютерной лингвистике. Наиболее полезными при продумывании формата аннотации стали следующие схемы разметки и таксономии явлений диалога: DAMSL [Core, Allen 1997], производная от SWBD-DAMSL схема MRDA [Shriberg et. al. 2004] и семантически-ориентированная схема DiAML [Bunt et. al. 2010]. Для императивных реплик учитывается подход, реализованный в базе данных «Интонация русского диалога» [Кодзасов и др. 2006].

При разметке имеет смысл учитывать три типа информации: (1) информацию о внутренних свойствах и диалогических функциях реплик (resp. intra-utterance features), (2) информацию о диалогическом контексте (resp. inter-utterance features), (3) информацию о сходстве между стимульными и реактивными репликами (в настоящей статье не рассматривается). В разделах 2 и 3 ниже предлагается формат разметки для передачи информации типа (1) и (2). В разделе 4 кратко обсуждается формат транскрипта и представление метаданных.

2. Информация о внутренних свойствах и функциях реплик

2.1. Императивные реплики

При передаче информации о свойствах и функциях **императивных реплик** предлагаю учитывать следующие позиции:

Каузация: фактитивная vs пермиссивная, resp. реплика стимульная vs реактивная (выбор из 2 тегов).

Лично-числовая форма императива, другие грамматические характеристики глагольной словоформы, лемма (в формате *говорить*=V,несов,пе=ед,пов,2-л, присваиваются в результате автоматической разметки).

Модус: положительный императив vs отрицательный императив (выбор из 2 тегов).

Наличие неопущенного подлежащего (субъектного местоимения) (приводится местоимение).

Инклюзивная vs эксклюзивная интерпретация употребления императива (для форм 1Pl, выбор из 2 тегов).

Семантика глагола — специально отмечаются глаголы, направленные на запрос информации, глаголы говорения, глаголы мысли и эмоциональных состояний etc., ср.: *скажи-ка*, *сама подумай*, но *ты не переживай* (выбор из набора тегов; возможно одновременное присвоение нескольких помет).

Подтип императива, неодинаковые наборы значений для разных лично-числовых комбинаций 1Sg, 1Pl vs 2Sg, 2Pl (выбор из набора тегов; например, для форм 1Pl с фактитивной каузацией используются пометы «пропозитив», «гортатив», «фактитивный пермиссив»).

Значение реплики в терминах теории речевых актов (используется классический набор из [Храковский, Володин 1986] «приказ», «просьба», «инструкция», «предложение», «совет» и «разрешение»).

Функция императивной реплики в диалоге с точки зрения управления **очерёдностью говорения** (turn management — ср. императивы *слушай*, *подожди*, *постой*), управления **временем и скоростью протекания диалога** (time management, ср. *ну говори! ну озвучивай / милая! *П быстрее!*), управления **тематикой разговора** (topic management), управления **коммуникативным поведением собеседника** (partner communication management) (выбор из 4 тегов, возможно одновременное присвоение нескольких помет).

Императивы в составе формул вежливости, показателей просьбы типа *разрешите*, *позвольте* и др. (один тег, указывающий на формулу).

Релевантные просодические характеристики императивных реплик, прежде всего — указание на словоформу-носитель акцентного пика в соответствии с [Янко 2008] (приводится словоформа-носитель, ей присваиваются значения «часть речи» и «синтаксическая функция»).

Первое слово императивной реплики (приводится словоформа).

Длина реплики в словах (число).

2.2. Вербальные реакции

При передаче информации о свойствах **вербальных реакций на императивные реплики** предлагаю учитывать следующие позиции:

Реактивная реплика (выбор из двух обязательных тегов, указывающих на наличие/отсутствие реактивной реплики).

Значение реплики в терминах диалогических актов (таксономия DAMS), главные теги «договорённость» и «понимание»; дочерние теги «договорённости»: «согласие», «частичное согласие», «может быть», «отказ», «частичный отказ», «удержание» (hold); дочерние теги «понимания»: «сигналы понимания», «сигнал непонимания», «корректировка предыдущего высказывания» (выбор из набора тегов; возможно одновременное присвоение нескольких помет).

Длина реплики в словах (число).

Разница в длине между стимульной и реактивной репликами, в словах (например, +5 если реактивная реплика длиннее, -5 если реактивная реплика короче стимульной).

3. Информация о дискурсивном контексте

При передаче информации о дискурсивном контексте предлагаю учитывать:

Длину паузы, которую делает говорящий после императивной реплики, так как если каузатор ожидает вербальной реакции, он делает паузу после своей реплики, позволяя каузируемому ответить (выбор из трёх тегов для пауз трёх степеней длительности: коротких (до 0,2 с.), средних (0,3 — 0,6 с.), длинных, с обязательным указанием длительности (более 0,7 с.), см. [Du Bois et al. 1992]).

Наложения речи, которые могут говорить о несогласованной интерпретации значения инициирующей реплики каузатором и каузируемым (отрезки одновременного говорения выделяются в транскрипте, с помощью квадратных скобок маркируется начало и конец фрагмента наложенной речи в репликах каждого из участников диалога).

4. Метаданные и формат транскрипта

При расшифровке в ELAN [Wittenburg et al. 2006] текст транскрипта ОРД делится на реплики и фрагменты реплик с помощью знаков фразового и синтагматического членения «/», «//». «?», «!» «/+», «…». Каждая реплика атрибутирована определённому говорящему, характеристики говорящего и его социальная роль в конкретной ситуации коммуникации известны (изначально — приведены в базе данных MS Access).

Для наших целей удобнее использовать несколько иной формат транскрипта, при котором он **разделён на строки, каждая строка соответствует одной реплике** (фрагменту реплики) и **имеет поряд-**

ковый номер. Деление на строки и присвоение номера происходит автоматически в результате экспорта из ELAN (*.eaf) (который используется при создании ОРД) в CLAN (*.cha) [MacWhinney 2017]. Существенно, что внутри каждого файла *.cha присутствуют метаданные об участниках диалога, степень подробности которых можно варьировать.

В обязательном порядке в разделе метаданных указывается **тип дискурса** (повседневный (бытовой) vs институциональный).

Литература

1. *Кодзасов С. В., Архипов А. В., Бонч-Осмоловская А. А., Захаров Л. М., Кривнова О. Ф.* (2006), База данных «интонация русского диалога»: побудительные реплики // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции, Бекасово, Диалог 2006, с. 236–242.
2. *Янко Т. Е.* (2008), Интонационные стратегии русской речи в сопоставительном аспекте, Языки славянских культур, М.
3. *Храковский В. С., Володин А. П.* (1986), Семантика и типология императива: Русский императив, Л., Наука.

References

1. *Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A.* (2016), Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech. In: Ronzhin, A. et al. (eds.) SPECOM 2016, Lecture Notes in Computer Science, LNCS 2016. Vol. 9811, Springer, Switzerland, pp. 659–666.
2. *Bunt H., Alexandersson J., Carletta J.; Choe J.-W., Fang A. Ch., Hasida K.; Lee K., Petukhova V., Popescu-Belis A., Romary L., Soria C., Traum D. R.* (2010), Towards an ISO standard for dialogue act annotation. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17–23 May 2010, Valletta, Malta. European Language Resources Association (ELRA), pp. 2548–2555.
3. *Core M. G., Allen J. F.* (1997), Coding Dialogues with the DAMSL annotation scheme. In: Traum D. (ed.), Working notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines, Menlo Park, pp. 28–35.
4. *Du Bois J. W., Cumming S., Schuetze-Coburn S, & Paolino D.* (1992), Santa Barbara Papers in Linguistics. Vol. 4. Discourse transcription. Santa-Barbara.
5. *Kodzasov S. V., Arkhipov A. V., Bonch-Osmolovskaia A. A., Zakharov L. M., Krivnova O. F.* (2006), Database "Intonation of Russian Dialogue": imperative utterances [Baza dannykh "Intonatsiya russkogo dialoga": pobuditel'nye repliki]. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2006" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2006"], Bekasovo, pp. 236–242.

6. *MacWhinney B.* Tools for Analyzing Talk. Part 2: The CLAN Program. Available at: http://talkbank.org/manuals/CLAN.pdf.
7. *Shriberg E., Dhillon R., Bhagat S., Ang J., Carvey H.* (2004), The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In: Strube M., SidnerC. (eds.), Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, MA, pp. 97–100.
8. *Wittenburg P., Brugman H., Russel A., Klassmann A., Sloetjes H.* (2006), ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.
9. *Xrakovskij V. S., Volodin A. P.* (1986), Semantics and typology of imperative: Russian imperative [Semantika i tipologija imperativa: Russkij imperativ]. Science [Nauka], Leningrad.
10. *Yanko T.* (2008), Intonational strategies of the Russian speech from a contrastive perspective [Intonatsionnye strategii russkoj rechi v sopostaviteljnom aspekte]. Yazyki slavyanskih kul'tur, Moscow.

Блинова Ольга Владимировна
Санкт-Петербургский государственный университет (Россия)
Blinova Olga
Saint Petersburg State University (Russia)
E-mail: o.blinova@spbu.ru

Н. В. Богданова-Бегларян
N. V. Bogdanova-Beglarian

УСТНАЯ СПОНТАННАЯ РЕЧЬ: СУДЬБА НЕКОТОРЫХ ГРАММАТИЧЕСКИХ ЕДИНИЦ[1]

SPONTANEOUS COLLOQUIAL SPEECH: SOME GRAMMAR ITEMS FUNCTIONS

Аннотация. В статье обсуждаются судьбы некоторых грамматических единиц, подвергающихся в устной речи процессам десемантизации и прагматикализации. На материале корпуса повседневной русской речи «Один речевой день» (ОРД) показаны новые функции таких весьма употребительных единиц, как этикетные формы *здрасьте, пожалуйста, извини*(*те*), вводное слово *скажем*, наречие *там*, частица *вот*, конструкции *как его* (*ее, их, это*). Показана полифункциональность многих таких прагматем и многоступенчатость эволюционных процессов, приведших к их рождению. Нет сомнения, что все такого рода единицы устной речи должны быть зафиксированы, детально описаны в специальном словаре и включены в учебные материалы для русской и иноязычной аудитории.
Ключевые слова: разговорная речь, повседневная речь, языковой корпус, семантика, прагматическое значение, десемантизация, прагматикализация.

Abstract. The article is dedicated to some grammar items functions in spontaneous speech when they lose their semantics and acquire some pragmatic meaning. The research is based on the corpora material, the Corpus of Spoken Russian Language, module «One Speaker's Day» (the ORD corpus). Some new functions are demonstrated in the article, such as courtesy forms *zdras'te, pozhalsta, izvini*(*te*), parenthetic word *skazhem*, adverb *tam*, particle *vot*, collocations *kak jego* (*jejo, ix, eto*). The main object is the multyfunctional character of such items and its long evolution. No doubt such features should be described in specific dictionaries and included into books for Rus-sian and foreign students.
Keywords. colloquial speech, everyday speech, speech corpus, semantics, pragmatic meaning, desemantization, pragmatikalization.

1. Введение

Устная речь, организованная в корпус и предоставляющая тем самым исследователю большие объемы эмпирического мате-риала, демонстрирует множество явлений, которые своими кор-нями уходят в грамматику, а сегодня уже не укладываются в эти традиционные рамки. Такие единицы можно описать в терминах *прагматем* [Богданова-Бегларян 2014], а процесс их рождения уместно связать с *прагматикализацией* — переходом некоторых грамматических форм или

[1] Исследование проведено в рамках проекта «Русский язык повседневного общения: особенности функционирования в разных социальных группах», поддержанного грантом РНФ № 14-18-02070.

конструкций на коммуникативно-прагматический уровень языка, заменой их грамматического значения на прагматическое (функцию в тексте) [Günther, Mutz 2004].

2. Судьба некоторых грамматических единиц

2.1. Междометия (этикетные формы) *здравствуйте* и *пожалуйста* в редуцированном виде могут выступать в речи в функции междометных прагматем [Graf 2011; Пальшина 2015]:

- *а ты же старше Сашки / да ?*П на один год //*П ну* ***здраст*** */ на десять лет!*[2]
- *а что ж это у меня получилось-то ?*П чистый лист /* ***драсьте пожалста****!*

В такой же функции междометной прагматемы можно встретить в речи и глагольную (также этикетную) форму *извини*(*те*) (*меня*),

- *ну* ***извините*** */ и подарок Даше и пылесос это слишком;*
- *а что / а толку-то / ты вот всё снимешь / а всё лицо /* ***извини меня*** */ и так дырявое* (…) *от этих / этих* (…) *серёжек.*

2.2. Вводное слово *скажем*, часто в комбинации с другими дискурсивными единицами, способно употребляться в устной речи в различных функциях, не замеченных словарями:

- вербальный хезитатив [Богданова-Бегларян 2013]: в *принципе значит как ? вот /* ***скажем вот*** */ п… что касается азиатов например*;
- рефлексив (маркер эвфемизации) [*Богданова-Бегларян* 2015а]: *он пытается сейчас пересмотреть // *П ему <u>помогают пересмотреть</u> // *П* ***скажем так*** */ в том числе / толкают // *П это ж надо пересматривать // * П пересматривать ему не очень хочется*;
- упреждающий рефлексив (маркер эвфемизации) [*там же*]: *ну просто здесь случай / *П* ***скажем так*** *<u>не самый простой</u>*;
- дискурсив (маркер финала) [Богданова 2012]: в *общем* / (…) *я и не волновался | выдавался*(*?*) *// *П* ***скажем так***.

[2] Об особенностях орфографического представления материалов корпуса ОРД см.: [Богданова-Бегларян 2016: 242-243].

Судьба этой формы демонстрирует не только полифункциональность прагматемы, но и два витка языковой эволюции, которые к ней привели: глагол → вводное слово → прагматема.

2.3. Частица *вот* часто образует в речи устойчивую дейктическую конструкцию *вот* (…) *вот*, включающую различные указательные слова и также не замеченную словарями:

- ***вот такое вот*** / *типа это самое ∫ типа такого чего-то*;
- *она бы подписала / это* ***вот здесь вот*** *видишь это* / (…) *допущены вот ошибки.*

Комбинация вопросительной частицы *как* с местоименными словами образует в речи вербальный (поисковый) хезитатив [Богданова-Бегларян 2013]:

- *это вот* () *после*(:) *того / как я*(:) *этот* ***/ как его ?*** *@ угу // *П пролечился*(:) *вот @ угу // @ от* (*э-э*) *мочекаменной* (*э-э*) *болезни ?*
- *обязательно сейчас закрепим / все эти самые // как и тампончиком рисовать // кистью торцовой /* ***как их*** */ общие силуэты;*
- *на работе конечно*(:) *это самое … *П* ***как это*** (…) *дурдом / ну как всегда.*

Здесь снова налицо полифункциональность прагматикализованной единицы [Богданова-Бегларян 2015б; Верховцева 2016].

2.4. Круг прагматических значений наречия *там в* устной речи оказывается очень широк:

- вербальный хезитатив: *ага-га-га // *П стильную* ***там*** *всякую мебель;*
- ритмообразующая прагматема [Богданова-Бегларян и др. 2013]: *я знаю что ничего не изменится / @ угу / @ чего* ***там*** () *психовать*;
- ксенопоказатель: 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*;
- маркер-аппроксиматор: но *она сейчас тоже будет голову морочить / адрес* ***там***;
- маркер приблизительности (в ряду других подобных, в примере подчеркнуты), *грубо говоря* ***там*** */ четыре с половиной на два восемьдесят где-то.*

3. Некоторые выводы

Список подобных новаций, выявленных при анализе корпусного материала повседневной устной речи, можно продолжать (см., например: междометная прагматема *щас* или вербальный хезитатив *щас-щас*(*-щас*) *из* редуцированной формы наречия *сейчас*; маркер-аппроксиматор *туда-сюда* также из соответствующего наречия (*он* () *он мне звонит* / *типа* / *короче* / *мы всё готово* / *короче* / ***туда-сюда*** / *а я говорю* / *я в отпуске*); разнообразные прагматические значения местоимения-прилагательного *такой* (ксенопоказатель, вербальный хезитатив, изобразительный маркер, маркер-интенсификатор или деинтенсификатор) и мн. др.), но и приведенных примеров достаточно, чтобы поставить задачу создания как грамматики, так и специальных словарей русской устной повседневной речи, в которых найдут описание все подобные единицы.

Литература

1. *Богданова Н. В.* (2012), Конструкция (…) *скажем* (…) в повседневной русской речи (материалы к словарю дискурсивных единиц) // Вестник Калмыцкого института гуманитарных исследований РАН, № 2, с. 153–157.
2. *Богданова-Бегларян Н. В.* (2013), Кто ищет — всегда ли найдет? (о поисковой функции вербальных хезитативов русской спонтанной речи) // Компьютерная лингвистика и интеллектуальные технологии. По м-лам ежегодной Международной конференции «Диалог». Вып. 12 (19). В двух томах. Том 1. Основная программа конференции. М., с. 125–136.
3. *Богданова-Бегларян Н. В.* (2014), Прагматемы в устной повседневной речи: определение понятия и общая типология // Вестник Пермского университета. Российская и зарубежная филология, № 3 (27), с. 7–20.
4. *Богданова-Бегларян Н. В.* (2015а), Рефлексив в системе дискурсивных единиц русской устной речи // Мир русского слова, № 3, с. 11–17.
5. *Богданова-Бегларян Н. В.* (ред.) (2015б), Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 2. Теоретические и практические аспекты анализа. Том 2. Звуковой корпус как материал для новых лексико-графических проектов. СПб.: Филол. ф-т СПбГУ, 364 с.
6. *Богданова-Бегларян Н. В.* (ред.) (2016), Русский язык повседневного общения: особенности функционирования в разных социальных группах. Коллективная монография. СПб.: ЛАЙКА, 244 с.
7. *Богданова-Бегларян Н. В., Кислощук А. И., Шерстинова Т. Ю.* (2013), О ритмообразующей функции дискурсивных единиц // Вестник Пермского университета. Российская и зарубежная филология, № 2 (22), с. 7–17.
8. *Верховцева Т. А.* (2016), Функционирование конструкции (…) *как это* (…) в устной спонтанной речи // Коммуникативные исследования, № 3 (9), с. 11–18.

9. *Пальшина Д. А.* (2015), От десемантизации к прагматикализации (о смысловых и функциональных трансформациях редуцированных форм частотных слов в устной повседневной речи) // Вестник Пермского университета. Российская и зарубежная филология, № 3 (31), с. 34–38.
10. *Graf, E.* (2011), Interjektionen im Russischen als Interaktive Einheiten. Frankfurt am Main, 2011. 328 s.
11. *Günther, S., Mutz, K.* (2004), Grammaticalization vs. Pragmaticalization? The Development of Pragmatic Markers in German and Italian. W. Bisang, N. P. Himmelmann, B. Wiemer (eds.). // What Makes Grammaticalization? A Look from its Fringes and its Components. Berlin: Language Arts & Disciplines, p. 77–107.

References

1. *Bogdanova, N. V.* (2012), Konstrukcia (…) *skazhem* (…) v povsednevnoj russkoj rechi (materualy k slovar'u diskursivnykh jedinic) [Construction (…) *skazhem* (…) in Everyday Russian Speech (Materials to the Dictionary of Discursive Items)]. In: Vestnik Kalmyckogo instituta gumanitarnykh issledovanij RAN [Kalmyk Institute of Human Studies Herald], no. 2, p. 153–157.
2. *Bogdanova-Beglarian, N. V.* (2013), Kto ishchet — vsegda li najd'ot? (o poiskovoj funkcyi verbal'nyx xezitativov russkoj spontannoj rechi [Those Who Search, Do They Always Find? (about Retrieval Function of Hesitatives in Russian Spontaneous Speech)]. In: Komp'iuternaia lingvistika i intellektual'nye tekhnologii: Trudy mezhdun. konf. «Диалог 2013» [Computational Linguistics and Intelligent Technologies: Proceedings of the Intern. Conf. "Dialogue 2013"]. Moscow, p. 125–136.
3. *Bogdanova-Beglarian, N. V.* (2014), Pragmatemy v ustnoj povsednevnoj rechi: opredelenie pon'atia i obshchaja tipologia [Pragmatic Items in Everyday Speech: Definition Of The Concept And General Typology]. In: Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia [Perm University Herald. Russian and Foreign Philology]. no. 3 (27), p. 7–20.
4. *Bogdanova-Beglarian, N. V.* (2015a), Refleksiv v sisteme diskursivnykh jedinic russkoj ustnoj rechi [Reflexive in System of Discursive Items of Russian Oral Speech]. In: Mir russkogo slova [The World of a Russian Word]. no. 3, p. 11–17.
5. *Bogdanova-Beglarian, N. V.* (ed.) (2015b), Zvukovoj korpus russkogo jazyka kak material dl'a analiza russkoj rechi. Kollektivnaja monografia. Chast' 2. Teoreticheskie i prakticheskie aspekty analiza. Tom 2. Zvukovoj korpus kak material dl'a novyx leksikograficheskix proektov [Speech Corpus as a Base for Analysis. Collective Monograph. Part 2. Theoretical and Practical Aspects of Analysis Vol. 2. Speech Corpus as a Base for a New Lexicographic projects]. St Petersburg, 364 p.
6. *Bogdanova-Beglarian, N. V.* (ed.) (2016), Russkij jazyk povsednevnogo obshchenia: osobennosti funkcionirovania v raznykh social'nykh gruppakh. Kollektivnaja monografia [Everyday Russian Language in Different Social Groups. Collective Monograph]. St Petersburg, 244 p.
7. *Bogdanova-Beglarian, N. V., Kisloshchuk, A. I., Sherstinova, T. Ju.* (2013), O ritmoobrazujushhej funkcii diskursivnykh jedinic [On Rhythm-Forming Function of Discourse

Markers]. In: Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia [Perm University Herald. Russian and Foreign Philology]. no. 2 (22), p. 7–17.
8. *Graf, E.* (2011), Interjektionen im Russischen als Interaktive Einheiten. Frankfurt am Main, 2011. 328 S.
9. *Günther, S., Mutz, K.* (2004), Grammaticalization vs. Pragmaticalization? The Development of Pragmatic Markers in German and Italian. W. Bisang, N. P. Himmelmann, B. Wiemer (eds.). In: What Makes Grammaticalization? A Look from its Fringes and its Components. Berlin: Language Arts & Disciplines, p. 77–107.
10. *Pal'shina, D. A.* (2015), Ot desemantizacii k pragmatikalizacii (o smyslovykh i funkcional'nykh transformaciakh reducirovannykh form chastotnykh slov v ustnoj povsednevnoj rechi [From Desemantization — to Pragmaticalization (about Semantic and Functional Transformations of Reduced Forms of Frequency Words in Everyday Communication)]. In: Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia [Perm University Herald. Russian and Foreign Philology]. no. 3 (31), p. 34–38.
11. *Verkhovtseva, T. A.* (2016), Funkcionirovanie konstrukcii (…) *kak eto* (…) v ustnoj spontannoj rechi [Construction (…) *kak eto* (…), Functional in Oral Spontaneous Speech]. In: Kommunikativnye issledovania [Communication Studies]. № 3 (9), p. 11–18.

Богданова-Бегларян Наталья Викторовна
Санкт-Петербургский государственный университет (Россия)
Bogdanova-Beglarian Natalia
Saint-Petersburg State University (Russia)
E-mail: n.bogdanova@spbu.ru

В. В. Бочкарев, В. Д. Соловьев, А. В. Шевлякова
V. V. Bochkarev, V. D. Solovyev, A. V. Shevlyakova

УСТОЙЧИВОСТЬ КОЛЛОКАЦИЙ: ДИАХРОНИЧЕСКИЙ ПОДХОД[1]

COLLOCATIONS: DIACHRONIC APPROACH

Аннотация. В данной работе исследовался большой массив коллокаций на базе корпуса Google Books. Для английского и русского языков осуществлялось выделение списков коллокаций по величине показателя ассоциативной связи MI. Отобранные списки включили по двести тысяч наиболее устойчивых словосочетаний. Затем были построены временные зависимости средних значений MI по данной выборке и проанализировано изменение показателя MI на больших интервалах времени. Было выявлено, что частота встречаемости указанных коллокаций возрастает, то есть со временем они становятся неотъемлемой частью языка. Это подтверждается и тем, что по выборке условное среднее значение MI в XX веке выше значения в XIX веке.
Ключевые слова. Коллокации, частоты словосочетаний, Google Books Ngram.

Abstract. In this paper, a big amount of collocations was investigated using the Google Books corpus. The lists of English and Russians collocations were made taking into account the value of the associative relationship MI. Each list included 200000 collocations. Time dependencies of average MI values were built based on the selected lists; the change of the MI value was analyzed. It was found out that frequency use of the investigated collocations increases with time. It is also confirmed by the fact that conventional average MI value is higher in the 20th century than in the 19th century.
Keywords. Collocations, collocation frequencies, Google Books Ngram.

1. Введение

Для языка характерным является наличие коллокаций, которые представляют собой сочетания двух или более слов, имеющих признаки синтаксически и семантически целостной единицы. Изучение коллокаций, динамики их употребления представляется важным как с теоретической, так и с практической точек зрения. Методы выявления коллокаций описаны в [Захаров и др. 2010], где отмечено, что коллокации в настоящий момент играют ведущую роль в лексикографической практике. В [Клышинский и др. 2010] отмечена важность коллокаций для компьютерной лингвистики, различных задач обработки текстов. В этой работе приводится описание построения словаря словосочетаний на основе корпусов текстов.

Создание больших корпусов текстов открыло широкие возможности для изучения коллокаций. Они позволяют с довольно высокой

[1] Работа выполнена при поддержке РФФИ, грант № 15-06-07402.

точностью выявить коллокации и особенности их использования. В данной работе исследуется большой массив коллокаций на базе корпуса Google Books Ngram как с точки зрения синхронического, так и диахронического подходов с целью получения выборки из наиболее употребляемых коллокаций (биграм), наиболее объективных статистических закономерностей их употребления и описания статистической меры для вычисления силы синтагматической связанности.

2. Метод

Объем корпуса Google Books Ngram сильно отличается в различные годы. Для того, чтобы провести диахронический анализ, нужны количественные показатели, значение которых не зависит от объема корпуса. По этой причине мы не можем использовать, например, показатели t-score и log-likelihood, хорошо зарекомендовавшие себя в задаче выделения коллокаций. Мы используем взаимную информацию (MI) [Church et al. 1990], так как этот показатель является комбинацией относительных, нормированных величин. Кроме MI, мы использовали нормированную частоту (далее NF), которая определяется следующим выражением:

$$NF = \frac{f_{12}}{\min\{f_1, f_2\}}. \quad (1)$$

Здесь f_{12} — относительная частота словосочетания, а f_1 и f_2 — относительные частоты входящих в него слов. Как показывает проведенный анализ, использование показателей MI и NF дает качественно сходные результаты.

Для английского и русского языков были отобраны списки, включающие по двести тысяч наиболее устойчивых словосочетаний, а также построены временные зависимости средних значений MI и NF по данной выборке. Выделение списков коллокаций осуществлялось по значению MI, с дополнительной проверкой выбранного словосочетания по критерию отношения правдоподобия. Последнее необходимо, так как словосочетание из двух редких слов, не являющееся коллокацией, может иметь высокий выборочный показатель MI. Также при количественном анализе мы исключали словосочетания, включающие хотя бы одно служебное слово.

3. Полученные результаты

Частота встречаемости найденных коллокаций для обоих рассматриваемых языков со временем в целом возрастает, что позволяет прийти к заключению о том, что они закрепляются в сознании говорящих и становятся неотъемлемой частью языка. Более точно, суммарная относительная частота по выборке из 200 тыс. наиболее устойчивых коллокаций выросла за период с 1800 по 2000 в 2,4 раза для английского языка и в 2,8 раза для русского. Среднее по выборке значение показателей MI и NF также существенно выросло за это время. Для английского языка среднее значение MI увеличилось с 8,8 до 9,8, среднее значение NF выросло приблизительно в 3 раза (с 0,037 до 0,113). Также увеличились и медианные значения. Это показывает, что рост показателей является типичным поведением для большей части рассматриваемых коллокаций. Для русского языка наблюдаются те же тенденции. Различие состоит в том, что для английского языка средние значения MI и NF со временем возрастают практически монотонно, а для русского языка на графике наблюдаются изломы, соответствующие крупным историческим событиям (в 1917–1920 и 1985–1991). Было также проанализировано изменение показателей MI и NF отдельных слов на больших интервалах времени. Для этого подсчитывались частоты выбранных словосочетаний и входящих в их состав слов в течение двух достаточно больших временных интервалов времени. Например, для английского языка первый интервал был выбран 1800–1900, второй 1900–2000 (рассматривались также интервалы 1900–1950 и 1950–2000). По полученным частотам для каждого слова вычислялись показатели MI(1) и MI(2) для обоих интервалов времени. Затем оценивалось условное среднее значение MI(2) (на выборке 200 тысяч коллокаций) при условии, что MI(1) равно (с некоторой точностью) заданной величине. Аналогичные расчеты были выполнены и для показателя NF.

Установлено, что для обоих языков значения MI во второй период выше значения в первый период. Исключение составляет группа слов с наибольшими значениями MI, для которой в среднем показатель несколько уменьшается. Показатель NF ведет себя аналогичным образом, причем уменьшение условных средних наблюдается только в области очень больших значений — выше 0,32–0,37 для английского языка. Среди коллокаций с такими большими значениями NF преобладают специальные термины (химические, медицинские, био-

логические и т. д.), собственные имена и названия. Если обратиться к отдельным словам, приращения NF при переходе от периода 1900–1950 к периоду 1950–2000 положительны в 64 % случаев, при переходе от 1800–1900 к 1900–2000 в 72 % случаев. Такое значительное преобладание положительных приращений над отрицательными также подтверждает тенденцию увеличения устойчивости коллокаций.

4. Заключение

С использованием меры ассоциативной связанности MI для английского и русского языков подготовлены выборки 200 тысяч наиболее устойчивых коллокаций. Проанализирована статистика частот употребления большого массива коллокаций за период 1800–2000 годов. Полученные результаты указывают, что употребление отобранных коллокации в среднем становится более устойчивым. В дальнейшем также предполагается контрастивное изучение коллокаций с точки зрения частотного подхода и для большего числа языков.

Литература

1. *Захаров В. П., Хохлова М. В.* (2010), Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". М.: РГГУ, с. 137–143.
2. *Клышинский Э. С., Кочеткова Н. А., Литвинов М. И., Максимов В. Ю.* (2010), Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". М.: РГГУ, с. 181–185.
3. *Church K., Hanks P.* (1990), Word association norms, mutual information, and lexicography // Computational Linguistics, no. 16(1), p. 22–29.

References

1. *Zakharov V. P., Khokhlova M. V*. (2010), Analiz ehffektivnosti stastisticheskih metodov vijavlenija kollokatsij v tekstah na russkom jazike [Analysis of efficiency of collocation determination methods in Russian texts]. In: Komp`juternaja lingvistika I intellectual`nye tehnologii: Trydy mezhdunarodnoj konferentsii "Dialog-2007" [Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"]. Moscow, RSHU, pp. 133–143.
2. *Klyshinskij Eh. S., Rjchetkova N. A., Litvinov M. I., Maksimov Ju. V.* (2010), Avtomaticheskoe formirovanije basy sochetaemosti slov na osnove ochen bolshogo korpusa tekstov. [Automatic generation of the word combinability base on the basis of an extremely large text corpus]. In: Komp`juternaja lingvistika I intellectual`nye

tehnologii: Trydy mezhdunarodnoj konferentsii “Dialog-2007” [Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”]. Moscow, RSHU, pp. 181–185.
3. *Church K., Hanks P.* (1990), Word association norms, mutual information, and lexicography. In: Computational Linguistics, no. 16(1), pp. 22–29.

Бочкарев Владимир Владимирович
Bochkarev Vladimir
E-mail: vbochkarev@mail.ru

Соловьев Валерий Дмитриевич
Solovyev Valery
E-mail: maki.solovyev@mail.ru

Шевлякова Анна Владимировна
Shevlyakova Anna
E-mail: anna_ling@mail.ru

Казанский федеральный университет (Казань)
Kazan Federal University (Russia)

А. М. Галиева, О. А. Невзорова
A. Galieva, O. Nevzorova

СИНОНИМИЯ ОБЩЕСТВЕННО-ПОЛИТИЧЕСКИХ ТЕРМИНОВ В ТАТАРСКОМ ЯЗЫКЕ (НА КОРПУСНЫХ ДАННЫХ)[1]

SYNONYMY IN TATAR SOCIAL-POLITICAL TERMS (ON CORPUS DATA)

Аннотация. В статье анализируются общественно-политические термины в татарском языке на корпусных данных. Выделяются основные структурные типы синонимических номинаций (на уровне отдельных лексем и словосочетаний), раскрываются причины их появления (экстралингвистические и интралингвистические), приводятся количественные данные, отражающие частотное распределение и особенности функционирования таких единиц.
Ключевые слова. Татарский язык, корпус, синонимы, общественно-политическая лексика.

Abstract. The paper studies Tatar social-political terms on corpus data. The authors distinguish basic structural types of synonymous items (on word and phrase level) and describe main intralinguistic and extralinguistic causes of their emerging. Quantitative data reflecting frequency distribution and some features of functioning synonymous items is given.
Keywords. The Tatar language, linguistic corpus, synonyms, social-political vocabulary.

1. Введение

Общественно-политическая сфера — одна из наиболее динамично развивающихся сфер современной жизни, соответственно, общественно-политическая лексика находится в непрерывном развитии и постоянно обогащается новыми номинативными единицами, отражающими реалии современной жизни.

Принятие законов о языках в Республике Татарстан в 1992 г., законодательное закрепление равноправия двух государственных языков — татарского и русского — существенно расширило сферы применения татарского языка. В настоящее время основная масса татарских общественно-политических терминов являются кальками с русского языка. При переводе русских терминов появляются многочисленные семантические дублеты (синонимы и квазисинонимы), эти дублеты активно функционируют в языке и требуют своего изучения.

Материалом для исследований послужили данные Письменного корпуса татарского языка (http://corpus.tatar) и Татарского национального корпуса «Туган тел» (http://corpus.antat.ru). Корпусные исследо-

[1] Исследование выполнено при финансовой поддержке РНФ (проект № 16-18-02074)

вания фиксирует все разнообразие современной общественно-политической лексики, тогда как существующие специальные словари являются либо устаревшими [Амиров 1996], [Ганиев 1997], либо не отражают все особенности процессов, происходящих в татарском языке в настоящее время [Тагирова, Амирханов 2014].

2. Особенности синонимии татарских общественно-политических терминов

Сосуществование в татарском языке слов тюрко-татарского, арабо-персидского, русского или западноевропейского происхождения, относящихся к одному и тому же референту, является причиной появления значительного числа синонимов на уровне отдельных лексем. Таблица 1 иллюстрирует распределение семантических дублетов — слов западноевропейского и арабского происхождения — в текстах корпусных коллекций.

Таблица 1. **Распределение разноязычных синонимов в татарских текстах на примере слов западноевропейского и арабского происхождения**

Лексемы	Перевод на рус.	Количество лексем в ПК	Количество лексем в ТНК
экономик	экономический	621	676
икътисади	экономический	27 351	22 274
политик	политический	1 005	1 536
сәяси	политический	25 011	24 489
республика	республика	316 667	258 433
җөмһүрият	республика	2 479	1 631

Появление подобных семантических дублетов вызвано совокупностью причин. В настоящее время большая часть татарских общественно-политических терминов создается путем калькирования (покомпонентного перевода) соответствующих русских терминов. Разработка и перевод терминологии осуществляется различными группами специалистов под влиянием несовпадающих мировоззренческих и идеологических установок. Различные предпочтения разработчиков терминологии (тюркоцентризм, ориентация на арабский/персидский, русский или западноевропейские языки) ведут к тому,

что в языке используются различные номинации для одной и той же реалии (экстралингвистические причины).

Появление синонимических терминов обусловлено также особенностями лексической, словообразовательной и грамматической систем самого татарского языка (интралингвистические причины). В частности, конкурирующими синонимами в ряде случаев являются лексические дублеты собственно-татарского происхождения, образованные от разных татарских основ при помощи татарских словообразовательных средств. Ниже приводится пример двух таких лексем — *эшкуар и эшмәкәр* со значением ‘предприниматель, бизнесмен’ с количественными характеристиками по корпусным коллекциям (ПК:ТНК), а также наиболее частотные производные слова, образованные на базе данных лексем.

эшкуар (8606:11280)/ предприниматель, бизнесмен;
эшкуар (22671: 13876)/ предприниматель, бизнесмен;
эшкуарлык (2641: 3 852)/ предпринимательство;
эшмәкәрлек (6454: 4207)/ предприниматель, бизнесмен.

На уровне многокомпонентных номинаций и словосочетаний лексическая синонимия осложняется структурными факторами, обусловленными типологическими особенностями татарской грамматики. В частности, в тюркских языках регулярными являются синонимические модели словосочетаний по типам:

ADJ +N и N + N, POSS_3.

Например, ‘экономическое преступление’:

икътисадый (ADJ) *җинаять* (N),
икътисад (N) *җинаяте* (N, POSS_3).

Наличие подобных регулярных структурных соответствий приводит к умножению грамматических вариантов словосочетаний. Как показывает таблица 2, в настоящее время под влиянием русского языка все большее распространение получают составные номинации и словосочетания, образованные путем сложения относительных и прилагательных.

Метаданные корпусов позволяют установить терминологические предпочтения того или иного средства массовой информации и во многих случаях определить если не источник возникновения термина, то основные очаги его распространения (сайты государственных структур, тексты законодательных актов, СМИ и пр.). Таблица 3 ил-

люстрирует степень распространенности синонимических терминов, обозначающих номинацию *средства массовой информации*, и показывает, какие издания и информационные агентства используют данный вариант.

Таблица 2. **Количество двухкомпонентных номинаций, образованных по моделям N + N, POSS_3 и ADJ +N**

Словосочетание	Структурная модель словосочетания	Русский перевод	Частотность в ПК	Частотность в ТНК
Җәмгыять палатасы	N + N, POSS_3	Общественная палата	10	2
Иҗтимагый палата	ADJ +N	Общественная палата	2 637	1 664
икътисад үсеше	ADJ +N	экономическое развитие	489	128
икътисади үсеш	N + N, POSS_3	экономическое развитие	5 663	3 242

Таблица 3. **Частотность использования терминов, обозначающих 'средства массовой информации', в коллекции Письменного корпуса татарского языка**

Номинация	Употребление в ПК	Основные источники
массакүләм мәгълүмат чаралары	5 050	Газета «Ватаным Татарстан» и мн. др.
гамәви мәгълүмат чаралары	2 530	Информационное агентство Татар-Информ и мн.др.
киңкүләм мәгълүмат чаралары,	550	Информационное агентство Татар-Информ, газеты «Безнең гәҗит», «Өмет»
күмәк мәгълүмат чаралары	111	Газета «Бөгелмә авазы»

Существование большого количества семантических дублетов на уровне отдельных лексем и многокомпонентных номинаций — одна из характерных черт лексической системы современного татарского языка.

Постоянный контакт с доминирующим русским языком является причиной непрекращающегося калькирования русской терминоло-

гии. Ситуация осложняется наличием огромного количества независимых очагов формирования терминологии, отсутствием реального координирующего центра и отсутствием общедоступных русско-татарских баз данных терминологии, содержащих все основные варианты перевода термина.

3. Заключение

Большое количество синонимических номинаций разного происхождения и структуры свидетельствует о том, что современный татарский язык находится в состоянии активного развития.

Существование независимых очагов разработки терминологии, различные предпочтения переводчиков, производящих отбор единиц (слов, словообразовательных моделей, моделей словосочетаний и т. п.) из богатого инвентаря слов и конструкций приводит к тому, что в языке используются различные номинации для одной и той же реалии.

На уровне монолексем наблюдается конкуренция синонимов различного происхождения: слов западноевропейского, арабо-персидского, русского и тюрко-татарского происхождения. На уровне многокомпонентных номинаций синонимия осложняется вариативностью не только компонентов словосочетаний, но и грамматических моделей построения словосочетаний.

Нам представляется, что наблюдаемое многообразие семантических дублетов является не столько доказательством богатства лексической системы татарского языка, сколько проявлением сложных социолингвистических процессов и неустойчивости норм современного татарского словоупотребления.

Материалы данного исследования являются важной составляющей для разрабатываемого в настоящее время Татарско-русского словаря сочетаемости общественно-политической лексики.

Литература

1. *Амиров Ф. К.* (1996), Русско-татарский юридический словарь. Казань.
2. Письменный корпус татарского языка. URL: http://corpus.tatar/, свободный (дата обращения: 31.03.17).
3. *Ганиев Ф. А.* (ред.) (1997), Русско-татарский общественно-политический словарь. Казань.
4. *Тагирова Ф. И., Тимерханов А. А.* (ред.) (2014), Русско-татарский словарь актуальной лексики. Казань.

5. Татарский национальный корпус «Туган тел».. URL: http://corpus.antat.ru/ свободный (дата обращения: 30.03.17).

References

1. *Amirov F.K.* (1996), Russko-tatarskiy yuridicheskiy slovar. [Russian-Tatar legal dictionary]. Kazan.
2. Corpus of Written Tatar. Available at: http://corpus.tatar/.
3. *Ganiyev F.A.* (ed.) (1997), Russko-tatarskiy obshchestvenno-politicheskiy slovar. Kazan.
4. *Tagirova F.I.. Timerkhanov A.A.* (ed.) (2014), Russko-tatarskiy slovar aktualnoy leksiki [Russian-Tatar dictionary of actual vocabulary]. Kazan.
5. Tugan Tel» Tatar National Corpus. Available at: http://corpus.antat.ru/

Галиева Альфия Макаримовна
Galieva Alfiya
E-mail: amgalieva@gmail.com

Невзорова Ольга Авенировна
Nevzorova Olga
E-mail: onevzoro@gmail.com

НИИ "Прикладная семиотика" Академии наук Республики Татарстан (Россия)
Research Institute of Applied semiotics of Tatarstan Academy of Sciences (Russia)

А. О. Гребенников
A. O. Grebennikov

К ВОПРОСУ ОБ АППРОКСИМАЦИИ ЗАВИСИМОСТИ ОБЪЕМА СЛОВАРЯ ОТ ОБЪЕМА ВЫБОРКИ

APPROXIMATION OF THE SAMPLE SIZE — VOCABULARY SIZE DEPENDENCE

Аннотация. Сравниваются результаты использования функции Вейбулла и функции Хаустайна для аппроксимации зависимости объема словаря от объема выборки на материале частотного словаря А. П. Чехова при различных способах организации используемого материала. Продемонстрировано предпочтительное использование функции Хаустайна. Также подтверждается целесообразностью хронологической организации текстов при анализе подобного рода.
Ключевые слова. Писательская лексикография, статистическое моделирование, стилеметрия.

Abstract. The use of Weibull and Haustein functions for the approximation of the dependence between sample size and resulting vocabulary size is analyzed. Frequency dictionary of the short stories by A. P. Chekhov was chosen as the material for the experiment. Haustein function is proved to be the preferable one, with the approximation results being far more precise. Chorological order of text processing is also justified for the analysis along similar lines.
Keywords. Authors' lexicography, statistical modeling, stylometry.

Выбор аппроксимирующей функции является одним из основных вопросов в стилеметрических исследованиях, в частности при моделировании зависимости объема словаря от объема выборки. Традиционно, прогностические результаты об объеме генеральной совокупности за пределами реального диапазона наблюдений представляются труднопроверяемыми из-за большого размера исходной генеральной совокупности. Проанализируем поведение различных аппроксимирующих функций, используя базу частотного словаря рассказов А. П. Чехова (150 рассказов, общий объем выборки — около 200 000 словоупотреблений), созданного на кафедре математической лингвистики СПбГУ в рамках «Автоматической антологии русского рассказа XIX — XX веков» [Гребенников 1999].

Для аппроксимации зависимости объема словаря от объема выборки были выбраны:

- функция Вейбулла:

$$y = N_{\max} - N_{\max} e^{-cx^d}, \qquad (1);$$

- и функция Хаустайна:

$$y = \frac{N_{\max} x^{\gamma}}{x^{\gamma} + q}, \quad (2)$$

где N — объем словаря, x — объем выборки, N_{max} — асимптотический объем словаря, c, d; γ, q — параметры распределения [Косарева и др. 2015; Haustein 1970].

На первом этапе рассказы последовательно объединялись порциями по 10 и для этих объединений составлялись ранговые частотные словари лексем. Полученные результаты нарастания объема словаря затем были аппроксимированы по авторизированной методике исследования, разработанной Г. Я. Мартыненко с использованием метода наименьших квадратов [Косарева и др. 2015; Гребенников1998]. Полученные результаты приводятся в таблице 1.

Таблица 1. **Результаты аппроксимации при объединении текстов в хронологическом порядке**

Кол-во рассказов	Словоупотр.	Лексемы	Апп. по Вейбуллу (N_{max} = 20 929)	Апп. по Хаустайну (N_{max} = 29 884)
10	5951	2016	2008	1985
20	11 725	3224	3245	3246
30	15 162	3907	3874	3886
40	20 352	4749	4722	4746
50	25 413	5477	5463	5492
60	33 499	6500	6513	6544
70	42 850	7400	7572	7597
80	50 321	8284	8326	8341
90	61 946	9385	9369	9368
100	74 506	10 335	10 353	10 335
110	87 079	11 350	11 219	11 186
120	105 019	12 269	12 290	12 244
130	123 046	13 302	13 211	13 161
140	161 074	14 774	14 768	14 748
150	198 246	15 889	15 925	15 976

Также возможно рассмотреть прогностические величины, полученные при помощи исследуемых формул, за пределами реального диапазона наблюдений (таблица 2).

Таблица 2. **Прогностические результаты об объеме словаря за пределами диапазона наблюдений при объединении текстов в хронологическом порядке**

Словоупотр.	Апп. по Вейбуллу	Апп. по Хауштайну
250 000	17 126	17 330
300 000	17 974	18 368
400 000	19 093	19 933
500 000	19 755	21 069
750 000	20 511	22 924
1 000 000	20 768	24 062

Из таблиц видно, что на эмпирическом интервале теоретические кривые весьма близки друг к другу, вплоть до объема 200–250 тыс. с/у. Расхождения с постепенным увеличением разрыва начинаются только после 500 тыс. с/у.

Сопоставим полученные нами данные с полученными ранее значениями, касающиеся полного объема словаря А. П. Чехова (лаборатория А. А. Поликарпова). Объем корпуса составляет 1 381 000 с/у, которому после лемматизации соответствует 36 153 лексем [Кукушкина и др. 2012]. Если из объема корпуса вычесть пьесы и повести, и сделать поправку на обрабатываемую нами выборку, то, скорее всего, и получился бы объем примерно соответствующий 30 тыс. лексем, что и прогнозируется с использованием функции Хауштайна.

Одновременно, можно заметить, что выборка объемом около 200 000 с/у представляется вполне достаточной для получения прогностических величин, максимально приближающихся к реально существующим при использовании функции Хауштайна.

Кроме того, представляется любопытным проверить влияние хронологической составляющей на результаты аппроксимации. Для повторного эксперимента рассказы объединялись уже в случайном порядке (аналогом которого выступил алфавитный по именам файлов, соответствующих названиям рассказов с некоторыми изменениями, записанным латиницей). При таком порядке объединения ряд зрелых произведений Чехова значительного объема (например, «Архиерей»)

оказались в первых кумулятивных группах. Полученные результаты приводятся в таблицах 3 и 4.

Таблица 3. **Результаты аппроксимации при объединении текстов в случайном порядке**

Кол-во рассказов	Словоупотр.	Лексемы	Апп. по Вейбуллу (N_{max} = 56 029)	Апп. по Хаустайну (N_{max} = 104 094)
10	12 256	3203	3130	3132
20	39 266	6203	6352	6351
30	54 129	7600	7684	7682
40	72 149	8939	9089	9085
50	84 571	9830	9962	9958
60	96 778	10 650	10 760	10 756
70	108 266	11 538	11 467	11 462
80	119 347	12 204	12 113	12 109
90	131 255	12 933	12 773	12 770
100	136 316	13 287	13 044	13 042
110	145 529	13 800	13 524	13 523
120	162 480	14 388	14 366	14 368
130	178 073	14 999	15 098	15 104
140	186 789	15 455	15 492	15 500
150	198 246	15 889	15 994	16 006

Таблица 4. **Прогностические результаты об объеме словаря за пределами диапазона наблюдений при объединении текстов в случайном порядке**

Словоупотр.	Апп. по Вейбуллу	Апп. по Хаустайну
250 000	18 071	18 104
300 000	19 842	19 902
400 000	22 874	23 009
500 000	25 418	25 649
750 000	30 392	30 943
1 000 000	34 106	35 055

Мы видим, в целом, аналогичный характер изменения данных, полученных с помощью анализируемых функций, однако асимптотические значения представляются мало реальными даже в случае использования функции Вейбулла (около 56 000 лексем), и уже совсем невероятными с использованием функции Хауштайна (около 104 000 лексем). Видимо, это обусловлено тем, что язык автора — это сложный целостный организм, произвольное нарушение законов развития которого приводит к искажению реальных тенденций.

Таким образом, представляется возможным сделать следующие выводы:

1) функция Хауштайна, в целом не отличающаяся значительно от функции Вейбулла на реально анализируемом материале, показывает гораздо более точные значения при прогнозировании, что свидетельствует о ее большей адекватности исследуемому материалу;
2) при анализе зависимости объема словаря от объема выборки следует предпочесть хронологический порядок объединения текстов;
3) подтверждается работоспособность объема выборки объемом около 200 000 словоупотреблений в стилеметрических исследованиях.

Дальнейший содержательный и сопоставительный анализ с привлечением расширенного материала (других словарей из «Автоматической антологии русского рассказа XIX — XX веков» [Гребенников 2015]) представляется перспективным.

Литература

1. *Гребенников А. О.* (1998), Исследование устойчивости лексико-статистических характеристик текста. Дис. … канд. филол. наук. СПб.
2. *Гребенников А. О.* (2015), О стилеразличительных возможностях частотных словарей языка писателя // Русский язык и литература в пространстве мировой культуры: Материалы XIII Конгресса МАПРЯЛ (г. Гранада, Испания, 13–20 сентября 2015 года), т. 7, с. 93–97.
3. *Гребенников А. О.* под ред. Мартыненко Г. Я. (1999), Частотный словарь рассказов А. П. Чехова. СПб.
4. *Косарева Е. О., Мартыненко Г. Я.* (2015), Отношение текст — словарь в повседневной устной речи // Структурная и прикладная лингвистика: межвуз. сб., вып. 11. СПб., с. 220–228.
5. *Кукушкина О. В., Рюдигер Д. Ю., Суровцева Е. В., Лапонина Л. В* под ред. проф.

Поликарпова А. А (2012), Частотный грамматико-семантический словарь языка художественных произведений А. П. Чехова (с электронным приложением) М.
6. *Haustein H.-D.* (1970), Prognoseferfahren in den sozialistischen Wiertschaft. Berlin.

References

1. *Grebennikov A. O.* (1998), Issledovanie ustoychivosti leksiko-statisticheskikh kharakteristik teksta [Validity of the Statistics for Fiction]. Dis. kand. filol. nauk [PhD (Linguistcs) Thesis]. Saint Petersburg.
2. *Grebennikov A. O.* (2015), O stilerazlichitelnykh vozmozhnostyakh chastotnykh slovarey yazyka pisatelya [Author's lexicon frequency dictionary and style distinguishing]. In: Russkiy yazyk i literatura v prostranstve mirovoy kultury: Materialy XIII Kongressa MAPRYaL (g. Granada, Ispaniya, 13–20 sentyabrya 2015 goda) [The Russian Language and Literature in the Context of World Culture. Proceedings of XIII MAPRYaL Congress, Granada, Spain, 2015], vol. 7, pp. 93–97.
3. *Grebennikov A. O. edited by Martynenko G. Ya.* (1999), Chastotnyy slovar rasskazov A. P. Chekhova [Frequency Dictionary of the Short Stories by A. P. Chekhov]. Saint Petersburg.
4. *Haustein H.-D.* (1970), Prognoseferfahren in den sozialistischen Wiertschaft. Berlin
5. *Kosareva E. O., Martynenko G. Ya.* (2015), Otnoshenie tekst — slovar v povsednevnoy ustnoy rechi [The Type-Token Ratio in Everyday Spoken Russian]. In: Strukturnaya i prikladnaya lingvistika [Structural and Applied Linguistics], vol. 11. Saint Petersburg, pp. 220–228.
6. *Kukushkina O. V., Riudiger D. I., Surovtseva E. V., Laponina L. V edited by prof. Polikarpov A. A.* (2012), Chastotnyi grammatiko-semanticheski slovar jazyka hudojestvennyh proizvedenii A. P. Chekhova (s elektronnym prilojeniem) [The frequency grammar-semantic dictionary of the Chekhov's fiction]. Moscow.

Гребенников Александр Олегович
Санкт-Петербургский государственный университет (Россия)
Alexander Grebennikov
Saint Petersburg State University (Russia)
E-mail: a.grebennikov@spbu.ru

П.Л.Гроховский, А.В.Добров, А.Е.Доброва, Н.Л.Сомс
P.L.Grokhovsky, A.V.Dobrov, A.E.Dobrova, N.L.Soms

КОРПУС-МЕНЕДЖЕР ДЛЯ МОРФОСИНТАКСИЧЕСКОЙ РАЗМЕТКИ: ОПЫТ РАЗРАБОТКИ КОРПУСА ТИБЕТСКИХ ГРАММАТИЧЕСКИХ СОЧИНЕНИЙ[1]

CORPUS MANAGER FOR MORPHOSYNTACTIC ANNOTATION: EXPERIENCE OF DEVELOPMENT OF CORPUS OF INDIGINEOUS TIBETAN GRAMMAR TREATISES

Аннотация. В данной статье представлен опыт разработки корпус-менеджера для морфосинтаксической разметки на материале корпуса тибетских грамматических сочинений. Рассматриваются проблемы токенизации и вертикальной разметки тибетского текста, обусловленные особенностями синтактики тибетских алломорфов. Предлагается новый подход к организации разметки корпуса, не требующий разбиения текста на словоформы и основанный на синтаксической разметке. Описывается созданная технология отладки морфосинтаксической разметки, объединяющая корпус-менеджер, формальную грамматику и лингвистический процессор и позволяющая эффективно дорабатывать языковые модули лингвистического процессора так, чтобы формальная модель объясняла все, а не только некоторые явления в корпусе.
Ключевые слова. Корпусный менеджер; тибетский язык; морфосинтаксическая разметка; токенизация; лингвистический процессор.
Abstract. The article presents the experience of developing a corpus manager for morphosyntactic annotation on the basis of The Corpus of Indigenous Tibetan Grammar Treatises. The problems of tokenization and vertical markup of Tibetan texts are considered, which are conditioned by Tibetan allomorph syntactics features. A new approach to organization of the corpus annotation is proposed, which does not require segmentation of text into word forms and is based on syntactic annotation. The developed technology of debugging morphosyntactic markup is described, which integrates the corpus manager, the formal grammar and the natural language processor and allows to effectively refine linguistic modules of the natural language processor so that the formal model can explain not just some, but all the phenomena in the corpus.
Keywords. corpus manager; Tibetan language; morphosyntactic annotation; tokenization; natural language processor.

В рамках проекта РФФИ «Морфосинтаксический анализатор текстов на тибетском языке» был создан корпусный менеджер (далее — КМ) для ранее подготовленного корпуса тибетских грамматических сочинений. Токенизация и морфологическая разметка были первоначально выполнены вручную: искусственное разделение тибетских текстов на токены характеризовалось некоторой произвольностью,

[1] Исследование выполнено в рамках научно-исследовательского проекта РФФИ «Морфосинтаксический анализатор текстов на тибетском языке» (16-06-00578 А).

так как графематическое или какое-либо иное деление на словоформы в тибетском языке отсутствует; поэтому было принято решение ограничиться инвентарём более чётко определяемых атомарных единиц морфосинтаксической структуры: алломорфов, знаков пунктуации, разделителей, цифр.

Синтактика алломорфов в тибетском языке переплетена с синтаксисом предложения, поэтому формальная грамматика должна моделировать все уровни грамматической системы от алломорфов до высказываний и текстов.

Существующие КМ ориентированы на языки иного строя и работают с токенизацией и морфологической разметкой, поэтому было принято решение разработать иной КМ, который бы позволил: 1) работать с синтаксической (морфосинтаксической) разметкой и 2) находить в ней места, требующие усовершенствования лингвистического процессора, её порождающего.

КМ позволяет загружать неразмеченные тексты или тексты в "вертикальном" формате для их дальнейшей автоматической разметки, отражающей структуры непосредственных составляющих и зависимостей, при этом единицы, ранее считавшиеся токенами, разбиваются на алломорфы, которые далее объединяются в древовидные структуры.

Поиск, организованный в КМ, позволяет находить морфосинтаксические структуры по заданным моделям. На данный момент доступен поиск по тибетским моделям слово- и формообразования, однако может быть реализован поиск по синтаксическим структурам любой сложности (в разметке других корпусов в КМ представлены структуры предложений и текстов, и расширение границ разметки тибетского корпуса планируется в будущем). Поскольку поиск осуществляется не по словам, а по морфосинтаксическим деревьям, результатом поиска являются фрагменты синтаксических структур (морфосинтаксические деревья с грамматическими характеристиками и морфемным наполнением).

КМ включает в себя поддержку для корпусов на различных языках документов корпусов и включает ряд инструментов для автоматической разметки корпуса и обнаружения фрагментов этой разметки, требующих усовершенствований лингвистического обеспечения («ошибок» разметки).

КМ предоставляет возможность просмотра разметки полностью размеченных фрагментов текста. Для частично размеченных фрагментов отображается три дополнительных вида помет: нераспознанные

единицы, разрывы и перекрытия синтаксических деревьев. Нераспознанными считаются фрагменты, для которых в разметке отсутствуют синтаксические деревья; разрывами — позиции, в которых дерево не может быть связано с соседним; перекрытиями — фрагменты текста, в которых пересекаются синтаксические деревья, не полностью покрывающие текст: фрагмент, покрытый одним деревом, включает позицию начала фрагмента, покрытого вторым деревом, но не позицию его конца.

Данный инструментарий позволяет одновременно работать над разметкой корпуса и совершенствовать формальную модель, стоящую за используемым лингвистическим обеспечением, что представляет собой новый подход к разработке модулей лингвистического процессора, обеспечивающий постоянную верифицируемость формальной модели и её соответствие корпусному материалу. Последовательно устраняя нераспознанные фрагменты, разрывы в разметке, перекрытия и комбинаторные взрывы путём усовершенствования лингвистического обеспечения, разработчик в итоге добивается не только полной разметки корпуса, но и такого состояния формальной модели, при котором она объясняет все наблюдаемые в корпусе явления.

При создании формальных грамматик изначально учитываются, как правило, лишь наиболее типичные и понятные разработчику явления языковой грамматики, однако при работе с корпусом текстов обнаруживается множество неочевидных, но частотных конструкций, не учтенных при создании модели. К числу таких явлений в тибетском корпусе относилось употребление различных сочетаний с цифрами и числительными, имён и именованных сущностей, окказионализмов и специфических конструкций, в том числе — металингвистические употребления экспонентов языковых единиц (например, *суффикс -pa-/-ba-*). Работа с данным корпусом осуществлялась именно методом последовательного устранения недостатков разметки путем последовательного пополнения и исправления словарей и формальной грамматики; при этом было устранено около 700 разрывов, 100 перекрытий и 200 комбинаторных взрывов.

При разработке КМ использовалась гибкая архитектура, включающая в себя СУБД PostgreSQL, серверное приложение (скрипты на языке Python, подключённые к веб-серверу Apache через mod_wsgi) для обработки JSON-RPC запросов и статический веб-интерфейс с адаптивной верстkой на основе технологии Bootstrap: КМ доступен как в десктопных, так и в мобильных браузерах. Лингвистический про-

цессор также подключён к КМ: серверный компонент КМ отправляет запросы gsoap-серверу AIIRE на обработку и автоматическую разметку текстов; разметка передаётся в формате XML. КМ обеспечивает сохранение разметки в СУБД; на стороне СУБД средствами хранимых процедур производится парсинг XML и сохранение разметки в таблицы непосредственных составляющих и грамматических признаков. Дальнейшая обработка разметки производится средствами серверного компонента КМ: производится поиск и сохранение в СУБД нераспознанных фрагментов, разрывов, перекрытий и комбинаторных взрывов. Кроме того, серверный компонент КМ обладает API для выдачи компонентов разметки в виде JSON-объектов, отражающих НС-структуры с грамматической информацией, информацией о зависимостях и морфемном наполнении. Отрисовка древовидных структур реализована на клиентской стороне на языке Javascript: НС-структуры изображаются в виде SVG-файлов. При существенном объеме данных этот процесс может занимать длительное время. Исходный код КМ, как и исходный код лингвистического процессора, является открытым и доступен по адресу: http://svn.aiire.org/repos/tproc/trunk/t/corpus_manager. КМ допускает вертикальное и горизонтальное масштабирование стандартными средствами Apache и PostgreSQL. В КМ реализован механизм аутентификации и распределения прав доступа; гостевой доступ обеспечивает возможность просмотра опубликованных корпусов и их разметки, а также поиска по ним; доступ разработчика и административный доступ позволяют работать с разметкой и инструментарием её отладки.

В ходе морфосинтаксической разметки корпуса 86 192 единицы размечены как атомарные, что почти в 2 раза превосходит количество единиц, ранее считавшихся токенами (48 166). При этом 44 837 из них получили автоматическую разметку в виде морфосинтаксических деревьев, полностью покрывающих исходные токены. Таким образом, суммарное покрытие автоматической разметки составило 97 %. КМ доступен по адресу http://corpora.spbu.ru/corman/.

Литература

1. *Beyer S.* (1992), The Classical Tibetan language. New York.
2. *Grokhovskii P. L., Zakharov V. P., Smirnova M. O., Khokhlova M. V.* (2015), The Corpus of Tibetan Grammatical Works // Automatic Documentation and Mathematical Linguistics. Vol. 49, no. 5, pp. 182–191.
3. *Haspelmath M.* (2011), The indeterminacy of word segmentation and the nature of

morphology and syntax // Folia Linguistica. Vol. 45, iss. 1, pp. 31–80.
4. *Захаров В. П.* (2016), Корпусная лингвистика // Прикладная и компьютерная лингвистика / под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. М.: УРСС. 320 с.; с. 138–155.

References

1. *Beyer S.* (1992), The Classical Tibetan language. New York.
2. *Grokhovskii P. L., Zakharov V. P., Smirnova M. O., Khokhlova M. V.* (2015), The Corpus of Tibetan Grammatical Works. In: Automatic Documentation and Mathematical Linguistics, vol. 49, no. 5, pp. 182–191.
3. *Haspelmath M.* (2011), The indeterminacy of word segmentation and the nature of morphology and syntax. In: Folia Linguistica, vol. 45, iss. 1, pp. 31–80.
4. *Zakharov V. P.* (2016), Corpus Linguistics. In: Applied and Computer Linguistics. Eds. I. S. Nikolaev, O. V. Mitrenina, T. M. Lando. Moscow, URSS, 320 p.; pp. 138–155.

Гроховский Павел Леонович
Grokhovskiy Pavel
E-mail: p.grokhovskiy@spbu.ru

Добров Алексей Владимирович
Dobrov Aleksei
E-mail: a.dobrov@spbu.ru

Санкт-Петербургский государственный университет (Россия)
Saint Petersburg State University (Russia)

Доброва Анастасия Евгеньевна
Dobrova Anastasia
E-mail: adobrova@aiire.org

Сомс Николай Леонидович
Soms Nikolai
E-mail: nsoms@aiire.org

ООО "АИРЕ" (Россия)
AIIRE LLC (Russia)

П. Л. Гроховский, М. О. Михайлова
P. L. Grokhovsky, M. O. Mikhailova

АВТОМАТИЗАЦИЯ СЕГМЕНТАЦИИ И ЧАСТЕРЕЧНОЙ РАЗМЕТКИ ТИБЕТСКОГО ТЕКСТА[1]

AUTOMATIZATION OF SEGMENTATION AND POS-TAGGING OF TIBETAN TEXT

Аннотация. Данный доклад посвящён методам автоматизации сегментации и разметки, разработанным и использованным при создании корпуса текстов на современном тибетском языке. Автоматизация, основанная на сочетании использования словаря и правил, позволяет значительно сократить время обработки, сделать разметку более единообразной. Кроме того, в ходе работы были созданы словари и скрипт для сегментации тибетского текста, область возможного применения которых значительно шире академической разметки текста.
Ключевые слова. Сегментация, разметка, корпус, тибетский язык, автоматизация.

Abstact. The research presents methods implemented for partial automatization of segmentation and POS-tagging of Tibetan text. The system is based on combination of vocabulary and rules, and has significantly improved speed and uniformity of text processing, as well as contributed to creation of text-based dictionary and script for tokenisation of Tibetan text, that can find broader usage in Tibetan studies.
Keywords. Segmentation, annotation, corpus, Tibetan, automatization.

В ходе работы над разметкой двух корпусов тибетских текстов (грант ФФЛИ № С-47 «Базовый корпус тибетского классического языка с русским переводом и лексической базой данных» и грант РФФИ № 13–06–00621 А «Пилотная версия электронного корпуса тибетских грамматических сочинений») возникла необходимость частично автоматизировать дальнейшую разметку. Разработка программных средств для предварительной автоматической разметки и сегментации тибетского текста осуществляется в рамках гранта РГНФ 16–04–12016 «Программные средства автоматической обработки текста на современном тибетском языке (морфологический уровень)».

За стандарт разметки и сегментации принят формат корпуса грамматических сочинений. Это исключало использование разметки по словарю с последующим снятием неоднозначности по правилам, разработанной группой британских исследователей [Garret et al. 2014] в связи с различиями в принципах сегментации и наборе тегов. На-

[1] Работа выполнена при поддержке гранта РГНФ № 16-04-12016 «Программные средства автоматической обработки текста на современном тибетском языке (морфологический уровень)».

пример, анализ данных имеющихся корпусов показывает, что частицы, выделяемые ими как глагольные [Garrett 2015: 70], могут следовать и за токенами других типов. Поэтому на основании реальных данных было принято решение выделять такие частицы в отдельный класс, а не относить к глагольным аффиксам. Другое отличие — решение объединить в один токен глагольную основу и формообразующие аффиксы глагола, поскольку между ними ничего невозможно вставить, и обособить аффиксы падежей существительных в отдельные токены, поскольку падежный аффикс присоединяется справа к именной группе и может быть отделен от существительного определениями.

Изначально предполагалась только предварительная автоматическая разметка заранее сегментированного текста, поскольку автоматическая сегментация тибетского текста мало разработана. Инструменты для сегментации текстов на языках на базе латинской письменности, использующей пробел в качестве разделителя слов (например, nltk), не подходят для языка, не содержащего никакого графического разделителя слов, даже при условии разработки тибетского языкового пакета для них, что теоретически возможно, хотя очень трудоёмко. Для китайского языка, похожего в этом отношении на тибетский, есть статистическое решение, основанное на рейтингах встречаемости конкретных слогов в начале, середине и конце слова [Nianwen Xue, 2002]. Китайские [Huidan Liua, 2011] и британские [Garret et al, 2015] исследователи провели эксперименты по применению этой системы к тибетскому тексту, показавшие высокую точность на тестовых данных, но открытой для использования системы нет, а её тренировка с нуля требует большой и специфически размеченной базы текстов.

На материале уже размеченных текстов использовавшихся корпусов был автоматически составлен словарь, содержащий для каждого токена следующую информацию: написание токена по-тибетски, написание токена в транслитерации, лемма по-тибетски, лемма в транслитерации, тег (все варианты разметки данного токена, присутствующие в корпусе в случае омонимии), количество вхождений в текстах, список текстов, в котором данный токен встретился хотя бы один раз, например:

འཕགས་ ‘phags འཕག ~ འཕགས་ ‘phag ~ ‘phags V ~ N 6 текст 1, текст 7, текст 2, текст 3, текст 5

Большинство токенов может встретиться в текстах как с графическим разделителем (tsheg) на конце, так и без него, поэтому было

принято решение вносить в словарь токены без этого разделителя. Данный словарь возможно автоматически пополнять при добавлении новых размеченных текстов, он позволяет сортировать и искать данные по любому столбцу — например, увидеть подряд все токены одной части речи или все токены с одной леммой. За пределами данного проекта этот словарь может использоваться как словарь частот, лемм и частей речи, а также для решения комплексных задач, например, для поиска омонимов разных частей речи (совпадающее написание глагола и существительного) можно делать поиск по паре лемма+тег. Общий объём корпусов, по которым составлен словарь, 91825 токенов, в том числе 9112 уникальных токенов и 6111 лемм. Второй словарь, который используется для токенизации и разметки — словарь форм глаголов Нейтена Хилла [Hill 2010], объединяющий данные нескольких ранее созданных глагольных словарей.

Программные средства написаны на Python 2.7. Скрипт присуждает каждому токену лемму (леммы) и тег (теги) и записывает удобный для редактирования и проверки файл .csv, где каждая строка содержит один токен и его атрибуты. Для упрощения проверки разметки написаны инструменты для поиска, в том числе по любому атрибуту или их сочетанию и поиск контекста. Для сочетаемости с корпус-менеджерами написаны конвертеры табличного файла в вертикальный формат, xml или формат для nltk)

Так как частичная автоматизация обработки предполагает последующую ручную проверку, программа для разметки должна была выдавать все варианты разметки, найденные для данного токена в доступном материале, и оставлять неразмеченными все новые токены — при разметке только по изначальному словарю они составляли около 16 %. После успешного проведения эксперимента по автоматической сегментации основным режимом работы стало одновременное выполнение сегментации и разметки неподготовленного текста, так как это значительно сокращает работу лингвиста.

Сегментация построена следующим образом. Полностью обрабатывалась каждая строка входящего текстового файла по отдельности. Она разбивалась на графические слоги по разделителю слогов tsheg. Если после этого строка была слишком длинной (более 20 слогов) — использовался генератор подстрок, который разбивал длинную строку на фрагменты по знакам препинания (вертикальная черта shad и пробел) и передавал для разбора эти фрагменты по одному. Далее такой фрагмент разбивался на единицы, подходящие для поиска по словарю.

На этой стадии от слогов, содержащих только буквы тибетского алфавита, отделялись знаки препинания и цифры, обособлялись элементы, которые не отделяются от предыдущего слога графически, но являются отдельными морфемами: падежный аффикс *'i*, финитная частица *'o*, вопросительная частица *'am* и глагольный аффикс *'ang*. Затем каждая строка в виде списка таких единиц передавалась в функцию, которая осуществляла по словарю поиск последовательностей максимальной длины. Это важно для тибетского языка, где почти все буквы могут формировать отдельный слог и слово.

После окончания сегментации и разметки скрипт выводит на печать общее количество токенов, количество размеченных токенов, количество неразмеченных токенов и время, затраченное на обработку. При наличии уже сегментированного текста возможно также отдельно использовать функцию разметки, без одновременной сегментации.

Анализ лакун и ошибок сегментации и разметки выявил следующие проблемы, послужившие основными направлениями доработки системы.

Во-первых, недостаточная полнота словаря. Этот недостаток исправляется по мере пополнения словаря, но лишь частично: с добавлением в словарь всего материала из корпуса текстов общего содержания, то есть, увеличении материала почти в три раза относительно изначального объёма, количество неразмеченных токенов сократилось всего на 5 %, до 11 %.

Во-вторых, многие ошибки относятся к открытым классам словоформ с очень большим варьированием (числительные, цифры, формы глаголов), образование которых которые можно описать правилами. Введение системы таких правил сократило количество неопознанных токенов до 7 %.

Эти правила используются на стадии поиска возможных комбинаций слогов в словаре. Использовались следующие правила:

- Выделение форм глаголов: скрипт объединял глагольную основу (только основы, найденные в словаре с тегом "V") с глагольными аффиксами из списка, и присуждал такой словоформе тег в зависимости от аффикса, что позволяет опознать и разметить даже те формы глаголов, которые не встретились в существующих корпусах и поэтому отсутствуют в словаре. Словарь аффиксов и тегов составлен по грамматическому корпусу.
- Выделение чисел как непрерывных последовательностей латинских или тибетских цифр.

- Выделение количественных числительных как единиц, состоящих только из слогов, которые образуют числительные, по списку возможных составляющих. Этот список взят из монографии [Beyer 1992: 221–226] и включает в себя как числа, так и специфические для тибетского языка разделители разрядов.
- Выделение порядковых числительных как токенов, которые на конце имеют тибетский аффикс -ра, а вся предшествующая часть которых является количественным числительным.
- Выделение сокращённых форм падежей: для двух падежей (терминатива и эргатива) после гласных используются сокращённые формы. Их особенность в том, что они сокращаются до одной буквы (‘r’ и ‘s’ соответственно) и встраиваются в предшествующий графический слог. Отделить в самостоятельные токены все ‘r’ и ‘s’ нельзя, так как таких случаев гораздо меньше, чем случаев, когда эти согласные являются финалью токена, поэтому при поиске последовательностей, оканчивающихся на эти согласные, также проверяется наличие в словаре этой последовательности без последней буквы. Если она есть, то выделяется два токена: найденная в словаре последовательность и падежный аффикс.

Система была протестирована на корпусе из 8 тибетских текстов научного стиля общим объемом 55 533 токена, 52 539 (94 %) токенов было размечено и 2994 (6 %) — нет. Точность (precision) и полнота (recall) совпадают в пределах округления. При тестировании только разметки на эталонной сегментации процент неразмеченных токенов составил 6,2 %, строгая точность 45 % (только один вариант разметки и он верный), нестрогая 92,3 % (наличие нескольких вариантов разметки, в том числе правильного), неправильно размечено 1,5 % токенов. Некоторая часть неразмеченных фрагментов оказалась результатами ошибок в текстах. При тестировании на тех же текстах после очистки от таких ошибок, количество неразмеченных токенов сократилось на 0,7 % (до 5,5 %), это примерно 240 токенов на весь корпус. При одновременном выполнении сегментации и разметки F-мера разметки при строгой оценке составляет 46 %, при нестрогой — 85 %, 7 % токенов не размечено, F-мера сегментации составила 88 %.

Что касается скорости разметки, то сегментация и разметка самого большого текста, который был сегментирован в ходе работы над этим проектом — 20 тысяч токенов в автоматически размеченном варианте — обрабатывался примерно две секунды. Работа только в режиме

разметки (при уже сделанной человеком сегментации) требует меньше времени: сама разметка занимает 0.1 секунды для того же текста.

References

1. *Beyer S.* (1992), The Classical Tibetan Language. New York.
2. *Garrett E., Hill N. W., Zadoks A.* (2014), A Rule-based Part-of-speech Tagger for Classical Tibetan. In: Hymalayan Listics 13(2), permalink: http://escholarship.org/uc/item/5jv3r0rn
3. *Garrett E., Hill N. Wngui., Kilgarriff A., Ravikiran V., Zadoks A.* (2015), The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries. In: Revue d'Etudes Tibétaines, 32. pp. 51–86.
4. *Garrett E., Hill N.* (2015), A Constraint Grammar POS-Tagger for Tibetan. In: Proceedings of the Workshop on "Constraint Grammar — methods, tools and applications" at NODALIDA, Vilnius, pp. 19–22.
5. *Hill N. W.* (2010), A lexicon of Tibetan verb stems as reported by the grammatical tradition, Munich.
6. *Huidan Liua, Minghua Nuo, Longlong Ma, Jian Wua, Yeping He* (2011), Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Field. In: 25th Pacific Asia Conference on Language, Information and Computation, Beijing, pp. 168–177.
7. *Nianwen Xue* (2002), Combining Classifiers for Chinese Word Segmentation. In: Proceedings of the first SIGHAN workshop on Chinese language processing, Stroudsburg, vol. 18, pp. 1–7.

Гроховский Павел Леонович
Grokhovskiy Pavel
E-mail: p.grokhovskiy@spbu.ru

Михайлова Мария Олеговна
Mikhailova Maria
E-mail: ariyatan@gmail.com

Санкт-Петербургский государственный университет (Россия)
Saint Petersburg State University (Russia)

М. Дебренн, А. М. Лаврентьев
M. Debrenne, A. Lavrentiev

КОРПУС РУССКИХ ФРАНКОЯЗЫЧНЫХ ДНЕВНИКОВ XIX В. КАК МАТЕРИАЛ ДЛЯ ИССЛЕДОВАНИЯ ВЗАИМОДЕЙСТВИЯ ЯЗЫКОВ И КУЛЬТУР[1]

A CORPUS OF 19TH CENTURY RUSSIAN FRANCOPHONE DIARIES AS DATA FOR RESEARCH ON INTERACTION OF LANGUAGES AND CULTURES

Аннотация. В докладе представлена работа по созданию корпуса франкоязычных дневников русских авторов XIX в. Рассмотрены проблемы адекватной исследовательским задачам разметки и предложены технические решения, позволившие оптимизировать работу коллектива проекта. Эксплуатация корпуса осуществляется с помощью платформы TXM – локальной версии и веб-портала.
Ключевые слова. Билингвизм, дневник, французский язык, разметка TEI, платформа TXM.

Abstract. This paper presents the work on a corpus of francophone diaries written by Russian authors in the 19th century. The problems of the markup necessary for research purposes are considered and technical solutions that allow optimizing the work of the project team are proposed. The local version and the web portal of the TXM platform can be used to work with the corpus.
Keywords. Bilingualism, diary, French language, TEI markup, TXM platform.

Франкоязычные дневники русской аристократии XIX в. предоставляют интереснейший материал для изучения билингвизма и взаимодействия русской и французской культуры. Создание корпуса таких дневников, снабженного адекватной исследовательским целям разметкой, может способствовать получению качественно новых научных результатов и обеспечить доступ к данным материалам широкому кругу представителей различных гуманитарных наук.

Такую цель ставит перед собой проект «Взаимодействие культур в пространстве русского франкоязычного дневника XIX века», проводимый коллективом исследователей Новосибирского государственного университета и лаборатории IHRIM Национального центра научных исследований Франции. Представленная в докладе работа сосредоточена на материале дневниковых тетрадей О. И. Орловой-Давыдовой (1814–1876), хранящихся в ГПНТБ СО РАН и озаглавленных «Journal d'Olga Davidoff». В корпус вошли пять тетрадей, содержащих

[1] Работа проведена с финансовой поддержкой РГНФ проект (№ 16-24-08001) и французского Фонда домов гуманитарных наук (FMSH).

записи на французском языке, датированные 1835–1845 гг, с небольшими более поздними вставками 1847, 1849 и 1869 гг.

На первом этапе проекта были определены виды разметки, необходимые для эксплуатации корпуса. К ним относится выделение имен собственных (топонимов и антропонимов), русскоязычных вкраплений (на кириллице или в транслитерации), различных видов ошибок и исправлений в тексте рукописи. Несколько маркеров разметки могут накладываться на один фрагмент текста (топоним может быть написан кириллицей и содержать исправление). Основу структуры корпуса составляет дневниковая запись. Все записи датированы (одним днем или периодом), однако указанные даты нуждаются в нормализации, так как наблюдается смешение старого и нового стиля и отдельные ошибки (несоответствие даты и дня недели). Начало каждой страницы и ее номер также маркируются, чтобы обеспечить возможность синоптического выведения на экран транскрипции и фотографии страницы рукописи.

Помимо транскрипции проект предполагает подготовку перевода дневника на русский язык. Разметка перевода сводится к минимуму, необходимому для постраничного выравнивания.

Для удобства участников проекта и с учетом технических возможностей был предложен рабочий процесс (workflow), включающий предварительную разметку в Microsoft Word с использованием технологии стилей и специальных сочетаний символов с последующим автоматическим преобразованием в формат XML с разметкой, соответствующей стандарту международной Инициативы по кодированию текстов (TEI)[2]. Трансформация осуществляется с помощью онлайн сервиса Odette[3] [Glorieux 2015] и последующего применения специально разработанной стилевой таблицы XSLT. Сервис Odette позволяет распознавать структуру документа (по стилям заголовков) и преобразовывать стили символов в стандартные теги TEI. Дополнительное преобразование необходимо для обработки сложных случаев (наложение нескольких видов разметки, нормализация дат и т. п.).

Файлы корпуса в формате XML-TEI импортируются в корпус-менеджер TXM [Heiden 2010] с помощью недавно разработанного модуля импорта XTZ. Этот модуль позволяет в автоматическом режиме применять серию трансформаций XSLT (до и после токенизации)

[2] http://www.tei-c.org

[3] http://obvil-dev.paris-sorbonne.fr/developpements/Odette

и создавать «синоптическое издание», включающее транскрипцию и фотографию источника. В процессе импортирования текст проходит автоматическую морфологическую разметку и лемматизацию с помощью программы TreeTagger[4], что существенно расширяет возможности качественного и количественного анализа.

На момент публикации сборника все пять тетрадей затранскрибированы и размечены в документе Word, тестовый корпус импортирован на платформу ТХМ. Он используется для проверки и корректировки транскрипции и предварительной разметки, а также для совершенствования автоматизации процедуры обновления корпуса и разработки сценариев его эксплуатации. Прототип издания корпуса размещен на демонстрационном портале ТХМ[5]. На основе размеченного текста дневников подготовлены два магистерских исследования, позволяющие уточнить использование имен собственных (антропонимов) в дневниках и особенности индивидуального идиолекта диаристов. Опубликован ряд статей с общим анализом особенностей французского языка О. И. Орловой-Давыдовой [Дебренн 2016a, Debrenne 2016b].

В дальнейшем корпус проекта будет пополняться за счет дневников других авторов. Первым из них является дневник поручика Александра Чичерина (1793–1813), работа над которым уже идет. Особенностью этого дневника является наличие большого числа авторских рисунков, тесно связанных с текстом. Анализ и аннотация этих рисунков и их отношений с фрагментами текста являются отдельной задачей, новой для текстометрических исследований.

Литература

1. *Дебренн М.* (2016a), Сопоставительный девиатологический анализ переписанных дневников О. И. Давыдовой и первичных текстов // Вестник НГУ, Серия Лингвистика и межкультурная коммуникация, 14 (3), с. 59–74.
2. *Debrenne M.* (2016b), The French Language in the Diaries of Olga Davydova. An example of Russian-French Aristocratic Bilingualism. In: M. van Strien-Chardonneau & M.-C. Kok (Ed.), Escalle Le français, langue de l'intime à l'époque moderne et contemporaine, Amsterdam : Amsterdam University Press B. V., pp. 125–142.
3. *Glorieux F.* (2015), Processing texts to produced structured data [Le traitement de textes pour produire des documents structurés (XML/TEI)]. URL: http://resultats.hypotheses.org/267.

[4] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

[5] http://portal.textometrie.org/demo/?command=documentation&path=/DAVYDOVA

4. *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 — PACLIC24, Sendai, pp. 389–398.

References

1. *Debrenne M.* (2016a), Sopostavitel'nyj deviatologicheskij analiz perepepisannykh dnevnikov O. I. Davydovoj i pervichnykh tekstov [Comparative error-based analysis of the copied Davydova's diary and its original]. In: Vestnik NGU, Serija Lingvistika i mezhkulturnaja kommunikacja [Vestnik SNU, Series Linguisitcs and intercultural communication], 14 (3), pp. 59–74.
2. *Debrenne M.* (2016b), The French Language in the Diaries of Olga Davydova. An example of Russian-French Aristocratic Bilingualism. In: M. van Strien-Chardonneau & M.-C. Kok (Ed.), Escalle Le français, langue de l'intime à l'époque moderne et contemporaine, Amsterdam: Amsterdam University Press B. V., pp. 125–142.
3. *Glorieux F.* (2015), Processing texts to produced structured data [Le traitement de textes pour produire des documents structurés (XML/TEI)]. Available at: http://resultats.hypotheses.org/267.
4. *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 — PACLIC24, Sendai, pp. 389–398.

Дебренн Мишель
Новосибирский государственный университет (Россия)
Debrenne Michèle
Novosibirsk State University (Russia)
E-mail: micheledebrenne@gmail.com

Лаврентьев Алексей Михайлович
Национальный центр научных исследований (Франция)
Lavrentiev Alexei
Centre national de la recherche scientifique (France)
E-mail: alexei.lavrentev@ens-lyon.fr

Н. Г. Зайцева, А. А. Крижановский, Н. Б. Крижановская, Н. А. Пеллинен, А. П. Родионова

N. G. Zaitseva, A. A. Krizhanovsky, N. B. Krizhanovsky, N. A. Pellinen, A. P. Rodionova

ОТКРЫТЫЙ КОРПУС ВЕПССКОГО И КАРЕЛЬСКОГО ЯЗЫКОВ (ВЕПКАР), ПРЕДВАРИТЕЛЬНЫЙ ОТБОР МАТЕРИАЛОВ И СЛОВАРНАЯ ЧАСТЬ СИСТЕМЫ[1]

OPEN CORPUS OF VEPS AND KARELIAN LANGUAGES (VEPKAR), PRELIMINARY SELECTION OF MATERIALS AND DICTIONARY OF THE SYSTEM

Аннотация. В статье описывается Открытый корпус вепсского и карельского языков (http://dictorpus.krc.karelia.ru/). Обсуждаются вопрос выбора текстовых материалов для наполнения корпуса и трудности, связанные с многообразием диалектов карельского языка. Рассмотрена структура словарной статьи, где словарь является частью системы корпуса текстов.
Ключевые слова. Вепсский язык, карельский язык, корпус, словарь.

Abstract. The project "Open corpus of Veps and Karelian languages" (http://dictorpus.krc.karelia.ru/) is described in this paper. Issues addressed in this paper include the selection of particular texts in constructing a corpus and difficulties associated with the rich diversity of dialects of Karelian language. The structure of the dictionary entry is presented, where the dictionary is a part of the corpus.
Keywords. Veps, Karelian language, corpus, dictionary.

1. Введение

По финно-угорским языкам известны лишь корпусы наиболее крупных из них: языковой банк Финляндии; фонетический корпус спонтанной эстонской речи (http://www.murre.ut.ee/phonetic-corpus/), венгерский национальный корпус (http://mnsz.nytud.hu/index_hun.html). Среди финно-угорских языков России имеется пилотная версия Корпуса удмуртского языка, представляющая лишь язык прессы и некоторое количество нехудожественных текстов.

Корпус вепсского языка (http://vepsian.krc.karelia.ru/) разрабатывался сотрудниками ИЯЛИ и ИПМИ КарНЦ РАН в 2009–2016 гг. Корпус и Словарь включают более тысячи текстов, более 800 библиографических источников, более 10 тысяч лемм и словоформ [Зайцева и др. 2015].

[1] Работа поддержана грантом Программы: Евразийское наследие и его современные смыслы. Направление 4. Мультимедийные технологии в филологических исследованиях (проект «Разработка морфологической базы и развитие корпуса вепсского языка»).

С 2016 года Корпус стал многоязычным: на базе компьютерной программы и базы данных Корпуса вепсского языка создан Корпус карельского языка. Объединенный корпус получил название: «Открытый корпус вепсского и карельского языков» (ВепКар). Корпус карельского языка включает себя три подкорпуса, деление осуществлено в соответствии с тремя основными наречиями (собственно карельского, ливвиковского, людиковского). Сайт корпуса ВепКар доступен по ссылке http://dictorpus.krc.karelia.ru/. Разработана табличная форма для указания морфологических признаков (падеж, число) для именных частей речи (см. пример для леммы astta (вепс.), aštuo, astua, astuda (кар.). Начата работа по семантической разметке текста: предложено автоматическое связывание слов текста со значениями лемм в словаре (рис. 1).

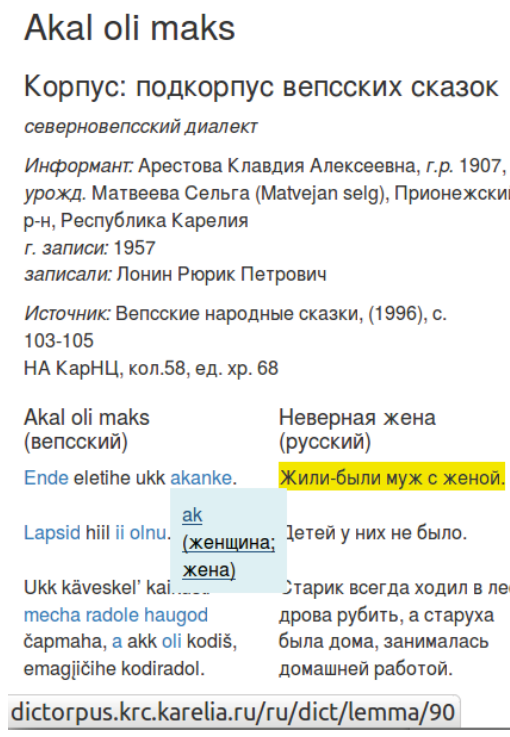

Рис. 1. **Пример текста из подкорпуса вепсских сказок, при наведении мышкой на строку текста подсвечивается соответствующая строка в переводе, при клике на слово в исходном тексте выпадает окошко с леммой, текстом значения и гиперссылкой на словарную статью, здесь словоформа *akanke*, лемма *ak* (вепс.)**

2. Отбор материалов по карельскому языку

Специалистам, работающим с материалами карельского корпуса, с одной стороны, значительно легче следовать опыту вепсского и подбирать текстовые материалы на примере предшественников. С другой стороны, при заполнении корпуса они сталкиваются с определённой спецификой, несколько отличающейся от работы с вепсским материалом. Это обусловлено, главным образом, наличием бóльшего в сравнении с вепсским числа диалектов и говоров, употребляемых в карельском языке. Как и в вепсском корпусе, перед исследователями стоит задача: какие говоры представить в корпусе и пример из какого говора взять за основу, так как в корпусе важнее и интереснее представить говоры с наиболее яркими и показательными фонетическими, морфологическими, синтаксическими и лексическими маркерами, что, помимо популяристических целей, будет иметь большое значение для лингвистических исследований. Остановиться на определённом говоре / группе говоров диктует необходимость связывать основную лемму со всеми текстовыми и грамматическими материалами корпуса, выбирая за основу тот или иной говор, исследователь помимо прочего анализирует его представленность и частотность в размещаемых материалах. Абсолютно все говоры, представленные в классификации П. М. Зайкова, отразить в корпусе невозможно, так как часть из них либо практически исчезла, либо находится на грани исчезновения. Речь идёт, прежде всего, о людиковских говорах, носителей которых осталось, по неофициальным данным, не более 300. Принято решение в качестве леммы указывать литературную форму слова, а в людиковском (поскольку он не обладает официально признанной литературной формой) за основу будет браться михайловский говор, так как на нём издана бóльшая часть людиковских текстов.

3. Словарная статья

При создании информационной системы встала задача определения структуры словарной статьи, где словарь является частью Корпуса. Есть ряд ограничений на словарь, а именно: многоязычность, необходимость сохранения словоформ разных диалектов одного языка, данные словарной статьи будут использоваться в дальнейшем для разметки текста Корпуса, фрагменты (предложения) текста Корпуса могут использоваться для иллюстрации разных значений слова.

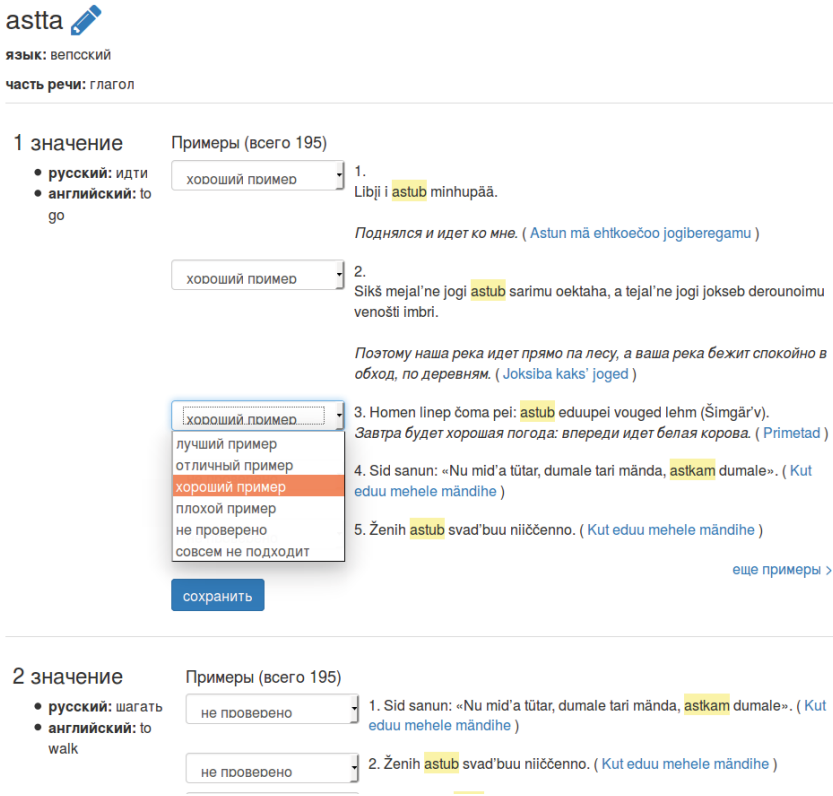

Рис. 2. **Пример леммы astta (вепс.) с двумя значениями, сопровождающимися переводом на русский и английский и примерами из корпуса текстов (http://dictorpus.krc.karelia.ru/ru/dict/lemma/56)**

На рис. 2 показан выбранный способ подачи материала в словарной статье. Вначале указана сама лемма, язык и часть речи. Далее материал разбит по значениям для многозначных слов. В каждом разделе, соответствующем одному значению, указаны:

1. Номер значения. Редактор может перенумеровать значения (рис. 3).
2. Толкование можно указать на любом языке, существующим в системе. Разница между толкованием и переводом здесь в том, что толкование — это произвольный текст, а перевод — это ссылка на конкретное значение.

3. Перевод на любой из языков в Корпусе. Сейчас в системе шесть языков (см. http://dictorpus.krc.karelia.ru/ru/dict/lang). В режиме редактирования леммы на рис. 3 при вводе перевода на английский язык для первого значения слова astta виден выпадающий список для выбора значения английских слов.

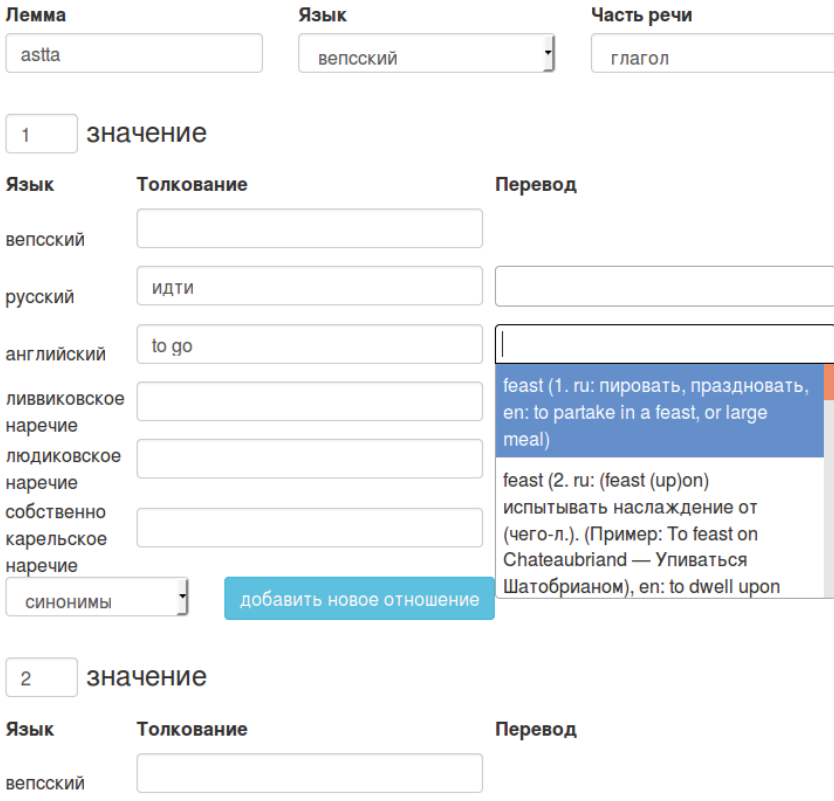

Рис. 3. **Поля словарной статьи на примере леммы astta (вепс.) в режиме редактирования**

4. Синонимы, антонимы и другие отношения для каждого значения даются отдельно.
5. Автоматически найденный список предложений из корпуса, содержащих словоформы леммы. Предложения сопровождаются ссылками на полные тексты в корпусе. В этом списке редактор

может вручную выбрать и указать предложения, наилучшим образом иллюстрирующие значение слова, может отметить ошибочно найденные предложения, чтобы они не выводились читателю. Вопрос остаётся открытым: как накопленная таким образом информация может быть использована для автоматической классификации предложений, разрешения лексической многозначности и вообще семантического поиска?

Литература

1. *Зайцева Н. Г., Филатова М. М., Шибанова Н. Л., Крижановский А. А.* Корпус вепсского языка // Корпусная лингвистика — 2015. Труды межд. конф. СПб.: С.-Петербургский гос. университет, Филологический факультет, 2015, с. 202–212. URL: http://scipeople.com/publication/121149/

References

1. *Zaitseva N. G., Filatova M. M., Shibanova N. L., Krizhanovsky A. A.* (2015), Korpus vepsskogo yazyka [Veps corpora]. In: International scientific conference "Corpus linguistics". Saint Petersburg, 2015, pp. 202–212.

Зайцева Нина Григорьевна
Zajceva Nina (*zng@ro.ru*)

Пеллинен Наталия Александровна
Pellinen Natalija (*nataliapellinen@gmail.com*)

Родионова Александра Павловна
Rodionova Aleksandra (*sashenka22@yandex.ru*)

Институт языка, литературы и истории Карельского научного центра РАН (Россия)
Institut jazyka, literatury i istorii Karel'skogo nauchnogo centra RAN (Russia)

Крижановский Андрей Анатольевич
Krizhanovskij Andrej (*andrew.krizhanovsky@gmail.com*)

Крижановская Наталья Борисовна
Krizhanovskaja Natal'ja (*nataly@krc.karelia.ru*)

Институт прикладных математических исследований Карельского научного центра РАН (Россия)
Institut prikladnyh matematicheskih issledovanij Karel'skogo nauchnogo centra RAN (Russia)

А. А. Зинина, А. А. Котов,
Н. А. Аринкин, Л. Я. Зайдельман

A. Zinina, A. Kotov,
N. Arinkin, L. Zaydelman

НАЛОЖЕНИЕ КОММУНИКАТИВНЫХ ФУКНЦИЙ: ИЗУЧЕНИЕ НА МУЛЬТИМОДАЛЬНОМ КОРПУСЕ «REC» И ПЕРЕНОС НА РОБОТА «Ф-2»

SUPERPOSITION OF COMMUNICATIVE FUNCTIONS: STUDIES ON "REC" MULTIMODAL CORPUS AND SIMULATION BY "F-2" ROBOT

Аннотация. Естественное поведение человека в коммуникации обладает богатством и сложностью: в частности, человек может одновременное выражать речью, мимикой и жестами различные эмоциональные состояния и коммуникативные намерения. На основе мультимодального корпуса REC мы описываем это явление как наложение коммуникативных функций. Анализ таких случаев и разработанное нами программное обеспечение позволяют имитировать наложение коммуникативных функций на роботе Ф-2.
Ключевые слова. Мультимодальная коммуникация, коммуникативные функции, робот компаньон.

Abstract. Natural human behavior in a communication is quite rich and compound, whereas a human can simultaneously express different emotional states and communicative intentions via speech, facial expression and gestures. Basing on REC multimodal corpus we describe this phenomenon as a superposition of several communicative functions. The analysis of these cases and the developed software allow us to simulate the superposition of functions on a robot F-2.
Keywords. Multimodal communication, communicative functions, robot companion.

Корпусные исследования представляют важное направление в современной лингвистике. Во многом это обусловлено самостоятельной теоретической значимостью таких работ, а также широким практическим применением, в частности — перспективой синтеза невербального коммуникативного поведения человекоподобным роботом.

Мы изучаем коммуникативное поведение человека на основе мультимодального корпуса REC (Russian Emotional Corpus), содержащего размеченные видеозаписи поведения человека в различных эмоциональных диалогах [Kotov, Budyanskaya 2012]. Для жеста или элемента мимики отмечается коммуникативная функция в соответствии с типологией, представленной в [Котов, Зинина 2015a; Котов, Зинина 2015b]. Анализ жестов, связанных с определённой коммуникативной функций, позволяет выделить жесты — типчные представители этой

КФ или инвариант поведения. Наиболее типичные для определённой КФ жесты зарисовываются и выполняются на роботе Ф-2.

КФ размечаются в корпусе для всех движений, которые (а) специально выполняются для выражения некоторой функции или (б) могут быть достаточно однозначно поняты адресатом как выражающие некоторую функцию. КФ для движения отмечается только при согласии не менее двух экспертов. В корпусе используются 35 тэгов функциональной разметки (например, *понимание-согласие-одобрение*, *отрицание-несогласие-возражение*, *апелляция*, *остановка-адресата*, *я-хезитация* и др.) [Котов, Зинина 2015a; Котов, Зинина, 2015b]. В программе ELAN вручную размечаются КФ для мимики, а также движений рук, головы и тела, при этом большинство таких функций не привязаны к единственному способу выражения. Например, КФ *апелляция* в 15,8 % случаев выражается с помощью мимики, в 46,1 % — движениями головы, в 24,5 % — жестами рук, и в 13,6 % — движениями тела. Вместе с тем, анализ разметки позволяет выделить спектр типичных представителей отдельной КФ, например, кивок головой или движение вниз кистью руки для выражения *понимания-согласия-одобрения*. Такие типичные представители КФ рисуются в графическом редакторе Blender и сохраняются в базу данных MySQL. Далее — жест может быть извлечён из базы с помощью скриптов на языке BML — Behavior Markup Language [Kopp, Krenn и др. 2006; Vilhjálmsson, Cantelmo и др. 2007].

Простые сценарии невербального коммуникативного поведения были построены нами с помощью генеративной грамматики в среде Prolog, где КФ рассматриваются как нетерминальные символы и при порождении поведения заменяются реальным жестами и элементами мимики [Котов, Зинина, 2015b]. При таком подходе на каждом шаге на исполнение передаётся один элемент поведения.

Однако в реальном поведении, которое мы наблюдаем в корпусе, проявляются более сложные коммуникативные реакции. Информант может одновременно выражать несколько КФ с помощью разных исполнительных органов: рук, головы, мимики и глаз. В этом случае мы говорим о *наложении* КФ. Например, в случае *20081215-fipp-a05*(*00:52.120–00:54.550*) информант смотрит вверх, крутит головой, машет рукой и говорит: *Нет, нет, нет, нет, нет. Сейчас объясню. Сейчас* (рис. 1).

Рис. 1. **Совмещение коммуникативных функций в корпусе REC: задумчивость (выражается головой и глазами) и отрицание (выражается жестом руки),** *20081215-fipp-a05*(*00:54.010*)

В этом примере движения глаз размечены как *я-размышление*, а движения головы и рук — как *отрицание-несогласие-возражение*. Наложение именно этих функций достаточно часто наблюдается в корпусе (всего 29 случаев). При таком пересечении *я-размышление*, как правило, выражается с помощью глаз (24 случая), а *отрицание-несогласие-возражение* — с помощью головы и рук (14 случаев). Такое сложное поведение невозможно реализовать на роботе только с использованием генеративной грамматики, где функции воспроизводятся последовательно. Поэтому, с целью синтеза правдоподобного поведения была разработана архитектура, которая позволяет комбинировать несколько элементов коммуникативного поведения (жестов, элементов мимики или даже их частей) с помощью тэгов BML, отмечающих важные для жеста исполнительные органы.

Поступающая на робота инструкция (жест) выполняется по временным меткам. При активации временной метки, связанные с ней инструкции передаются для исполнения на приводы робота (для жестов), на экран (для мимики) и на аудиосистему (для синтеза речи). У нарисованного жеста могут вызываться только те теги BML, которые задействуют важный для понимания КФ исполнительный орган. Например, у жеста с функцией *я-размышление* активируются временные метки, связанные с мимикой, а у жеста с функцией *отрицание-несогласие-возражение* — с жестами головы и рук (рис. 2).

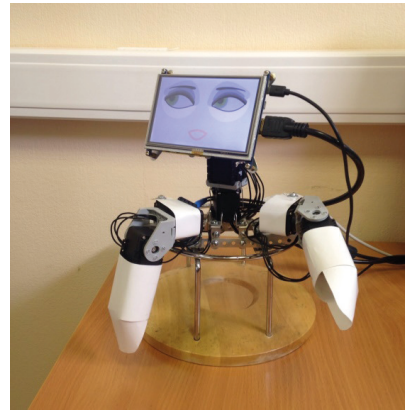

Рис. 2. **Робот Ф-2, совмещающий коммуникативные функции задумчивости (*я-размышление* — направление взгляда) и отрицания (движение левой рукой) из фрагмента корпуса REC** *20081215-fipp-a05*(*00:54.010*)

Таким образом, наложение коммуникативных функций — это важная особенность действий человека в реальной коммуникации, которая может быть реализована на роботе и позволяет приблизить поведение робота к реальному поведению человека.

Литература

1. *Котов А. А., Зинина А. А.* (2015a), Функциональная разметка коммуникативных действий в корпусе «REC» // Труды международной конференции «Корпусная лингвистика — 2015». СПб.: СПбГУ, с. 287–295.
2. *Котов А. А., Зинина А. А.* (2015b), Функциональный анализ невербального коммуникативного поведения // Компьютерная лингвистика и интеллектуальные технологии. Вып. 14. Т. 1. М.: РГГУ, с. 299–310.
3. *Kopp S., Krenn B., Marsella S., Marshall A., Pelachaud C., Pirker H., Thórisson K., Vilhjálmsson H.* (2006), Towards a Common Framework for Multimodal Generation: The Behavior Markup Language in Intelligent Virtual Agents, pp. 205–217.
4. *Kotov A., Budyanskaya E.* (2012), The Russian Emotional Corpus: Communication in Natural Emotional Situations // Компьютерная лингвистика и интеллектуальные технологии. Вып. 11 (18). Т. 1. М.: РГГУ, pp. 296–306.
5. *Vilhjálmsson H., Cantelmo N., Cassell J., E. Chafai N., Kipp M., Kopp S., Mancini M., Marsella S., Marshall A., Pelachaud C., Ruttkay Z., Thórisson K., van Welbergen H., van der Werf R.* (2007), The Behavior Markup Language: Recent Developments and Challenges in Intelligent Virtual Agents, pp. 99–111.

References

1. *Kotov A., Zinina A.* (2015a), Funkcional'naja razmetka kommunikativnyh dejstvij v korpuse "REC" [Functional annotation of communicative actions in REC corpus]. In: Trudy mezhdunarodnoj konferencii "Korpusnaja lingvistika — 2015" [Publications of the international conference "Corpora Linguistics — 2015"]. SPb., Univ of SPb, pp. 287–295.
2. *Kotov A., Zinina A.* (2015b), Funkcional'nyj analiz neverbal'nogo kommunikativnogo povedenija [Functional analysis of nonverbal communicative behavior]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii [Computer linguistics and intellectual technologies] Issue. 14. Vol. 1. Moscow, RSUH, pp. 299–310.
3. *Kopp S., Krenn B., Marsella S., Marshall A., Pelachaud C., Pirker H., Thórisson K., Vilhjálmsson H.* (2006), Towards a Common Framework for Multimodal Generation: The Behavior Markup Language in Intelligent Virtual Agents, pp. 205–217.
4. *Kotov A., Budyanskaya E.* (2012), The Russian Emotional Corpus: Communication in Natural Emotional Situations. In: Komp'juternaja lingvistika i intellektual'nye tehnologii [Computer linguistics and intellectual technologies]. Issue. 11 (18). Vol. 1. Moscow: RSUH, pp. 296–306.
5. *Vilhjálmsson H., Cantelmo N., Cassell J., E. Chafai N., Kipp M., Kopp S., Mancini M., Marsella S., Marshall A., Pelachaud C., Ruttkay Z., Thórisson K., van Welbergen H., van der Werf R.* (2007), The Behavior Markup Language: Recent Developments and Challenges in Intelligent Virtual Agents, pp. 99–111.

Зинина Анна Александровна
Zinina Anna
E-mail: Zinina_aa@nrcki.ru

Котов Артемий Александрович
Kotov Artemy
E-mail: kotov_aa@nrcki.ru

Аринкин Никита Алексеевич
Arinkin Nikita
E-mail: Arinkin_aa@nrcki.ru

Зайдельман Людмила Яковлевна
Zaydelman Lyudmila
E-mail: Zaydelman_ly@nrcki.ru

Национальный Исследовательский Центр «Курчатовский институт» (Россия)
National Research Center "Kurchatov Institute" (Russia)

С. А. Зуева
S. A. Zueva

ПОИСК КЛАСТЕРОВ В СЕТИ ТЕКСТУАЛЬНЫХ СВЯЗЕЙ СЛОВ

SEARCH FOR CLUSTERS IN A NETWORK OF TEXTUAL LINKS OF WORDS

Аннотация. Доклад основывается на корпусе политических статей газеты «Гардиан» за пять лет (16,6 млн словоупотреблений). Для изучения корпуса строится сеть текстуальных связей слов на основе сравнения математического ожидания и наблюдаемой частота совместного употребления каждый пары слов. Грандиозный объем полученной сети (более полумиллиона связей) заставляет искать способы получения кластеров. Кластеризации начинается с формирования трех взаимосвязанных пар слов (троек). Примененный метод позволил выделить 2330 кластеров.
Ключевые слова. Политический дискурс, сеть текстуальных связей.

Abstract. The paper is based on the corpus of political articles of the Guardian newspaper of 5 years (16,6 million word tokens). For the purpose of the corpus analysis, the network of textual links of words is created using comparison of expected and observed frequency of co-occurrences of each pair of words. The size of the received network is huge (more than half a million of links), which compels us to seek ways of defining clusters. Clusterization starts with connecting three interrelated pairs of words (triplets). Under the approach 2330 clusters were identified.
Keywords. Political discourse, network of textual connections.

Данный доклад строится на материале корпуса политических статей газеты «Гардиан» за пять лет (2000, 2005, 2010, 2013 и 2015 год), который в совокупности составляет 16,6 млн словоупотреблений. Ставится задача определить текстуальные связи пар слов, представленных в корпусе. При помощи специально написанной программы корпус делится на равные отрезки по 50 слов, каждый такой отрезок получил свой номер (адрес). Если число общих адресов двух слов существенно отличается от ожидаемого в предположении их независимости (нулевая гипотеза), делается вывод о текстуальной связи этих слов. Рассмотрим пример из политического корпуса «Гардиан» за 2015, где всего 82 650 адресов (N = 82650). Примем нулевую гипотезу, что слова CUT и BUDGET употребляются в тексте независимо друг от друга. Зная частоту[1] этих двух слов CUT (f_1 = 4117), а BUDGET (f_2 = 1981) рассчитаем математическое ожидание (m) того, что они встретятся на одном отрезке по формуле:

$$m = \frac{f_1 \cdot f_2}{N}. \quad (1)$$

[1] Далее под частотой будем подразумевать количество адресов, где встретилось слово.

Подставляя фактические значения имеем $m \approx 98.67$, в действительности частота их совместного употребления в 5 раз больше — $F = 530$. Далее мы будем использовать следующую формулу, для определения меры неслучайности встречи двух слов на одном отрезке [Shaikevich 2001]:

$$S = \frac{F - m - 1}{\sqrt{m}}. \quad (2)$$

Применяя формулу (2) получаем:

$$S = \frac{(530 - 98.678 - 1)}{\sqrt{98.678}} \approx 43.$$

Пары слов со значениями $S = 3$ следует считать полезными для анализа, при $S \geq 4$ можно с уверенностью говорить, что использование данных слов характерно для представленного корпуса[2]. В среднем для каждого года количество полученных текстуальных связей при значении коэффициента $S \geq 3$ составляет 150 000 пар. В дальнейшем будем работать с сетью, образованной стабильными текстуальными связями, наблюдаемыми в 4 или 5 годах одновременно (16 511 текстуальных связей, 5110 разных слов). При работе с такой внушительной сетью возникает необходимость поиска способов выделения кластеров. Один из возможных алгоритмов, которые помогут нам провести кластеризацию на полученном множестве начинается с анализа трех взаимосвязанных пар, например, следующие пары связаны друг с другом и образуют циклический граф: (*voting system*); (*system first-past-the-post*); (*first-past-the-post voting*). Будем считать такую группу минимальным кластером, назовем ее тройкой (triplet), и изображать следующим образом: (“voting”–“system”–“first past the post”).

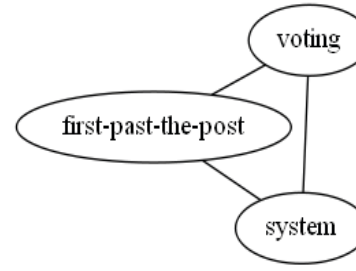

Рис. 1. **Пример минимального кластера**

Минимальные кластеры часто являются только отправной точной для построения кластеров большего размера. Большая часть минимальных кластеров образует сложную

[2] В нашем случае значение $S = 43$ относится к рекордным для политического словаря газеты.

сеть связей вокруг одной вершины, которая является общей для нескольких троек. Будем называть такие кластеры по ключевому слову, например, кластер *"voting"*, который объединил 133 троек в том числе *"voting"–"system"–"first-past-the-post"*); (*"voting"–"system"–"electoral"*); (*"voting"–"system"–"representation"*).

Построенная сеть текстуальных связей позволяет выделить 2330 кластеров и допускает процедуру сравнения между собой для определения на сколько процентов они совпадают. Так, оказалось, что кластеры *"state-school"* и *"headtecher"* состоят на 100 % из одинаковых слов (school, education, teacher, pupil, academy, children). Соответственно мы можем смело провести их объединение и получить кластер *"State school" — "Headtecher"* (*school, education, teacher, pupil, academy, children, state-school, headteacher*). Допустимо проводить и более сложные объединения, так, например, с кластером «*general-election*» совпадают больше, чем на 35 % следующие кластеры[3]: (*byelection 38 %, election 47 %, swing 38 %, win 37 %*).

Опыт построения кластеров вокруг одной вершины с последующим объединением в расширенные кластеры кажется перспективным. Рассмотрим кластер, образованный вокруг вершины *debate*, который состоит из 10 троек (рис. 2.). На графе отчетливо выделяются 3 основные области использования слова *debate* в политическом дискурсе: СМИ, предмет дебатов (поправки, предложения, вносимые в парламенте, какой-либо вопрос и пр.), а также слова, связанные непосредственно с семантикой слова debate (дискутировать, вовлекать, предмет).

Как видим, выбранный алгоритм поиска кластеров позволяет строить такие кластеры, которые включают в себя все области употребления слова типичные для дискурса. Можно использовать модифицированные вариант алгоритма, осуществляющий поиск кластеров, объединенных одним общим ребром. Будем называть такие кластеры по двум вершинам, например, *"Campaign-Party"*. Так, кластер debate разбивается на 3 более узконаправленных кластера: *Commons-Debate, Discussion-Debate и TV-Debate.* Использования данного алгоритма позволило выделить 8848 кластеров, которые объединяют минимум 2 тройки.

[3] В скобках указаны кластеры и в каком отношении они совпадают с основным кластером.

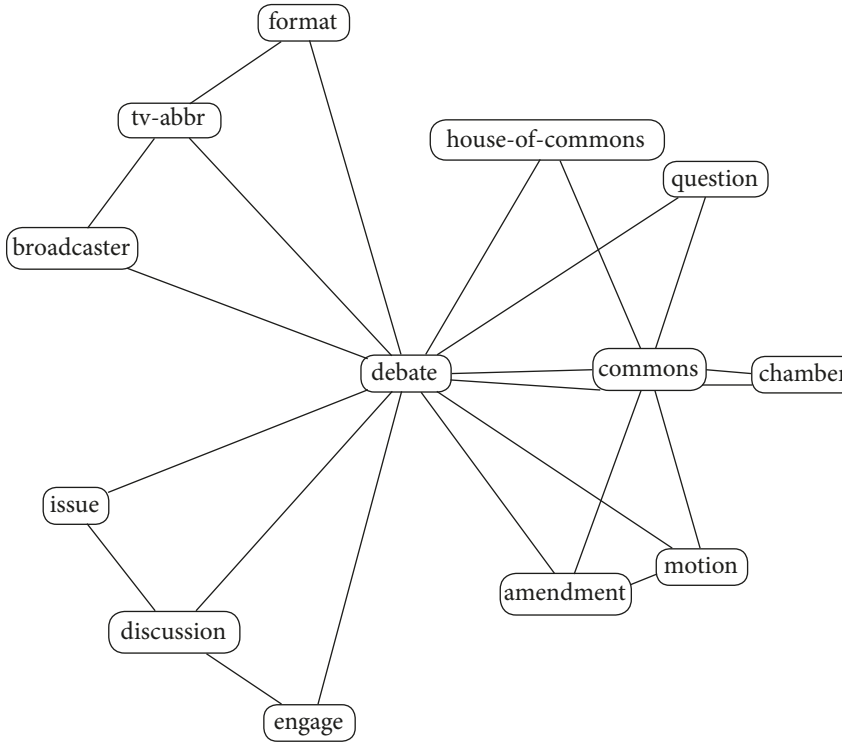

Рис. 2. **Кластер "Debate"**

Аналогично описанному выше методу получения расширенных кластеров по вершине, можно сравнивать между собой кластеры полученные по ребру. Так, например, кластеры *Finance-Economic* и *Borrowing-Finance* совпадают на 95 %. Существуют кластеры, которые полностью поглощаются большими. Так кластер *Election-Poll* одновременно входит в разные кластеры: *Vote-Byelection; Election-Byelection; Seat-Byelection; Poll-Election; Seat-Poll;* и др. Кластер *Budget-Spending* включает в себя кластеры: *Udget-Announce; Budget-Financial; Budget-Health; Chancellor-Finance; Cut-Finance;* и др. Таким образом, возможно проводить объединение кластеров по двум принципам, меньший кластер входит в больший или несколько меньших кластеров входят в состав большего, также два приблизительно равнозначных по количеству слов кластера могут совпадать полностью или частично.

Использование описанных выше методом поиска кластеров в полученной сети текстуальных связей дает позитивные результаты и кажется многообещающим.

References

1. *Shaikevich A. Y.* (2001), Contrastive and comparable corpora: quantitative aspects. In: International journal of corpus linguistics, vol. 6, pp. 229–255.
2. The Guardian Digital Archive. Available at: http://pqasb.pqarchiver.com/guardian/advancedsearch.html

---

Зуева Светлана Александровна
Российский государственный гуманитарный университет
Zueva Svetlana
Russian State Humanitarian University
E-mail: gzueva@gmail.com

Л.Л.Иомдин
Leonid L. Iomdin

МИКРОСИНТАКСИЧЕСКАЯ РАЗМЕТКА В КОРПУСЕ РУССКИХ ТЕКСТОВ[1]

MICROSYNTACTIC TAGGING IN A RUSSIAN TEXT CORPUS

Аннотация. Представлен новый вид разметки в корпусе русского языка «СинТагРус». Эта разметка идентифицирует два типа фразеологических единиц, относящихся к области микросинтаксиса – нестандартные синтаксические конструкции (например, конструкции с повторяющимися лексическими элементами типа *Читать не читал*) и синтаксические фраземы (лексикализованные фразеологические единицы, отличающиеся синтаксической спецификой). При микросинтаксической разметке используются две стратегии: сплошной просмотр текста и целенаправленный поиск последовательностей слов или синтаксических поддеревьев, которые с большой вероятностью относятся к искомым элементам.
Ключевые слова. Микросинтаксис, размеченные корпуса, лексикография, семантический анализ, многозначность фразем.

Abstract. A new type of annotation in a Russian corpus, SynTagRus, is presented. The annotation identifies two types of idiomatic units belonging to the area of microsyntax: nonstandard syntactic constructions (e.g. those with recurring lexical elements like *Čitat' ne čital* ≈ 'I didn't exactly read it') and syntactic idioms (lexicalized idiomatic units characterized by considerable syntactic individuality). Microsyntactic tagging uses two strategies: continuous examination of the whole text and targeted search for word sequences or subtrees which are likely to contain the desired elements.
Keywords. Microsyntax, tagged corpora, lexicography, semantic analysis, ambiguity of idioms.

1. Вводные замечания

В настоящей работе представлен новый вид разметки, которая производится в глубоко аннотированном корпусе русского языка «СинТагРус» (о современном состоянии корпуса см. [Дяченко и др. 2015]). Эта разметка идентифицирует два присутствующих в тексте типа фразеологических единиц, относящихся к области микросинтаксиса[2].

[1] Автор выражает признательность за поддержку данной работы Российскому фонду фундаментальных исследований (грант № 15-04-00562).

[2] Микросинтаксис — раздел лингвистики, занимающий промежуточное положение между словарем и грамматикой, исследуемый автором и несколькими коллегами в течение последних полутора десятилетий и описывающий синтаксически мотивированную идиоматику языка (см., в частности, [Иомдин 2015, Iomdin 2016, Marakasova-Iomdin 2016]). Идейно микросинтаксис близок грамматике конструкций Ч.Филмора и его последователей.

Первый тип — это нестандартные синтаксические конструкции, обладающие нетривиальной семантической спецификой, например, конструкции с повторяющимися словесными элементами типа «X как X» (ср. ***Мальчик как мальчик**, таких много в каждом классе*), «X есть X» (ср. ***Дети есть дети**, они быстро устают*), «X-овать не X-овал» (ср. ***Видеть не видел,*** но *много слышал о нём*) и т. п.

Второй тип — это синтаксические фраземы, или лексикализованные фразеологические единицы, отличающиеся, помимо семантической некомпозициональности, той или иной синтаксической спецификой. Значительную их часть составляют многозначные единицы, такие как *всё равно* (ср. *Он всё равно[1] не послушается*, где имеет место сентенциальное наречие, означающее 'независимо ни от чего', *Мне всё равно[2], куда ехать*, где присутствует предикативное наречие со значением 'безразлично' или «*Сняться в плохом фильме — все равно[3] что плюнуть в вечность*» (Ф. Г. Раневская), где присутствует другое предикативное наречие со значением «равносильно») или *как бы* (ср. *Он как бы[1] предчувствовал опасность*, где присутствует дискурсивная частица со значением сравнения или осторожной номинации и *Мы боялись, как бы[2] он не заболел*, где выступает сильноуправляемый союз, встречающийся при предикатах лексического класса со значением опасения).

Очевидно, что корпус, оснащенный такой разметкой, весьма полезен как для теоретической и практической лексикографии и грамматики, поскольку он позволяет исследовать многообразные контексты, в которых выступают микросинтаксические единицы, так и для широкого класса компьютерно-лингвистических задач, в частности, для задач глубоко семантического анализа текста. До последнего времени в распоряжении исследователей не было корпусов с фразеологической разметкой, хотя нужда в них отчетливо осознается (см., например, [Grzybek & Jesenšek 2014]). По этой причине появление такой разметки в СинТагРус'е можно считать первым опытом частичной фразеологической аннотации — во всяком случае, для русского языка.

Микросинтаксическое аннотирование корпуса представляет собой достаточно сложную задачу. Одна из причин состоит в том, что сколько-нибудь полного списка микросинтаксических единиц русского языка еще не существует. Поэтому мы при разметке использовали две стратегии:

1) сплошной просмотр текста и отыскание в нем кандидатов в микросинтаксические элементы;

2) целенаправленный поиск в корпусе линейных последовательностей слов или же синтаксических поддеревьев, состоящих из таких слов, которые заведомо могут образовывать микросинтаксические единицы. Это, например, такие последовательности и поддеревья, как *всё равно*, *как бы*, *как будто, коль скоро, разве что, пока что, только лишь, мало ли, ни разу, черт знает* + вопросительное слово и многие другие.

В обеих стратегиях мы опираемся на материалы создаваемого автором Микросинтаксического словаря русского языка. Как в том, так и в другом случае мы получаем предварительный вариант разметки текста, с которым производится дальнейшая работа. В настоящее время вся такая разметка проводится вручную: у нас пока нет ни правиловых, ни статистических критериев, которые позволили бы автоматизировать эту работу хотя бы частично.

2. Первые результаты

Уже в начале работы над микросинтаксической разметкой корпуса выяснилось, что встречаемость в текстах интересующих нас единиц достаточно высока: в среднем тексте[3] почти четверть предложений содержит хотя бы одну такую единицу [Маракасова-Иомдин 2016].

С технической точки зрения микросинтаксическая разметка корпуса выглядит так: в XML-представление предложения, содержащего микросинтаксический элемент или элементы, вводится специальное поле, в котором отражается имя элемента (обычно это словосочетание или просто последовательность слов) и указываются его линейные границы. Как выбор имени для микросинтаксического элемента, так и идентификация его границ — не совсем элементарные задачи. Неясно, например, как правильно называть элементы, которые могут содержать переменные части (скажем, единицу типа *какого чёрта, в* котором в качестве второго слова вместо *чёрта* могут выступать существительные *дьявола, рожна, хрена* и нек. др. или единицу типа *вот* + вопросительное слово, которая может реализоваться вари-

[3] Для опытной разметки мы использовали типичные тексты СинТагРус'а — научно-популярные, новостные и публицистические. Можно предположить, что в художественных текстах встречаемость микросинтаксических единиц, как и идиоматики в целом, еще выше, а в научных и технических текстах таких единиц меньше, но количественными данными мы пока не располагаем.

антами *вот что, вот кто, вот где*, *вот какой* и т. д. (Мы называем эти единицы, соответственно, *какого черта* и *вот* + ВОПР). Что же касается неочевидных линейных границ конструкции, то в качестве примера можно привести уже упомянутое предикативное наречие *все равно*[2], между словесными элементами которого могут появиться посторонние вставки, ср. *Не все **ли вам** равно, когда он жил, если вы не знаете, кто он такой?* (Ю. Н. Тынянов) или конструкцию типа в (*чьих-либо*) *силах*, как в предложении *Оградить двери от взлома, письма от вскрытия не в **моих** силах.* (Л. К. Чуковская).

Снимок экрана, воспроизведенный ниже на рис. 1, дает представление о микросинтаксической разметке СинТагРус'а. Здесь представлены две микросинтаксические единицы — в *силах*[1] (ср. *Хирург был не в силах помочь ему* — с подлежащим, обозначающим агента) и в *силах*[2] (ср. *Помочь ему было не в силах хирурга* — с инфинитивным подлежащим, выражающим событие или действие). Вторая единица появляется реже, она встречается в корпусе лишь один раз (см. предложение 10). Первая единица выступает во всех остальных предложениях, за исключением 9 и 11, в которых встречаются выражения в *Военно-морских силах и уверенность в собственных силах,* очевидным образом не имеющих отношения ни к одной из двух синтаксических фразем.

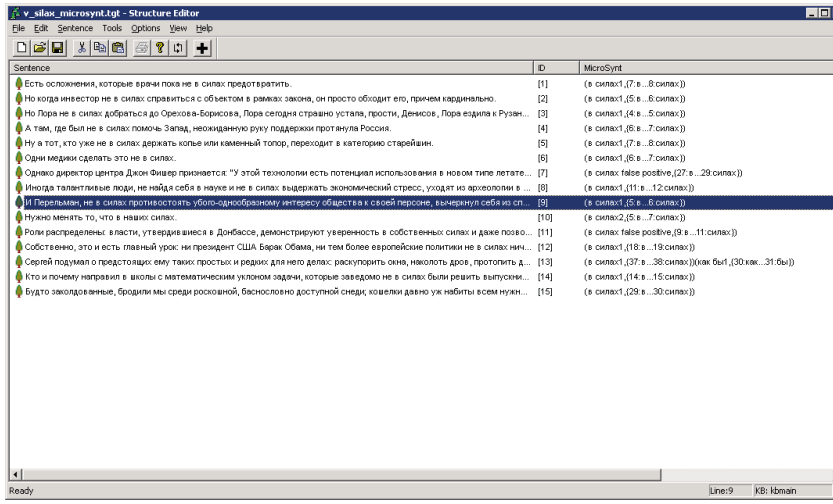

Рис. 1. **Микросинтаксическая разметка корпуса СинТагРус. Фразы с микросинтаксическими единицами в *силах*.**

На нынешней стадии создания корпуса мы считаем целесообразным оставлять в разметке и такие случаи, помечая их признаком false positive («ложное срабатывание», в расчете на то, что такую информацию впоследствии можно будет использовать для целей машинного обучения и автоматизированной идентификации фразеологических элементов и разрешения многозначности.

Принципиально, что микросинтаксическая разметка корпуса СинТагРус производится в дополнение к другим типам аннотации, в первую очередь, к синтаксической разметке, что дает исследователю возможность увидеть, как именно фразема встраивается в синтаксическую структуру предложения.

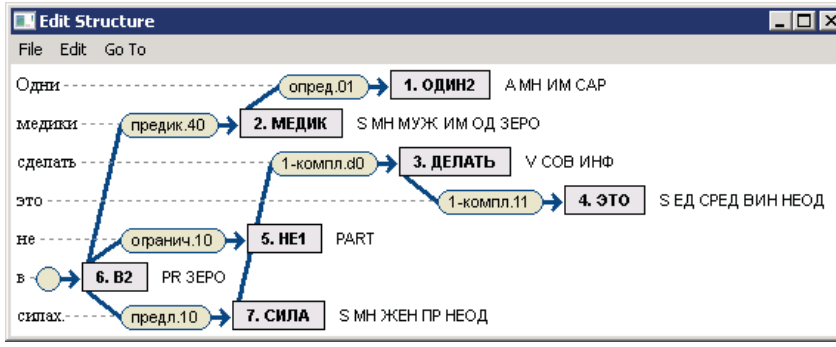

Рис. 2. **Полная древесная синтаксическая структура фразы, содержащей микросинтаксический элемент в *силах*[1].**

Рис. 2 иллюстрирует синтаксическую структуру предложения (9) из снимка экрана на рис. 1. Видно, что предлог в из фраземы в *силах* выступает в качестве вершины сказуемого, подлежащее при котором выражено одушевленным существительным *медик*. Что же касается инфинитива *сделать*, то он подчиняется второму элементу фраземы — слову *силах* по 1-му комплетивному синтаксическому отношению и по существу выражает вторую синтаксическую валентность этой фраземы.

3. Case study: конструкции типа *мало что*

Хотя корпус СинТагРус невелик по объему (он содержит около 1 млн. словоупотреблений (около 67 тыс. предложений), а микросинтаксическая разметка в нем затрагивает всего несколько тысяч пред-

ложений, с ее помощью уже можно получать нетривиальную лингвистическую информацию.

Показательным примером могут служить синтаксические фраземы типа *мало что*. Вхождений этих конструкций в корпусе немного, но и их достаточно, чтобы обнаружить интересную закономерность. Эти конструкции имеют три разновидности: первое слово в них — это *мало, много и редко, а* второе слово — вопросительное местоимение (*что, кто, где, какой* и др., причем ненаречные слова могут стоять в разных падежах и даже сопровождаться предлогами — *мало что, мало чего, мало чему, мало о чем*). Обратим внимание на две оппозиции: *мало чего — мало что и много чего — много что*. Материал корпуса обнаруживает несимметричность в их поведении: варианты *мало что и мало чего* встречаются с соизмеримой частотой, а вариант *много чего* заметно превышает по частоте вариант *много что*. Как кажется, этот факт можно объяснить переосмыслением в языке конструкции *много что* — вместо подчиненного форманта при вопросительном слове (ср. *мало что и кое-что*) слово *много* стало восприниматься как количественное наречие, подчиняющее местоимение *чего в* родительном падеже. Для слова *мало* этот процесс, возможно, начался, но еще не завершился.

Литература

1. *Дяченко П. В., Иомдин Л. Л., Лазурский А. В.* и др. (2015), Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) // Национальный корпус русского языка. 10 лет проекту. Труды Института русского языка им. В. В. Виноградова. М. Вып. 6, с. 272–299.
2. *Иомдин Л. Л.* (2015), Конструкции микросинтаксиса, образованные русской лексемой раз // SLAVIA, časopis pro slovanskou filologii, ročník 84, 2015, sešit 3, s. 291–306.
3. *Маракасова А. А., Иомдин Л. Л.* (2016), Микросинтаксическая разметка в корпусе русских текстов СинТагРус // Информационные технологии и системы 2016 (ИТиС'2016). Сборник трудов 40-ой междисциплинарной школы-конференции ИППИ РАН. Репино, Санкт-Петербург, с. 445–449.
4. *Iomdin L.* (2016), Microsyntactic Phenomena as a Computational Linguistics Issue // Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8–18. (http://aclweb.org/anthology/W/W16/W16–38.pdf). ISBN 978-4-87974-706-8
5. Grzybek P., Jesenšek V. (2014), Phraseology in Dictionaries and Corpora. Introductory Remarks // Phraseologie im Wörterbuch und Korpus. ZORA 97. Maribor, Bielsko-Biała, Budapest, Kansas, Praha, pp. 19–25.

References

1. *Djachenko P. V., Iomdin L. L., Lazursky A. V.* et al. (2015), The State-of-the-Art of a deeply annotated corpus of Russian texts (SynTagRus) [Sovremennoe sostojanie gluboko annotirovannogo korpusa tekstov russkogo jazyka (SinTagRus)]. In: Nacional'nyj korpus russkogo jazyka. 10 let proektu. Trudy Instituta russkogo jazyka im. V. V. Vinogradova. M. Vyp. 6, pp. 272–299.
2. *Iomdin L. L.* (2015), Microsyntactic constructions formed by the Russian word RAZ [Konstrukcii mikrosintaksisa, obrazovannye russkoj leksemoj RAZ]. In: SLAVIA, časopis pro slovanskou filologii, ročník 84, 2015, sešit 3, s. 291–306.
3. *Iomdin L.* (2016). Microsyntactic Phenomena as a Computational Linguistics Issue. In: Grammar and Lexicon: Interactions and Interfaces. Proceedings of the Workshop. Osaka, Japan. 2016, pp. 8–18. (http://aclweb.org/anthology/W/W16/W16–38.pdf). ISBN 978-4-87974-706-8
4. Grzybek P., Jesenšek V. (2014), Phraseology in Dictionaries and Corpora. Introductory Remarks. In: Phraseologie im Wörterbuch und Korpus. ZORA 97. Maribor, Bielsko-Biała, Budapest, Kansas, Praha, pp. 19–25.
5. Marakasova A. A., Iomdin L. L. (2016), Microsyntactic Tagging in the Corpus of Russian Texts SynTagrus. [Mikrosintaksicheskaja razmetka v korpuse russkix tekstov SinTagRus]. In: Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj mezhdisciplinarnoj shkoly-konferencii IPPI RAN. Repino, St Petersburg, pp. 445–449.

Иомдин Леонид Лейбович
ИППИ РАН им. А. А. Харкевича (Москва, Россия)
Iomdin Leonid
Institute for Information Transmission Problems, Russian Academy of Sciences (Moscow, Russia)
E-mail: iomdin@gmail.com

Е. О. Каллас, К. Р. Коппель, Р. Г. Каллас
E. O. Kallas, K. R. Koppel, R. G. Kallas

АВТОМАТИЧЕСКОЕ СОСТАВЛЕНИЕ СЛОВАРЯ КОЛЛОКАЦИЙ НА ОСНОВЕ КОРПУСА

AUTOMATIC CORPUS-BASED COMPILATION OF THE COLLOCATIONS DICTIONARY

Аннотация. В статье рассматриваются вопросы, связанные с автоматическим составлением базы данных словаря коллокаций. В качестве примера используется «Словарь коллокаций эстонского языка», составляемый в Институте эстонского языка г. Таллинна. Словарь содержит 10 000 наиболее частотных слов эстонского языка. База словаря (словник, коллокации, иллюстративные предложения) была автоматически сгенерирована на основе Национального корпуса эстонского языка. Данные были извлечены из системы Sketch Engine [Kilgarriff et al. 2004] в формате XML и импортированы в систему составления словарей Института эстонского языка EELex [Langemets et al. 2006]. Для распознавания и извлечения коллокаций нами были разработаны специальные правила описания лексико-синтаксических конструкций. Дополнительно были разработаны параметры для извлечения иллюстративных предложений. Словарь будет доступен онлайн в 2018 году.
Ключевые слова. Учебный словарь, коллокация, корпус, корпусная лексикография, эстонский язык.

Abstract. This article aims to introduce new resources and methods used in Estonian corpus lexicography to create monolingual Estonian dictionaries. The paper focuses on features offered by Sketch Engine [Kilgarriff et al. 2004], a state-of-the-art lexicographic tool for corpus analysis. For Estonian, Sketch Engine contains different types of corpora, including 463-million-word Estonian National Corpus. Through the example of the Estonian Collocations Dictionary, we analyse how corpus data (headwords, collocations and example sentences) can be automatically extracted from the Estonian National Corpus. The data was extracted in an XML format and imported into the EELex dictionary-writing system [Langemets et al. 2006], where it will be examined, edited and complemented by lexicographers. The Estonian Collocations Dictionary will be published online in 2018.
Keywords. Pedagogical dictionary, collocation, corpus, corpus lexicography, Estonian language.

1. Современные методы автоматического составления словарей

Корпусная лексикография представляет собой направление электронной лексикографии, тесно связанное с компьютерной и корпусной лингвистикой. Объектом ее исследования является возможность использования корпусных материалов при составлении словарей и лексикографических баз данных. Задачей этого направления является разработка методов, позволяющих обеспечивать автоматическое распознавание и отбор лексикографических единиц.

Использование методов корпусной лексикографии предполагает наличие значительного количества корпусов различного содержания.

Если же корпус уже создан, его инсталлируют в систему осуществления корпусных запросов (*Corpus Query System*). К последнему поколению подобных систем принадлежит, например, система Sketch Engine [Kilgarriff et al. 2004; Kilgarriff, Kosem 2012], применяемая также и в Институте эстонского языка. Функции программы позволяют составлять конкорданс и подвергать его всевозможной обработке (*Search*), выявлять статистические коллокаты (*Collocations*), составлять перечни слов (*Word List*), анализировать лексико- синтаксическую сочетаемость слова (*Word Sketch*), осуществлять отбор иллюстративных предложений (*Good Dictionary Example*), составлять тезаурус (*Thesaurus*) и др. Часть функций программы имеет универсальный характер и они легко применимы в отношении любых языков, другая часть функций предполагает создание специальных языковых разработок. Так, например, исследование сочетаемости слова предполагает описание лексико-синтаксических конструкций для каждого языка и написание специальной грамматики (*Sketch Grammar*), [см. подробнее Kallas 2013]. Для автоматического извлечения иллюстративных предложений также необходимо принять во внимание ряд характеристик, свойственных тому или другому языку (например, порядок слов) [см. подробнее Kilgarriff et al. 2008; Kozem et al. 2013]. Исследование Kozem et al. 2013 показало, что автоматическое составление базы данных снижает затраты рабочего времени лексикографов примерно вдвое.

2. Принципы составления базы данных «Словаря коллокаций эстонского языка»

«Словарь коллокаций эстонского языка» является новым проектом Института эстонского языка в сфере учебной лексикографии. Целевой группой словаря являются изучающие эстонский язык на уровнях B2-C1.

В основе базы данных словаря лежит Национальный корпус эстонского языка, объем которого составляет 463 млн слов. На данный момент это самый большой и разнообразный в жанровом отношении корпус эстонского языка. Подкорпусами национального корпуса эстонского языка являются, например, тексты периодических изданий, художественные тексты, юридические тексты, научные тексты, религиозные тексты, интернет-форумы, блоги и др.

Основными информационными единицами словаря являются заглавное слово, толкование (только для многозначных слов), коллокации и иллюстративные предложения.

В базе данных словаря мы группируем коллокаты заглавного слова на основе их частеречной принадлежности. Так, для существительных, например, указывается с какими другими существительными, прилагательными, глаголами, наречиями и предлогами они сочетаются. На рисунке 1 в качестве иллюстративного примера приводятся автоматически выявленные коллокаты слова *arutlus* «обсуждение».

ARUTLUS nimisõna OMADUSSÕNAD • teoreetiline, avalik, pikk, filosoofiline, loogiline, tõsine, huvitav, sisuline **arutlus** NIMISÕNAD • **arutluse** objekt, tulemus, taust, tase TEGUSÕNAD • **arutlus** käib, toimub, algab, keskendub *millele*, jätkub, kestab, tekib • **arutlust** jätkama, alustama, korraldama, kuulama	**ОБСУЖДЕНИЕ** сущ. ПРИЛАГАТЕЛЬНЫЕ • теоретическое, открытое, долгое, философское, логичное, интересное, содержательное **обсуждение** СУЩЕСТВИТЕЛЬНЫЕ • объект, итог, подоплека, уровень о**бсуждения** ГЛАГОЛЫ • **обсуждение** идет, ведется, начинается, концентрируется *на чем*, продолжается, тянется, возникает • продолжать, начинать, организовывать, слушать **обсуждение**

Рис. 1. **Набор коллокатов заглавного слова *arutlus* «обсуждение»**

3. Процесс генерирования базы данных

Генерирование базы данных «Словаря коллокаций эстонского языка» состояло из следующих этапов: 1) разработка структуры базы данных в системе составления словарей Института эстонского языка EELex[1]; 2) генерирование словника словаря; 3) анализ грамматики лексико-синтаксических конструкций (*Sketch Grammar*) и уточнение параметров, необходимых для импорта данных из системы Sketch Engine в систему EELex; 4) описание параметров и классификаторов, необходимых для извлечения иллюстративных предложений; 5) экспорт данных из системы Sketch Engine в виде файла XML; 6) импорт данных в словарную систему EELex.

3.1. Словник

В словник словаря входят 10 000 наиболее частотных существительных, прилагательных и глаголов эстонского языка, а также наре-

[1] Vt http://eelex.dyn.eki.ee/ (01.04.17).

чия образа действия (например, *«salaja» — тайно*, *«kiiresti» — быстро*). В качестве составных заглавных слов в словаре выступают составные (*«alla andma» — сдаваться*) и фразеологизированные глаголы (*«silmas pidama» — иметь в виду*).

При составлении словника мы использовали функцию перечня слов *Word List*, которая позволяет на основе регулярных выражений генерировать перечни различного содержания. В результате мы получили перечень, состоящий из 10 000 наиболее частотных существительных, прилагательных, глаголов и наречий эстонского языка, который подвергли дополнительному анализу. Пришлось удалить из списка заглавные слова, представленные в двоякой форме написания (*«šokk»* vs *«shokk» — шок*), сокращения (*eek*, *eur*, *toim*) и термины (*«süsinikdioksiid» — двуокись углерода*).

3.2. Коллокаты

Для извлечения из корпуса коллокатов заглавного слова нами была использована функция *Word Sketch*. Грамматика *Sketch Grammar* включает в себя 109 правил, позволяющих выявить лексико-синтаксическую сочетаемость заглавных слов словаря. На рисунке 2 представлен пример выявленных словосочетаний существительного *diskussioon* — дискуссия.

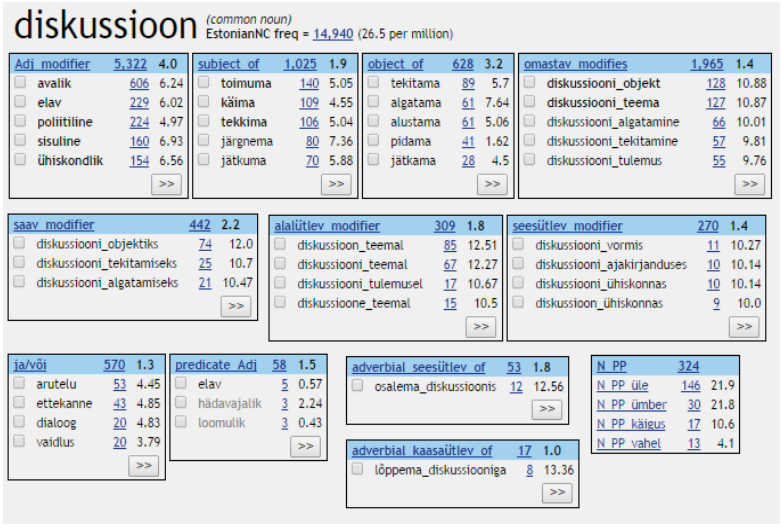

Рис. 2. **Пример словосочетаний с заглавным словом *«diskussioon»* — дискуссия**

3.3. Иллюстративные предложения

Для отбора иллюстративных предложений нами была использована функция Good Dictionary Example [Kilgarriff jt 2008]. GDEX работает как фильтр, позволяющий автоматически распознавать в массиве корпуса предложения, которые подходят для определённого типа словаря в качестве иллюстративных примеров. Нами был разработан целый ряд параметров и классификаторов, позволивших существенно улучшить первоначальные результаты [см. подробнее Koppel, Kallas 2016].

4. Заключение

В данной статье мы описали процесс автоматического генерирования базы данных словаря коллокаций на основе корпуса. Нами были использованы такие функции системы Sketch Engine как составление перечня слов (*Word List*), выявление лексико-синтаксической сочетаемости слова (*Word Sketch*) и отбор иллюстративных предложений (*Good Dictionary Example*). В результате была создана база данных, объём которой составил 10 939 заглавных слов, 493 971 коллокатов и 2 469 855 иллюстративных предложений.

Автоматическое составление позволяет существенно ускорить процесс составления словарей. Проделанная нами работа помогла выявить слабые стороны использованных нами методов и те аспекты, которые требуют существенной доработки в будущем. Перспективным направлением корпусной лексикографии является не только составление баз данных словарей различного типа, но и конечных лексикографических продуктов, которые не будут требовать дополнительной редактуры.

References

1. *Kallas, J.* (2013), Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias [Syntagmatic relations of Estonian content words in corpus and pedagogical lexicography]. Tallinn: Tallinna Ülikool.
2. *Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D.* (2004), The sketch engine. In: Proceedings of the 11th EURALEX international congress. Lorient, France: Université de Bretagne Sud, 105–115.
3. *Kilgarriff, A., Husák, M., McAdam, Rundell M., Rychlý, P.* (2008), GDEX: Automatically finding good dictionary examples in a corpus. In: Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425–432.
4. *Kilgarriff, A., Kosem, I.* (2012), Corpus tools for lexicographers. In: Electronic lexicography. Oxford: Oxford University Press, 31–55.

5. *Koppel, K., Kallas, J.* (2016), Õppijasõbralik korpuslause: automaatse valiku võimalusi [User-friendly corpus sentence: Parameters for automatic selection]. Lähivõrdlusi. Lähivertailuja, 26, 222–250, 10.5128/LV26.07.
6. *Kosem, I., Gantar, P., Krek, S.* (2013), Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In: Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia, 17–19.
7. *Langemets, M., Loopmann, A., Viks, Ü.* (2006), The IEL dictionary management system of Estonian. In: G.-M. De Schryver (ed.) DWS 2006: Proceedings of the 20 Fourth International Workshop on Dictionary Writing Systems. Turin, 5th September 2006. Turin: University of Turin, pp. 11–16.

Каллас Елена
Kallas Jelena
E-mail: jelena.kallas@eki.ee

Коппель Кристина
Koppel Kristina
E-mail: kristina.koppel@eki.ee

Институт эстонского языка (Эстония)
Institute of the Estonian language (Estonia)

Каллас Роман, доктор филологии
Kallas Roman, PhD
E-mail: roman.kallas@mail.ee

С. Н. Карпович
S. N. Karpovich

ВИЗУАЛИЗАЦИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ С ПОМОЩЬЮ IPYTHON

VISUALIZATION TOPIC MODEL WITH IPYTHON

Аннотация. В работе предложен подход к представлению результатов тематического моделирования, отличающийся своими возможностями по выбору метода визуализации. Проведен анализ существующих подходов к визуализации тематических моделей. Разработана система, позволяющая выбирать источник данных для создания тематической модели, изменять параметры моделирования и представлять результат тематического моделирования с помощью iPython. Сделана визуализация тематической модели, построенной на корпусе SCTM-ru.
Ключевые слова. Корпус текстов, обработка текста на естественном языке, тематическое моделирование.

Abstract. The paper introduces an approach to topic model visualization that is characterized by wide possibilities of choosing a method of visualization. The existing approaches to topic models visualization have been analyzed, and a system, which allows choosing data source for topic models, changing modeling parameters and visualizing the result of topic modeling with IPython has been developed. The example of topic model visualization has been built using the SCTM-ru corpus of original news text.
Keywords. Text corpora, topic model, natural language processing.

1. Введение

Алгоритмы тематического моделирования являются одним из перспективных направлений машинного обучения и дистрибутивного анализа текстов на естественном языке. Дистрибутивный анализ — это метод исследования языка, основанный на изучении окружения отдельных единиц в тексте, не использующий сведений о полном лексическом или грамматическом значении этих единиц. Наиболее широко известны следующие дистрибутивно-семантические модели: модель векторных пространств, латентно-семантический анализ, тематическое моделирование, предсказательные модели дистрибутивной семантики.

Тематическое моделирование — это способ построения тематической модели коллекции текстовых документов. Тематическая модель (далее ТМ) позволяет группировать текстовые документы, определять, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему.

При создании ТМ специалисту необходимо оценить качество построенной модели, получить представление о распределении слов и тем, выявить ошибки.

2. Обзор существующих подходов к задаче визуализации результатов тематического моделирования

В работе [1] предложен метод визуализации ТМ, включающий программное обеспечение с открытым исходным кодом. Основная идея метода заключается в том, что визуализация модели обобщает и организует коллекцию документов. Каждая тема связывает несколько документов, и каждый документ связывает несколько тем. Пример представления показан на рисунке 1.

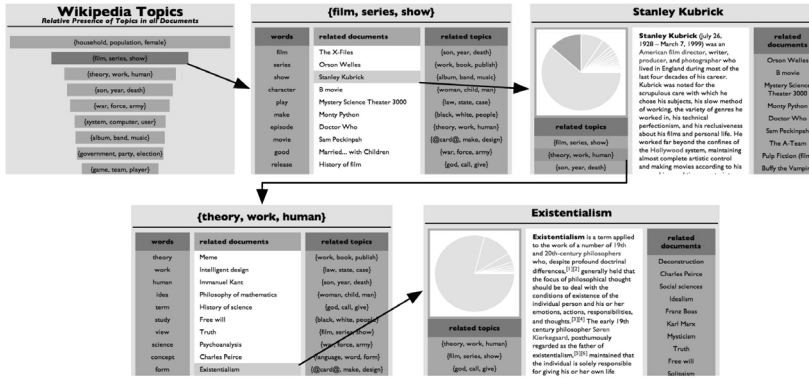

Рис. 1. **Пример интерфейсов визуализации ТМ**

В работе [2] представлены инструменты для визуального анализа и оценки качества тематического моделирования. Предлагаемый подход содержит два основных интерфейса: отображение матрицы отношений слово-тема и отображение документа. ТМ представлена в виде матрицы, где строки — это слова, а столбцы — это темы. Система обладает продвинутым интерактивным графическим интерфейсом, реализованным с помощью JavaScript библиотеки d3.js.

Существуют несколько систем и инструментов для визуального анализа ТМ, но каждый из них обладает существенными недостатками. Поэтому задача визуализации результатов тематического моделирования, и разработка системы, позволяющей создавать и перестраи-

вать ТМ, а затем представлять в интуитивно понятном виде промежуточные и итоговые результаты, по-прежнему актуальна.

3. Подход к визуализации результатов тематического моделирования

ТМ задает отношение между темами и документами в корпусе текстов. Одна из самых распространенных ТМ — это латентное размещение Дирихле (LDA) [3], эта модель является обобщением вероятностного семантического индексирования. Другие ТМ, как правило, являются расширением LDA. Результатом тематического моделирования являются вероятностные распределения слов и документов по одной группе кластеров.

Требования к разрабатываемой системе: управление параметрами ТМ, выбор источника данных, модульность системы, подключение и замена необходимых программных библиотек, представление значимых слов для темы, отображение близости тем друг другу, интерактивная визуализация, представление темпоральной ТМ.

Для реализации системы были выбраны следующие средства: языковая платформа Python, интерактивная оболочка iPython, дистрибутив python Anaconda, библиотеки: gensim, pyldavis, matplotlib. Вместе выбранные программные средства обеспечивают необходимую функциональность для реализации программного комплекса. На рисунке 2 представлена архитектура системы визуализации ТМ с помощью iPython. Источником данных является корпус текстов SCTM-ru, модуль TM-LDA с помощью библиотеки gensim создает ТМ, модуль Calculate осуществляет предварительный расчет данных для визуализации ТМ на временном ряду, модуль Visualization отвечает за представление результата тематического моделирования.

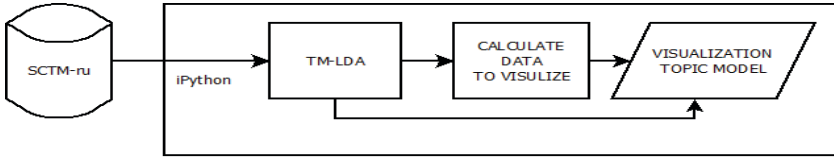

Рис. 2. **Архитектура системы визуализации ТМ**

В качестве примера визуализации создадим ТМ с помощью библиотеки gensim на специальном корпусе текстов SCTM-ru, впервые

представленном в работе [4]. Первая версия корпуса была создана на данных сайта Русские Викиновости с 2005 по 2014 год. Используем вторую версию корпуса, в который добавлены новости до 2017 года. Стандартный вывод ТМ — это список наиболее вероятных слов для каждой темы с численной оценкой вероятности.

На рисунке 3 представлен интерактивный высокоуровневый интерфейс, созданный с помощью библиотеки pyLDAvis. В левой части с помощью диаграммы Элерра-Венна представлены темы. Представление позволяет оценить, насколько близки темы, те, что расположены рядом, находятся ближе, те, в которых общие слова встречаются реже, находятся дальше друг от друга. С правой стороны представлен набор слов, наиболее характерных для выбранной темы.

В практических задачах часто приходится иметь дело с анализом потока текстовых документов, например, потока новостей. В данной работе под текстовым потоком понимается последовательность текстовых документов с определенным для каждого документа временем создания. ТМ, анализирующие поток текстовых документов, называют темпоральными ТМ. Чтобы визуализировать динамику изменения тем во времени, необходимо провести предварительный расчет данных ТМ.

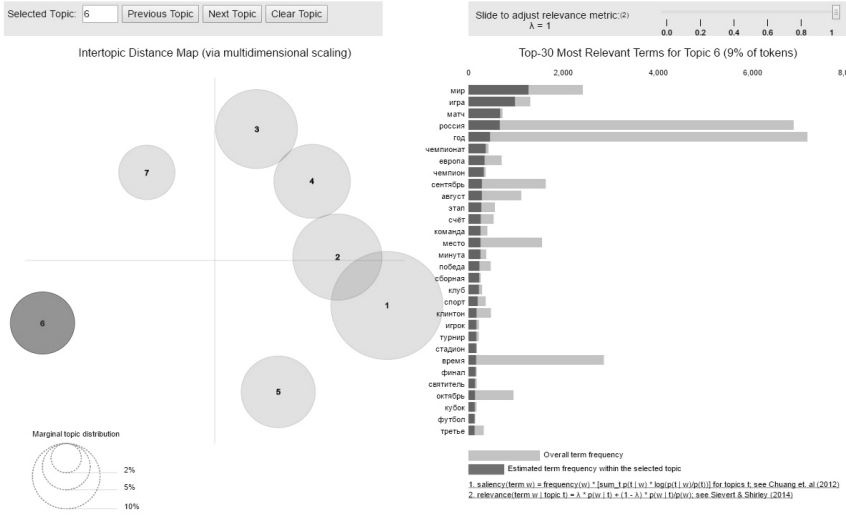

Рис. 3. **Высокоуровневый интерфейс, построенный с библиотекой pyLDAvis**

На рисунке 4 представлен следующий метод визуализации ТМ во времени, диаграмма-области построена на нормированных данных. Каждой теме соответствует свой цвет на диаграмме. Представление позволяет проследить повышение популярности или зарождение новой темы и снижение популярности другой.

С помощью библиотеки matplotlib возможна реализация представления темпоральных ТМ с помощью гистограмм и с диаграмм-линий.

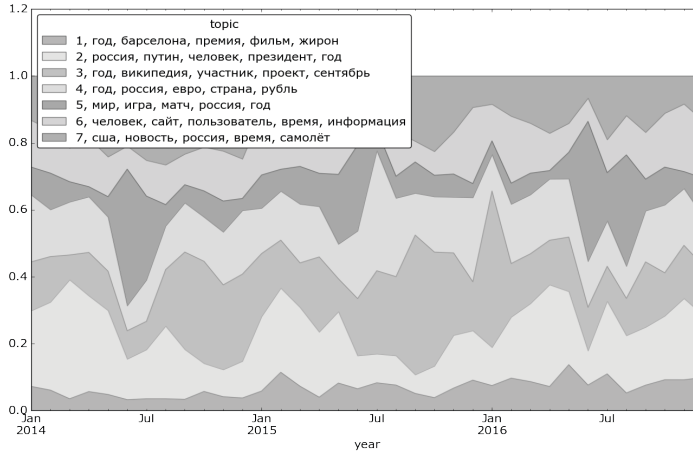

Рис. 4. **Нормализованное представление тематического моделирования во времени**

4. Заключение

В результате проделанной работы создана система для простого внедрения в рабочий процесс, с возможностью выбора подходящего метода визуализации и интуитивно понятным представлением результатов тематического моделирования. Особенностями системы являются гибкие возможности по настройке и управлению процесса создания ТМ, выбор источника данных и использование подходящих библиотек для визуализации. Наличие высокоуровневого интерфейса позволяет специалисту взаимодействовать с результатами тематического моделирования, в удобном виде представлять ТМ.

Поставленные в работе цели достигнуты. Разработанный подход может быть расширен новыми подходящими методами. Система доступна для свободного использования в некоммерческих целях https://github.com/cimsweb/Topic-Model-Visualization-With-IPython.

Литература

1. *Chaney A. J. B., Blei D. M.* (2012), Visualization Topic Models. ICWSM
2. *Chuang J., Manning C. D., Heer J.* (2012), Termite: Visualization techniques for assessing textual topic models // Proceedings of the International Working Conference on Advanced Visual Interfaces. ACM, 2012, pp. 74–77.
3. *Blei D. M., Ng A. Y., Jordan M. I.* (1999), Latent Dirichlet Allocation. Journal of Machine Learning Retrieval, pp. 993–1022
4. *Карпович С. Н.* Русскоязычный корпус текстов SCTM-RU для построения тематических моделей // Труды СПИИРАН. 2015. Т. 2, № 39, с. 123–142.

References

1. *Chaney A. J. B., Blei D. M.* (2012), Visualization Topic Models. ICWSM
2. *Chuang J., Manning C. D., Heer J.* (2012), Termite: Visualization techniques for assessing textual topic models. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. ACM, 2012, pp. 74–77.
3. *Blei D. M., Ng A. Y., Jordan M. I.* (1999), Latent Dirichlet Allocation. In: Journal of Machine Learning Retrieval, pp. 993–1022.
4. *Karpovich S. N.* (2015), The Russian language text corpus for testing algorithms of topic model. In: Trudy SPIIRAN — SPIIRAS Proceedings. Vol. 39, pp. 123–142.

Карпович Сергей Николаевич
ООО «Олимп» (Россия)
Karpovich Sergey
Olimp LLC (Russia)
E-mail: cims@yandex.ru

Г. Е. Кедрова, С. Б. Потемкин
G. E. Kedrova, S. B. Potemkin

СРЕДСТВА МАШИННОГО ПЕРЕВОДА ПРИ ОБРАБОТКЕ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ

MACHINE TRANSLATION TOOLS FOR THE PARALLEL TEXTS PROCESSING

Аннотация. Выравнивание параллельных текстов на языках с малыми лингвистическими ресурсами (словари, тезаурусы, корпуса и пр.) представляет значительные трудности, но бывает необходимо именно для расширения доступности текстов. Особенно полезно выравнивание текстов если один язык (напр. английский) имеет развитое обеспечение, а другой, (напр. непальский) недостаточно представлен в интернете (отсутствие двуязычных словарей, словарей синонимов, грамматической структуры). В этом случае полезно провести предварительное выравнивание параллельных текстов на двух языках по предложениям, а затем выровнять полученную заготовку вручную. В статье предлагается метод для такого выравнивания на базе онлайн-переводчика и средств динамического программирования.
Ключевые слова. Выравнивание, эквивалент, разделение на предложения, машинный перевод.

Annotation Alignment of parallel texts in languages with few linguistic resources (dictionaries, thesauri, corpora, etc.) presents considerable difficulties, but it is necessary precisely to increase the availability of texts. Especially important the case when one text language (e.g. English) has developed support, and the other one (e.g. Nepalese) is not sufficiently represented on the Internet (absence of bilingual dictionaries, dictionaries of synonyms, codified grammatical structure). If so it is useful to pre-align the parallel texts in two languages on the level of sentences, and then align the resulting sketch manually. This article proposes a method for sketch alignment based on the online translator and dynamic programming tools.
Keywords. Alignment, equivalent, division into sentences, machine translation.

Введение

Для параллельных текстов, а именно когда имеется текст на исходном языке и его перевод на целевой язык, часто требуется выполнить выравнивание исходного и переведенного текстов на уровне предложений. Это необходимо для построения базы данных памяти переводов, а также для дальнейшего анализа уже на уровне предложений, выделения однословных эквивалентов и фрагментов в двух предложениях, не имеющих пословного перевода. Процедура выравнивания текстов по предложениям успешно решается с использованием двуязычного словаря [Потемкин, Кедрова 2008]. Ситуация усложняется, если в распоряжении исследователя нет соответствующего машиночитаемого словаря или даже словаря на бумажном носителе, который

можно было бы оцифровать. Мы предлагаем процедуру, использующую механизм онлайн переводчика, свободно доступный на многих сайтах. Мы использовали Google транслятор, который предлагает большое число пар исходный язык — целевой язык и поддерживает API интерфейс, что позволяет автоматизировать весь процесс выравнивания.

Этапы выравнивания

Процесс выравнивания состоит из следующих этапов:

— Разделение исходного и целевого текстов на предложения. В первом приближении можно в качестве границ предложений принять точку, восклицательный, вопросительный знак и двоеточие. Необходимо позаботиться о том, чтобы не возникало шума, связанного с инициалами, многоточием, сокращениями вида и *т. д., пр., etc., i.e., Mr., St.*
— Перевод каждого предложения исходного текста с использованием онлайн переводчика. Необходимо обеспечить перевод именно отдельных предложений, без связывания их в цельный текст. Этого можно добиться, вводя после каждого предложения двойного интервала, непереводимой последовательности символов, напр. ################ и т. п.
— Выравнивание по предложениям двух переводов текста — выполненного онлайн переводчиком и имеющимся текстом, выполненным, как правило, профессиональным переводчиком. Это не простая процедура, потому что два перевода могут иметь различное количество предложений и границы предложений могут не совпадать в обоих текстах. Для выравнивания мы используем метод динамического программирования, где в качестве меры близости принято число совпадающих слов в предложении целевого текста и текста перевода, выполненного онлайн-транслятором.
— Сопоставление предложений исходного текста — предложениям текста, выполненного онлайн-переводчиком.
— Представление выровненных текстов в удобочитаемом виде, напр., в виде таблицы, для окончательного редактирования человеком.

Пример выравнивания: русско-португальский

В качестве иллюстрации выбраны параллельные тексты: рассказ А. П. Чехова *«Анна на шее»* и его перевод на португальский *«Anna no pescoço»* (*Tchekhov Anton 2009*). Число слов в переводе около 5000 в 300 предложениях; в русском оригинале около 5000 слов в 280 предложениях. Очевидно, что прямое сопоставление предложение — предложение невозможно.

Пример выровненных предложений из рассказ А. П. Чехова *«Анна на шее»* и перевода рассказа на португальский *«Anna no pescoço»*

a) 161 А от отца она унаследовала темный цвет волос и глаз, нервность и эту манеру всегда прихорашиваться (Исходное предложение)
b) 161 Um **pai** que *herdou cabelo escuro* e **olhos** nervos , e este estilo e **sempre** animar .(GOOGLE — перевод)
c) 179 Do **pai** , por outro lado , ela *herdara* os *cabelos* e **olhos** *escuros* , o nervosismo e o costume de andar **sempre** embonecada .(перевод, выполненный человеком)

Совпадающие слова выделены жирным шрифтом; слова с одним и тем же корнем выделены курсивом. Мера близости между предложениями b) и c) равна 3 (*pai, olhos, sempre*) *а с* учетом однокоренных слов (*herdou cabelo escuro*) — 6.

Динамическое программирование как инструмент выравнивания

Метод динамического программирования широко известен в различных областях вычислительной математики. Его сущность заключается в поисках среди всех возможных путей оптимального пути между двумя точками — начальной и конечной точками. Поиск оптимального пути (в нашем случае пути с максимальным весом) выполняется двойным проходом — вначале для каждой точки находим предыдущую, путь до которой уже является оптимальным и присваиваем этой точке ссылку на предыдущую. Затем — для конечной точки строим путь по ссылкам до начальной точки. В нашем случае за вес пути принимается число слов, совпадающих в двух предложениях (*pai, olhos, sempre*), т. е. вес данного предложения, участвующего в оптимальном пути равен 3.

Заключение

Нами подготовлены для дальнейшей обработки предварительно выровненные тексты переводов Чехова на английский (210 рассказов), немецкий (94 рассказа), французский (188 рассказов), испанский (71 рассказ), итальянский (44 рассказа), португальский (13 рассказов). Выбор языков связан с тем, что к работе по окончательному выравниванию текстов предполагается привлечь студентов романо-германского отделения филологического факультета МГУ в рамках летней производственной практики.

Литература

1. *Потемкин С. Б, Кедрова Г. Е* (2008), Выравнивание неразмеченного корпуса параллельных текстов // Труды Международной конференции «Диалог 2008», место издания Москва, с. 431–437.
2. *Tchekhov Anton* (2009), A Dama do Cachorrinho Tradução: Maria Aparecida Botelho Pereira Soares L&PM, 2009, pp. 62–71.

References

1. *Potemkin S. B., Kedrova G. E.* (2008), Alignment of the Unannotated Corpora of Parallel Texts. In: Proceedings of the International Conference "Dialogue 2008", published in Moscow, pp. 431–437.
2. *Tchekhov Anton* (2009), A Dama do Cachorrinho Tradução: Maria Aparecida Botelho Pereira Soares [A Madame with a Dog, Translated by: Maria Aparecida Botelho Pereira Soares], published by L&PM, pp. 62–71.

Кедрова Галина Евгеньевна
E-mail: kedr@philol.msu.ru

Потемкин Сергей Борисович
E-mail: prolexprimr@gmail.com

Московский Государственный Университет им. М. В. Ломоносова

Э. С. Клышинский, Д. В. Королев, А. А. Власова
E. S. Klyshinsky, D. V. Korolyov, A. A. Vlasova

МЕТОД ПОИСКА ГРУПП СЛОВ В ТЕКСТАХ С НЕСНЯТОЙ ОМОНИМИЕЙ[1]

A METHOD FOR SHALLOW PARSING OF GRAMMATICALLY AMBIGUOUS TEXTS

Аннотация. Существующие системы поиска фрагментов в текстах требуют предварительного снятия омонимии или работа с токенами. Мы разработали систему для поиска в текстах с учетом омонимичности слов. Система была использована для сравнения употребления слов в текстовых корпусах различных стилей. Также с ее помощью была сформирована база синтаксически связанных слов большого объема.
Ключевые слова. Грамматическая неоднозначность, частичный синтаксический анализ, информационный поиск.

Abstract. Existing software system for shallow parsing are processing texts that was previously disambiguated. We have developed a notation and software system that allows retrieve contact groups of words without text disambiguation. The system was used for comparative analysis of words usage in corpora written in different styles. We also used the system for extraction of syntactically connected words from large corpora.
Keywords. Grammatical ambiguity, shallow parsing, information retrieval.

Программные инструменты для синтаксического анализа текстов являются рабочим инструментов лингвистов на протяжении уже долгих лет. Однако для решения частных задач не всегда требуется проводить полный синтаксический анализ текста или его фрагментов. Частичный синтаксический анализ позволяет извлечь только интересующие исследователя связи, при этом занимает меньше вычислительных ресурсов и проще поддается формализации. В качестве примера практического применения здесь можно привести автоматическое извлечение терминов из текстов предметной области. Так, в работе [Захаров 2014] показано, что термины чаще всего представляют собой именную группу, извлекаемую именно с использованием частичного синтаксического анализа.

Для проведения частичного синтаксического анализа текстов разработаны такие системы, как NLTK [Bird 2009], CQL [CQL 2017] и LSPL [Большакова 2010]. Одним из недостатков (или достоинством) этих систем является необходимость работы с текстом со снятой омо-

[1] Данная работа выполнена при финансовой поддержке гранта РГНФ 15-04-12019

нимией. С одной стороны, снятие омонимии позволяет существенно ускорить поиск. Однако, с другой стороны, само по себе исследование омонимии составных конструкций в текстах может являться предметом научного поиска. Помимо этого, как было показано в наших предыдущих работах [Клышинский 2015], тексты на русском языке обладают большим (порядка 80 %) количеством слов, не омонимичных по части речи. Этот факт позволяет искать некоторые синтаксические конструкции (например, упоминавшиеся уже именные группы) в размеченных текстах без снятой омонимии с высокой степенью полноты (хотя и с падением точности).

В связи с этим мы разработали новую нотацию для поиска контактных фрагментов текста, ориентированную на работу с омонимичной разметкой и основанную на формализме регулярных выражений.

Помимо стандартного синтаксиса регулярных выражений, нотация включает в себя ряд особенностей. Терминальный символ описывает искомое слово в тексте и позволяет находить среди результатов морфологического анализа этого слова один вариант из множества, несколько вариантов из множества или единственный вариант разбора. Терминальный символ в нотации ограничен квадратными скобками и включает начальную форму, часть речи и грамматические параметры (разделяются точкой с запятой и могут опускаться). Например, [;prep;] — любой предлог, а [БЫТЬ;verb;] — глагол «быть» в любой форме.

Имя и значение грамматического параметра задаются раздельно. Равенство параметра некоторому значению обозначается как «название=значение». Так, глагол «быть» в прошедшем времени и мужском роде записывается как [БЫТЬ;verb; tense=past, gender=m], а произвольное существительное в именительном падеже [;noun; case=i]. При такой записи слово должно обладать хотя бы одним вариантом разбора, попадающим под заданный шаблон.

Для обозначения уникальности набора значений используется восклицательный знак, который ставится перед значением. Например, все варианты анализа слова являются существительными: [;!noun;]. Поиск уникальных значений реализован для начальной формы, части речи или всего набора значений в целом: [!КРАСНЫЙ ;adj; case=i, gender=f] — прилагательное в именительном падеже и женском роде, начальная форма — только «красный».

Варианты анализа одного слова записываются при помощи знака &. Например, существительное в именительном падеже, омонимичное глаголу: [;noun; case=i & ;verb;] .

Для проверки согласования используется знак + после названия параметра. Именная группа, в которой прилагательные согласуются с существительным: ([;!adv;]?[;!adj; gender+, number+, case+])+[;!noun; gender+, number+, case+] — положительная итерация наречия (может отсутствовать) и прилагательного после которой идет существительное; род, число и падеж слов совпадают; все слова однозначны по части речи. Отсутствие согласования указывается при помощи знака минус: [;noun; case+][;adj; case-] — существительное, за которым следует прилагательное в другом падеже.

Если значение параметра не равно определенному значению, перед ним ставится минус: [;adj; comp=-com] — прилагательное, не в сравнительной степени.

Нами было разработано программное обеспечение, находящее нужные фрагменты в корпусах большого размера. Его плюсом является возможность обрабатывать тексты на локальной машине в режиме он-лайн (за счет отсутствия модуля снятия омонимии его скорость работы около 1 млрд токенов в час). Программа извлекает как сочетания, так и предложения, в которых они употребляются.

Программа была использована для исследования лексики русского языка. На примере слова «гордый» были исследованы особенности статистики употребления слов в корпусах различных стилей. [Lukashevich 2016]

Однако, основным предназначением программы было извлечение синтаксически связанных словосочетаний из больших корпусов текстов. Извлекались связи существительных и прилагательных, входящих в именные и предложные группы, глагольные группы (глагол + предлог + существительное, причастия и деепричастия считались формами глагола), связи глагола с наречием. Здесь скорость работы программы оказалась важна, так как общий размер корпусов достиг 17 млрд словоупотреблений, извлеченных из корпусов разного стиля: новостных, научных, беллетристика, Википедия. Применение полного синтаксического анализа сделало бы данный проект нереальным, а снятие омонимии замедлило бы работу как минимум на порядок.

Основной проблемой здесь оказалась разработка правил, гарантирующих наличие синтаксической связи между словами. Так, например, помимо очевидного требования к согласованию существительных и прилагательных по роду, числу и падежу, необходимо было определить, что прилагательное не находится в краткой форме. В противном случае в результаты попадают фразы с прилагательным, подчиняю-

щимся глаголу, а не существительному: *Будет ли* ***прилично пальто****, которое мы купили в прошлом году?*

Проверка соседей слов позволила также извлекать конструкции, содержащие омонимичные слова. Так, например, при помощи правил, записанных ниже, извлекались именные конструкции, в которых существительное было омонимично глаголу или прилагательному, однако следующее за ним слово четко указывало на разбор в качестве существительного.

([;prep;] | [!;verb;] | [!;deepr;] | [!;participle;] | [;poss_pron;]) ([!;adv;]?[2:;!adj; gender+, number+, case+, comp=-com])+([3:;noun; gender+, number+, case+ & ;verb;] | [3:;noun; gender+, number+, case+ & ;adj;])[!;prep;]|

^([!;adv;]? [;!adj; gender+, number+, case+, comp=-com])+([;noun; gender+, number+, case+ & ;adj;] | [;noun; gender+, number+, case+ & ;verb;])[!;prep;]

Данная конструкция означает, что если именная группа ограничена справа предлогом, а слева предлогом, глаголом в одной из форм, притяжательным местоимением или началом предложения, то существительное в именной группе может быть омонимично прилагательному или предлогу.

Общий объем полученных сочетаний существительное + прилагательное: 383 млн сочетаний (12,1 млн уникальных сочетаний в начальных формах) для 67000 существительных и 41000 прилагательных. Глагол+предлог+существительное: 349 млн сочетаний (29,2 млн уникальных сочетаний в начальных формах) для 28000 глаголов и 73000 существительных. Извлеченные сочетания выложены на сайте http://cosyco.ru/ .

С одной стороны, разработанное программное обеспечение позволяет быстро извлекать сочетания из новых корпусов. С другой стороны, используемый подход существенно ограничивает полноту извлекаемых конструкций: из 17 млрд словоупотреблений в базу данных попало не более 1,5 млрд, тогда как по самым скромным подсчетам полный синтаксический анализ дал бы как минимум в пять раз большие результаты, хотя их достижимость всё еще остается под вопросом.

Литература

1. *Захаров В. П., Хохлова М. В.* (2014), Автоматическое выявление терминологических словосочетаний // Структурная и прикладная лингвистика. Вып. 10. СПб.: Изд-во С.-Петерб. ун-та, 2014, с. 182–200.

2. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python. O'Reilly Media.
3. Corpus Query Language (CQL) documentation (2017). http://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying
4. *Большакова Е. И., Носков А. А.* (2010), Анализ текста на основе лексико-синтаксичеких шаблонов с сокращением многовариантности // Сборник трудов «Новые информационные технологии в автоматизированных системах-13», с. 62–69.
5. *Клышинский Э. С.* и др. (2015), Исследование неоднозначности употребления слов в европейских языках // Препринты ИПМ им. М. В. Келдыша. № 4. 31 с.
6. *Lukashevich N. Y., Klyshinsky E. S., Kobozeva I. M.* (2016), Lexical Research in Russian: are Modern Corpora Flexible Enough? // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", pp. 427–439.

References

1. Zakharov V. P., *Khohlova M. V.* (2014), Automatic Term Extraction. In: Structural and Applied Linguistics. Vol. 10, pp. 182–200.
2. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python. O'Reilly Media.
3. Corpus Query Language (CQL) documentation (2017). Available at: http://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying
4. *Bolshakova E. I., Noskov A. A.* (2010), Analysis of Texts Using Lexical-Syntactical Templates with Reduction of Multivariance. In: In Proc. of "New Information Technologies in Automated Systems-13", pp. 62–69.
5. *Klyshinsky E. S.* et al. (2015). Analysis of Words' Ambiguity in European Languages. In: Preprints of Keldysh IAM, no. 4. 31 p.
6. *Lukashevich N. Y., Klyshinsky E. S., Kobozeva I. M.* (2016), Lexical Research in Russian: are Modern Corpora Flexible Enough? In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", pp. 427–439.

Клышинский Эдуард Станиславович
Klyshinsky Eduard
E-mail: eklyshinsky@hse.ru

Королев Дмитрий Владиславович
Korolyov Dmitiy
E-mail: dimakorolev94@yandex.ru

Власова Алина Алексеевна
Vlasova Alina
E-mail: ms.aavlasova@mail.ru

Национальный исследовательский университет
«Высшая школа экономики», МИЭМ (Россия)
National Research University "Higher School of Economics», MIEM (Russia)

А. Е. Колесников, Л. А. Малахова
A. Colesnicov, L. Malahov

О РАЗРАБОТКЕ ДИАХРОНИЧЕСКОГО ИСТОРИЧЕСКОГО КОРПУСА РУМЫНСКОГО ЯЗЫКА

ON DEVELOPMENT OF A DIACHRONIC HISTORICAL CORPUS FOR ROMANIAN

Аннотация. В работе рассмотрены технологические аспекты создания корпусов на примере диахронического исторического корпуса румынского языка, охватывающего печатные источники XVI–XIX веков.
Ключевые слова. Румынский язык, диахронический исторический корпус, технология создания корпусов.

Abstract. The paper discusses techniques of corpus development. A diachronic historical corpus for Romanian based on printed sources from the XVI–XIX centuries is taken for the case study.
Keywords. The Romanian language, diachronic historical corpus, techniques of corpus development.

1. Введение

Диахронический корпус RoDia [Colesnicov et al. 2016] отвечает на запрос филологов, изучающих динамику развития румынского языка. Имеется также в виду покрыть географические вариации. Это первый такой корпус для румынского языка. Языков, для которых есть диахронические корпуса, не более двух десятков, например, английский, шведский, итальянский, финский и японский.

2. Источники

В качестве источников использовались сканы старых книг и периодики до конца XIX века включительно. Основной проблемой при работе со старыми источниками является отсутствие регламентированного правописания и вытекающая вариативность письменного выражения явлений языка. Например, для румынского языка таким является слитное или раздельное написание частиц.

Хронология:

1508: начало книгопечатания в Румынии.

1560–1561: первая книга на румынском языке («Четвероевангелие»). Используется румынская кириллица (RC) — 47-символьный алфавит уставного начертания на основе кириллицы.

Конец XVIII — начало XIX веков: упрощённая румынская кириллица (SRC).

1830–1862: постепенный переход от SRC к румынской латинице (RL). Источники насчитывают до 17 вариантов так называемого переходного алфавита (TR) — смеси кириллицы и латиницы.

1862: RL строилась по фонетически-этимологическому принципу, не оправдавшему себя.

1904: современная румынская латиница (MRL); основана на фонетическом принципе, с небольшими изменениями используется до настоящего времени.

Очень важной особенностью всех упомянутых румынских алфавитов является существование однозначного отображения в одну сторону RC → SRC → TR → MRL; RL → MRL. Это позволяет приводить все слова к современному алфавиту, используя его как общее представление письма различных эпох, а также использовать в какой-то мере современные словари и средства обработки естественных языков.

3. Сканирование и распознавание

Была разработана технология распознавания сканов [Burtseva et al. 2016] с помощью программы ABBYY Finereader (AFR).

Пользователю предоставляется пакет, содержащий: алфавиты для AFR; словари (списки слов) для AFR; шаблоны, полученные тренировкой AFR на старых румынских алфавитах; утилита транслитерации кириллицы в латиницу и обратно; виртуальные клавиатуры; шрифт, содержащий редкие глифы старой румынский кириллицы; руководство.

Используются внешние (не включённые в пакет) программы: AFR, Notepad++, Scan Tailor, MS Word & Excel, математическое обеспечение сканера, специальные шрифты.

4. Морфологическая и синтаксическая разметка

После распознавания скана из текста удалялись фрагменты на современном языке (предисловия, комментарии и пр.). С помощью программы Lucon [Mititelu 2016] из оставшегося «старого» текста собиралась конкорданция. Было собрано около 120 тыс. словоформ.

Также выделялись стабильные словосочетания. При этом использовался словарь румынских выражений [Mărănduc 2010].

Далее использовался гибридный POS-таггер университета им. А. И. Кузы в Яссах [Simionescu 2011]. Для его правильной работы необходимы аннотированный лексикон и «золотой» корпус тренировочных текстов. Отсутствие таковых составляет в настоящее время основную трудность.

Метод раскрутки, когда часть текста используется как источник для пополнения лексикона POS-таггера, другая часть для его тренировки, а оставшийся текст обрабатывается POS-таггером перед занесением в корпус, требует очень больших трудозатрат.

Были использованы другие доступные источники, например, некоторые многоязычные параллельные корпуса. Особую роль сыграли некоторые переиздания старых книг, которые сопровождались выполненной экспертами вручную транслитерацией в MRL, а также индексами слов, например, румынское Евангелие 1648 г. из Алба-Юлии. Это позволило заложить базу лексикона «старого» языка для POS-таггера. Также использовался генератор парадигм склонения и спряжения старых слов [Gîfu et al. 2016].

После тщательной проверки результатов работы POS-таггера применялся синтаксический парсер, также разработанный в Яссах.

Одним из вопросов, которые пришлось решать в процессе работы, является выбор стандарта для аннотаций слов в POS-таггере. Был выбран несколько упрощённый вариант (примерно 430 из 600 вариантов аннотации) конвенции MULTEXT-East [Erjavec 2004].

5. Заключение

Создаваемый корпус послужит средством для изучения специалистами-филологами развития румынского языка, а также станет частью общерумынского корпуса.

References

1. *Burtseva L., Colesnicov A., Malahov L., Ciubotaru C., Cojocaru S., Demidov V., Petic M., Bumbu T., Ungur Ş.* (2016) On Technology for Digitization of Romanian Historical Heritage Printed in the Cyrillic Script. In: Proceeding of MFOI-2016 Conference on Mathematical Foundations of Informatics, July 23–30, 2016, Chisinau, Moldova, pp. 144–159.
2. *Colesnicov A., Malahov L., Maranduc C., Perez C-A.* (2016), RoDia: Project of a Regional and Historical Corpus for Romanian. In: Proceedings of MFOI-2016 Conference on Mathematical Foundations of Informatics, July 23–30, 2016, Chisinau, Moldova, pp. 268–284.

3. *Erjavec T.* (2004), MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proceedings of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'2004.
4. *Gîfu D., Simionescu R.* (2016), Tracing Language Variation for Romanian. In: Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2016, 3–9 Apr. 2016, Konya, Turkey.
5. *Mărănduc C.* (2010), Dictionary of Romanian Expressions, Syntagms and Set Phrases (DELS). Corint Publishing, Bucharest, 558 p.
6. *Mititelu C.* (2016), Lucon. Available at: https://sourceforge.net/projects/lucon/ (April, 3, 2017)
7. *Simionescu R.* (2011) Hybrid POS Tagger. Language Resources and Tools in Industrial Applications, Eurolan 2011 summer school.

Колесников Александр Евгеньевич
Малахова Людмила Андреевна
Институт математики и информатики АН Молдовы (Молдова)

Colesnicov Alexandru
Malahov Ludmila
Institute of Mathematics and Computer Science (Moldova)
E-mail: acolesnicov@gmx.com, lmalahov@gmail.com

С. А. Крылов, А. В. Тер-Аванесова
S. A. Krylov, A. V. Ter-Avanesova

ТЕКСТОВЫЕ БАЗЫ ДАННЫХ ПО РУССКИМ ГОВОРАМ[1]
TEXT DATA BASES OF RUSSIAN DIALECTS

Аннотация. В докладе дается краткий обзор корпусов и баз данных на материале русских народных говоров. Дается описание диалектных лексико-грамматических баз данных (микрокорпусов), построенных в формате SRARLING. Представлены результаты исследования языкового варьирования на материале базы данных по говору села Роговатое Старооскольского р-на Белгородской обл.
Ключевые слова. База данных, говоры русского языка, языковое варьирование.

Abstract. In the abstract we give a brief overview of the Russian dialectal corpora and data bases and of the dialectal lexico-grammatica data bases (microcorpora) built with the help of STARLING. The results of the investigation of language variation based on the microcorpus of the dialect spoken in the village of Rogovatoye (Staryy Oskol district, Belgorod region) are presented.
Keywords. Data base, Russian dialects, language variation.

1. Электронные корпуса и базы данных по русским говорам начали создаваться в 2000-е гг. Большинство таких работ из-за трудоемкости пока не доведено до конца; некоторые из них, продолжая пополняться, доступны в интернете: диалектный подкорпус Национального корпуса русского языка (www.ruscorpora.ru); Саратовский диалектный корпус (не общедоступен); электронная библиотека по русским народным говорам, созданная в Казани (www.dialekt.rx5.ru.); электронная картотека Архангельского областного словаря (нет в общем доступе); акустическая база данных «Русские регионы», созданная в университете Бохума (www.rureg.de); корпус говоров бассейна реки Устья (www.parasolcorpus.org/Pushkino); база данных по русской диалектной фонетике С. Л. Николаева и М. Н. Толстой (dialect-phon.ruslang.ru); лексико-грамматические базы данных по трём русским народным говорам (www.starling.rinet.ru) авторов настоящей статьи. Названные опыты ставили перед собой разные задачи и сильно различаются по принципам построения, способам подачи текста (фонетическая транскрип-

[1] Работа написана при поддержке грантов РГНФ №17-04-00485 «Текстовые базы данных по южнорусским говорам» и №17-04-18022 «Диалектологические исследования центра Европейской части России и восточнославянского пограничья», а также № 17-04-00594 «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика» (2017—2019гг.).

ция разной степени подробности или стандартная орфография), по наличию звуковой дорожки. Размер диалектных корпусов и баз данных, как правило, относительно невелик; однако устьянский корпус приближается к 1 млн словоформ, свыше 1 млн словоупотреблений — объём Саратовского корпуса и электронной картотеки Архангельского словаря. Корпусная и шире — компьютерная диалектология решает задачи хранения, обработки и исследования диалектного материала, создавая электронные формы диалектной фонотеки, картотеки, хрестоматии текстов.

2. Диалектные микрокорпуса (ДМК) по русским народным говорам основаны на расшифровке аудиозаписей речи носителей говоров — жителей села Пустоша, деревни Якушевичи Шатурского р-на Московской области и села Новоселки Рыбновского р-на Рязанской области (около 120000 текстоформ; 2005 г., 2012 г.), деревень Арзубиха, Захариха и Злобиха Харовского района Вологодской области (около 20000 текстоформ; 2006 г.), села Роговатое Старооскольского района Белгородской области (более 80000 текстоформ; 2015 г.). Информанты — представители старшего поколения (1910–1950 гг. р.). Эти ДМК построены как морфологически размеченные текстовые базы данных в формате STARLING (*.dbf). Доступные широкому пользователю диалектные материалы, в таком объеме представляющие данные одного говора, достаточно редки. ДМК представляют богатые возможности для системного описания, сопоставительного изучения говоров и изучения вариативности в говорах, обусловленной внутрисистемным развитием, контактными влияниями и влиянием литературного языка.

ДМК обеспечивают многоуровневое членение текста: на сверхфразовые единства, предложения, пунктуационные клаузы, пунктуационные синтагмы, макротакты, мезотакты, микротакты, графические слова (об уровнях членения и единицах см., напр., в: Крылов 2008). Так, в созданном в 2012 г. ДМК (говоры с. Пустоша, Новоселки и д. Якушевичи) представлены 2562 сверхфразовых единства (абзаца), 12918 предложений, 13057 пунктуационных клауз, 29149 пунктуационных синтагм, 69749 макротактов, 70356 мезотактов, 78263 микротакта, 100967 графических слов. Проведена морфологическая разметка баз данных: лексико-грамматический разбор с частичным снятием грамматической и лексико-грамматической омонимии. Словарь лексем содержит около 7000 вокабул. На основе морфологического ана-

лизатора (парсера), базирующегося на подкорпусе пустошенских текстов, созданном в 2005 г., разработана новая улучшенная версия парсера, отличающаяся от версии 2005 г. своей опорой на корпус 2012 г.

3. Основной лингвистической задачей создания перечисленных ДМК, определившей принципы сбора и характер диалектного материала, первоначально было изучение акцентных систем архаичных русских говоров с семью или восемью гласными фонемами (в литературном языке их 5 или 6); распределение фонем /o/ и /ô/ в говорах такого типа определяется характером праславянских тональных акцентов. Поэтому во всех ДМК содержится материал, собранный при помощи вопросников; это обеспечивает сопоставимость данных нескольких говоров и лексическую «представительность» ДМК, включающих целенаправленно собранные в нескольких говорах непроизводные лексемы праславянского словарного фонда. В результате создания ДМК подробно изучены диалектные системы с различением двух фонем «типа *о*».

4. Включение в ДМК образцов «спонтанной» речи носителей говоров дает возможность изучения грамматики говоров и вариативности на всех языковых уровнях. Адекватное описание системы словоизменения в говоре возможно лишь на исчерпывающем материале корпуса текстов — записей «спонтанной речи» носителей говора; применение вопросников неизбежно искажает реальную картину.

Ниже мы покажем, какие возможности предоставляет диалектная текстовая база данных для исследования вариативности форм словоизменения в говоре. В качестве примеров выбраны фрагменты глагольного и именного словоизменения: формы 3 лица наст. времени и формы местного падежа ед. числа имен мужского и среднего рода в слобожанском говоре села Роговатое — те морфологические позиции, в которых происходит вытеснение исконных для говора флексий под влиянием рус. лит. языка; сама же исконная система, очевидно, содержала вариативность показателей, появившуюся под более ранним влиянием восточноукраинских говоров.

Корпус роговатовских текстов представляет собой сплошные расшифровки нескольких аудиозаписей, выполненные в формате WORD в упрощенной фонетической транскрипции и при этом последовательно отражающей фонемные противопоставления и аллофонические чередования, свойственные говору. Аудиозаписи, выполненные в 2010–2014 гг., представляют речь 15 коренных жителей Роговат-

ки старшего поколения (1919–1948 гг. р.). Информанты до 1931 г. р. не имеют регулярного образования, среди них — две не учившиеся в школе женщины, однако одна женщина 1926 г. р. имеет образование 10 классов; все родившиеся в середине 1930-х гг. имеют образование 5 классов; информанты конца 1940-х гг. р. имеют среднее специальное или высшее образование. Все информанты являются носителями местного диалекта; однако идиолекты их несколько различаются как в отношении фонетики, так и грамматики и лексики. За редкими исключениями, речь каждого информанта представлена записями не менее полутора часов звучания.

Система флексий 3 лица настоящего/будущего времени невозвратных глаголов в говоре с. Роговатого отличается стабильностью; у некоторых информантов самого старшего возраста, она встречается практически без отклонений. Для системы флексий 3 лица характерно 1) отсутствие /т'/ *в* формах 3 л. ед. ч. глаголов I спряжения (*несé, мôже*) и 3 л. мн. ч. глаголов II спряжения (*сид'á*) и наличие /т'/ в формах 3 л. мн. ч. I спряжения (*мôгут'*), 3 л. ед. ч. II и «атематического» спряжений (*сидúт', дас'т', йêс'т', йес'т'*), 2) переход в I спряжение глаголов на *-ить, -ать и -êть* (исконно II спр.) с безударными окончаниями: 3 ед. *хôдет*, 3 мн. *хôд'ут'*.

Отклонения от исконной для говора системы, как правило, заключаются в том, что конечный элемент /т'/ присутствует в некотором количестве примеров тех форм 3 лица, которые исконно не имели /т'/. Лишь в одном примере отклонением является форма 3 мн. *пъстрóйа*. Можно поэтому считать, что основная масса отклонений появляется под влиянием системы русского литературного языка, однако влияние это «преломляется» через южнорусскую (или восточноукраинскую) систему флексий 3 лица с мягким конечным /т'/.

Процент инновационных форм 3 л. ед. ч. с конечным /т'/ (*несéт'*) невелик: от 0,7 до 8,5 % у разных информантов; общее число примеров форм л. ед. ч. I спр. у каждого информанта — около 150. Процент инновационных форм 3 л. мн. ч. (*сид'áт'*) обычно значительно выше: 20–47 %, (у одной информантки 1926 г. р. инновационных форм 3 мн. нет), однако и число форм 3 л. мн. ч. II спряжения невелико: от 15 до 25 у разных информантов.

Высокий процент инновационных флексий в форме 3 л. мн. ч. глаголов II спряжения при сравнительно низком — в формах 3 л. ед. ч. в говоре с. Роговатого, расположенного в северной части Слобожанщины, прежде всего объясняется ареальным фактором: восточноу-

краинским говорам свойственны формы 3 л. ед. ч. типа *несé, хóдe — сидúm'* и 3 л. мн. ч. типа *несým', хóдят', стойám'*. Восточноукраинское диалектное влияние на этапе формирования русских слобожанских говоров поддерживало стабильность диалектной флексии 3 ед. I спряжения *-e* и, по-видимому, обусловливало вариативность показателя 3 мн. II спр. *-á/-ám'* до начала интенсивного влияния на эти говоры рус. лит. языка.

Безударная флексия 3 ед. *-е у* глаголов II спряжения — по-видимому, инновация, возникшая на Слобожанщине: эта флексия представлена не только в русских оскольских говорах, а также говорах Северного Подонья, но и украинских слобожанских говорах, а в качестве одного из варьирующих показателей 3 л. ед. ч. глаголов II спряжения встречается в восточноукраинских говорах на обширной территории от Восточного Полесья до Одесской области.

5. Результаты восточноукраинского влияния можно усматривать в формах местн. падежа ед. ч. имен муж. и ср. рода, где у существительных в твердой разновидности склонения варьируют безударные окончания *-е и -у*, а в адъективном и местоименном склонении варьируют безударные окончания *-ъм и -ъму* (*ув автóбусе и ув автóбусу, на этом и на этому*; возможно *на этом/этому автóбусу /автóбусе*). Формы местн. ед. м.-ср. рода, совпадающие с формами дат. ед. м.-ср., которые выступают в конструкциях с предлогами предлогов в *и на*, типичных для местного падежа, мы усматриваем результат украинского влияния на русский говор. Сегодня в речи жителей Роговатого такие формы вытесняются под влиянием рус. лит. языка. У сущ. м. рода они представлены в 7,4 % примеров (всего 189 примеров) и зафиксированы только у лиц 1920-х гг. р. (независимо от уровня образования); у существительных ср. рода — в 21 % примеров (в 5 из 24 примеров), они зафиксированы у лиц 1920 и 1940-х гг. р. У прилагательных безударные окончания местн. п. ед. числа, совпадающие с окончаниями дат., зафиксированы в 4,6 % примеров (всего 43) лишь у старшего поколения говорящих; у местоимений — в 16 % примеров (всего 44 примера) у лиц всех исследованных поколений.

Литература

1. *Крылов С. А.* (2008), О частотном словаре фонетических слов // Фонетика и нефонетика. К 70-летию С. В. Кодзасова. М., с. 387–399.

References

1. *Krylov S. A.* (2008), O chastotnom slovare foneticheskikh slov (About the frequency dictionary of phonetic words). In: Fonetika i nefonetika. K 70-letiyu S. V. Kodzasova. Moscow, pp. 387–399.

Крылов Сергей Александрович
Институт востоковедения РАН (Россия)
Krylov Sergej
Institute of Oriental Studies of the Russian Academy of Sciences
E-mail: krylov-58@mail.ru

Тер-Аванесова Александра Валерьевна
Институт русского языка имени В. В. Виноградова (Россия)
Ter-Avanesova Alexandra
Institute of Russian Language (Russia, Moscow)
E-mail: teravan@mail.ru

С. А. Крылов, О. Е. Фролова
S. A. Krylov, O. E. Frolova

О КОРПУСЕ ОФИЦИАЛЬНО-ДЕЛОВЫХ ТЕКСТОВ РУССКОГО ЯЗЫКА[1]

ABOUT THE CORPUS OF OFFICIAL BUSINESS TEXTS OF RUSSIAN LANGUAGE

Аннотация. В докладе обсуждается структура и содержание корпуса официально-деловых текстов русского языка (ОДКРЯ). Потребность в таком корпусе продиктована несистемной представленностью юридических и законодательных текстов в диахронии и на синхронном уровне в существующих корпусах русского языка: они не позволяют проследить состав юридической терминологической системы и особенности формирования и изменения официально-делового функционального стиля русского языка.
Ключевые слова: корпус официально-деловых текстов русского языка, национальный корпус русского языка, законодательные и юридические тексты в синхронии и диахронии.

Abstract. The report discusses the structure and content of the Russian official-legislative texts corpus (ROLTC). The need for such a corpus is dictated by non-systemic representation of legal and legislative texts in diachrony and at a synchronous level in the existing corpora of the Russian language. They do not allow us to trace the composition of the legal terminology system and the features of the formation and change of the official Russian language of law.
Keywords: Russian official-legislative texts corpus, Russian National Corpus, legislative and legal texts in synchrony and diachrony.

Ещё на заре возникновения «корпусной лингвистики» в нашей стране (когда она ещё не обзавелась этим самоназванием) её фактическим идеологом был академик Андрей Петрович Ершов (1931–1988), один из первопроходцев отечественного программирования. Корпусная русистика отсчитывает своё начало с тех времён, когда А. П. Ершов в 1982 г. предложил создать корпус русской деловой прозы. По его мнению, «Деловая проза… — это языковый носитель производственных отношений человека». Она «всегда внутренне формализована», это «лингвистический феномен, который, сохраняя многие свойства языка в целом, в то же время самой своей сутью подготовлен для того, чтобы стать объектом механизации», благодаря «регламентирующему действию формальной модели, лежащей в основе данной области производственных отношений». ([Ершов 1982]; см. также [Большаков 1985]).

[1] Исследование выполнено при поддержке РГНФ: Проект РГНФ № 17-04-00594 «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика» (2017–2019гг.).

В докладе обсуждается необходимость создания официального и делового корпуса русского языка (ОДКРЯ), включающего не образцы жанров официального дискурса, а конкретные юридические документы («послепетровского» периода — XVIII–XXI в.):

- Свод законов Российской Империи [Свод законов Российской империи];
- Императорские указы и манифесты [Свод законов Российской империи];
- Конституции СССР и РФ [Советское законодательство];
- Постановления ЦК КПСС, Верховного Совета СССР и советского правительства [Советское законодательство];
- современные Уголовные Кодексы (СССР и РФ) и Кодекс Российской Федерации об административных правонарушениях, Административный кодекс СССР [УК ЗФ, КОФП].

Официально-деловому стилю русского языка посвящены работы русистов «докорпусной» эпохи — Е. М. Иссерлин ([Иссерлин 1966]; [Иссерлин 1970]); А. А. Ушакова ([Ушаков 1967]); Б. С. Шварцкопфа ([Шварцкопф 1968]; [Шварцкопф 1998]); А. С. Пиголкина ([Пиголкин 1972]). Он изучался и в диахроническом аспекте: канцелярский стиль XX в. (А. П. Романенко, см. [Романенко 1980]), советской эпохи (К. А. Логинова, см. [Логинова 1968]). Специальные труды посвящены специфике юридических текстов ([Губаева 1983]), законодательных текстов ([Губаева 1987]), вторичных деловых текстов ([Ширинкина 2001]).

Существующие корпуса не представляют официально-деловой функциональный стиль адекватно, поскольку соответствующие тексты представлены в них не систематически.

Это особенно становится видно, когда мы пытаемся проследить время первой фиксации слова или словосочетания — ср. такие конструкции и слова, как «установка на» (1922?) в значениях ‘цель, направленность к чему-н., ориентация на что-н. (нов.); принципы, директивы, руководящее указание (нов.)’ [Ушаков], «лжеколхоз» (1930?), «наплевизм» (1946?), или сочетание «спор хозяйствующих субъектов» (начало XXI в.). А это необходимо для прослеживания «филиации» цитат, лежащей в основе интертекстуальных связей.

*«Ремизов — **установка на** народный язык, народный образ» [Лев Луниц. На запад! (1922). [НКРЯ]*

*«Важным моментом в деле борьбы за коллективизацию сельского хозяйства является борьба с **лжеколхозами**, т. е. борьба с колхозами, идущими не по социалистическому, а капиталистическому пути развития» [Из инструктивного письма Гражданской кассационной коллегии Верховного суда РСФСР о судебной политике по делам, связанным с коллективизацией сельского хозяйства. 6 октября 1930 г. Исторические материалы. http://istmat.info/node/31783].*

*«Советский строй не может терпеть воспитания молодежи в духе безразличия к советской политике, в духе **наплевизма** и безыдейности» [Постановление Оргбюро ЦК ВКП(б) О журналах «Звезда» и «Ленинград» 14 августа 1946 г].*

Есть разрыв между непрофессиональной речью (широко представленной, напр., в НКРЯ) и юридическим терминологическим узусом. При этом лингвистические корпуса позволяют вести поиск по словам и конструкциям, а специальные — по названиям документов. ОДКРЯ, включающий законодательные документы, позволяет наблюдать за актуализацией единицы в специальных и неспециальных текстах. Потребность в таком корпусе продиктована несистемной представленностью юридических и законодательных текстов в существующих корпусах русского языка: они не позволяют проследить состав юридической терминологической системы и особенности формирования и изменения официально-делового русского языка. ОДКРЯ позволит сохранить связь диахронного и синхронного подходов и сможет стать эффективным инструментом исследования стилистики и узуса.

Литература

1. Библиотека нормативно-правовых актов Союза советских социалистических республик URL: http://www.libussr.ru/ (дата обращения: 06.02.2017).
2. *Большаков И. А.* (1985), О некоторых лингвистических особенностях деловой прозы. Семиотика и информатика. Вып. 26. М.
3. *Ершов А. П.* (1982), К методологии построения диалоговых систем: феномен деловой прозы. Вопросы кибернетики. Общение с ЭВМ на естественном языке. М. (Переизд.: Избранные труды. Новосибирск, 1994. с. 314—330).
4. *Иссерлин Е. М.* (1966), Лексика и фразеология современных дипломатических документов. М.
5. Исторические материалы. URL: http://istmat.info/node/31783 (дата обращения: 06.02.2017).
6. Кодекс Российской Федерации об административных правонарушениях. КонсультантПлюс. URL: http://www.consultant.ru/document/cons_doc_LAW_34661/ (дата обращения: 06.02.2017).

7. Национальный корпус русского языка. URL: http://www.ruscorpora.ru/en/index.html (дата обращения: 06.02.2017).
8. Постановление Оргбюро ЦК ВКП(б) О журналах «Звезда» и «Ленинград» 14 августа 1946 г. URL: http://www.hist.msu.ru/ER/Etext/USSR/journal.htm (дата обращения: 06.02.2017).
9. Свод законов Российской империи. URL: http://civil.consultant.ru/code/ (дата обращения: 06.02.2017).
10. Советское законодательство. URL: http://www.prlib.ru/Lib/Pages/authority_1–2–5.aspx (дата обращения: 06.02.2017).
11. Уголовный кодекс Российской Федерации. КонсультантПлюс. URL: http://www.consultant.ru/document/cons_doc_law_10699/ (дата обращения: 06.02.2017).
12. *Ушаков А. А.* (1967), Очерки советской законодательной стилистики. Пермь.
13. *Ушаков Д. Н* (*ред.*) (1935–1940) Толковый словарь русского языка. Т. 1–4. М. URL: http://ushakovdictionary.ru/ (дата обращения: 06.02.2017).
14. *Шварцкопф Б. С.* (1968), Культура деловой речи. Русская речь. № 3.
15. *Шварцкопф Б. С.* (1998), Глава V. Культура деловой речи // Граудина Л. К., Ширяев Е. Н. (отв. ред.). Культура русской речи. Учебник для вузов. М., с. 159–171.

References

1. Biblioteka normativno-pravovykh aktov Soyuza sovetskikh sotsialisticheskikh respublik [Library of normative legal acts of the Union of Soviet Socialist Republics]. Available at: http://www.libussr.ru/ (reference date: 06.02.2017).
2. *Bol'shakov I. A.* (1985), O nekotorykh lingvisticheskikh osobennostyakh delovoy prozy [On some linguistic features of business prose]. In: Semiotika i informatika. Vyp. 26. [Semiotics and Informatics. Issue. 26]. Moscow.
3. *Ershov A. P.* (1994), K metodologii postroenia dialogovyh system: fenomen delovoj prozy [To the methodology of building dialog systems: the phenomenon of business prose]. Moscow. (2nd ed. Selected Works. Novosibirsk, pp. 314—330).
4. *Isserlin Ye. M.* (1966), Leksika i frazeologiya sovremennykh diplomaticheskikh dokumentov [Vocabulary and Phraseology of Modern Diplomatic Documents]. Moscow.
5. Istoricheskiye materialy [Historical materials]. Available at: http://istmat.info/node/31783 (reference date: 06.02.2017).
6. Kodeks Rossiyskoy Federatsii ob administrativnykh pravonarusheniyakh. Konsul'tantPlyus [Code of the Russian Federation on Administrative Offenses]. [ConsultantPlus]. Available at: http://www.consultant.ru/document/cons_doc_LAW_34661 (reference date: 06.02.2017).
7. Natsional'nyy korpus russkogo yazyka. [National Russian Corpus]. Available at: http://www.ruscorpora.ru/en/index.html (reference date: 06.02.2017).
8. Postanovleniye Orgbyuro TSK VKP(b) O zhurnalakh «Zvezda» i «Leningrad» 14 avgusta 1946 [Decree of the Orgburo of the Central Committee of the CPSU (b) On the magazines «Zvezda» and «Leningrad» on August 14, 1946]. Available at: http://www.hist.msu.ru/ER/Etext/USSR/journal.htm (reference date: 06.02. 2017).
9. Svod zakonov Rossiyskoy imperii. [Code of Laws of the Russian Empire]. Available at: http://civil.consultant.ru/code/ (reference date: 06.02.2017).

10. Sovetskoye zakonodatel'stvo [Soviet legislation]. In: Prezidentskaya biblioteka [Presidential Library] Available at: http://www.prlib.ru/Lib/Pages/authority_1–2–5.aspx (reference date: 06.02.2017).
11. Ugolovnyy kodeks Rossiyskoy Federatsii. Konsul'tantPlyus [The Criminal Code of the Russian Federation]. In: Consultant Plus. Available at: http://www.consultant.ru/document/cons_doc_law_10699/ (reference date: 06.02.2017).
12. *Ushakov A. A.* (1967), Ocherki sovetskoy zakonodatel'noy stilistiki [Essays on Soviet legislative stylistics]. Perm'.
13. *Ushakov D. N* (*red.*) (1935–1940), Tolkovyy slovar' russkogo yazyka. [The explanatory dictionary of the Russian language.]. Vol. 1–4. M. Available at: http://ushakovdictionary.ru/ (reference date: 06.02.2017).
14. *Schwarzkopf B. S.* (1968), Shvartskopf B. S. (1968), Kul'tura delovoy rechi. Russkaya rech'. № 3. [Culture of business speech]. In: Russian speech, no. 3.
15. *Shvartskopf B. S.* (1998), Glava V. Kul'tura delovoy rechi [Chapter V. Culture of business speech]. In: Graudina L. K., Shiryayev Ye. N. (otv. red.). Kul'tura russkoy rechi. Uchebnik dlya vuzov [Culture of Russian speech. Textbook for high schools]. Moscow, pp. 159–171.

Крылов Сергей Александрович
Институт востоковедения РАН (Россия)
Krylov Sergej
Institute of Oriental Studies of the Russian Academy of Sciences
E-mail: krylov-58@mail.ru

Фролова Ольга Евгеньевна
МГУ имени М. В. Ломоносова (Россия)
Frolova Olga
Moscow State University (Russia)
E-mail: olga_frolova@list.ru

О. Ю. Крючкова, В. Е. Гольдин
O. Ju. Kryuchkova, V. E. Goldin

ДИАЛЕКТНЫЙ ТЕКСТОВЫЙ КОРПУС: ПРОБЛЕМЫ РЕПРЕЗЕНТАТИВНОСТИ, СБАЛАНСИРОВАННОСТИ, ЕДИНИЦ ХРАНЕНИЯ И ВЫДАЧИ

DIALECT TEXTUAL CORPUS: PROBLEMS OF REPRESENTATIVENESS, BALANCE, UNITS OF STORAGE AND DELIVERY

Аннотация. В докладе обосновываются следующие положения: система корпусов отдельных говоров и создание их коммуникативных моделей – путь к достижению репрезентативности диалектного корпуса; единицей хранения в диалектном корпусе является речевой фрагмент, выделенный по формальному критерию соответствия непрерывному участку говорения; единицей выдачи является абзац или целый текст.
Ключевые слова. Корпус, диалект, текст, репрезентативность.

Abstract. The report substantiates the following theses: the system of individual dialects corpora is the way to reach the representativeness of a dialect corpus; the dialect corpus's unit of storage is a speech fragment, picked out according to a formal criterion of correspondence to an uninterrupted parcel of speaking; the unit of delivery is a paragraph or a whole text.
Keywords. Corpus, dialect, text, representativeness.

1. Введение

Вопрос о критериях репрезентативности (достаточного объема) и сбалансированности (коммуникативной пропорциональности) корпуса является основополагающим для любого текстового корпуса. В идеале текстовый корпус — это модель языка или подъязыка в его функционировании. Характер единиц хранения в текстовых корпусах и типов выдачи определяется целями корпуса и спецификой его материала.

Решение названных вопросов имеет особую сложность при создании текстовых корпусов диалектной речи, представленной, во-первых, в многообразии своих относительно самостоятельных территориальных вариантов и, во-вторых, существующей в принципиально устной форме, в виде спонтанных диалогических и монологических речевых фрагментов.

В настоящее время апробируются разные модели построения диалектных корпусов, в которых практикуются различные подходы к обозначенным вопросам.

2. Репрезентативность и сбалансированность диалектного корпуса

Условием репрезентативности текстового диалектного корпуса (как и любого текстового корпуса) является представление в нем реально функционирующей коммуникативной системы. Что является реально функционирующими коммуникативными системами, если речь идет о диалектах? Это, безусловно, конкретные говоры, «диалектные микросистемы» со своеобразной внутренней организацией. В связи с этим репрезентативность диалектного корпуса по отношению ко всему «русскому диалектному языку» может быть достигнута только через систему корпусов отдельных говоров, представляющих важнейшие диалектные типы русской речи (наречия, группы говоров, диалектные зоны). При этом текстовая база корпуса каждого отдельного говора в соответствии с принципом пропорциональности должна стремиться к моделированию коммуникации в данном говоре, отражая важнейшие типы и формы диалектной речи, социальную дифференциацию носителей говора, жанрово-тематическую структуру диалектного общения.

Данные принципы лежат в основе создания Саратовского диалектологического корпуса (СарДК) [Крючкова, Гольдин 2008], в котором разрабатываются в настоящее время подкорпусы отдельных говоров разных типов (в настоящее время это подкорпус севернорусского говора с. Мегра Вытегорского района Вологодской области, подкорпус среднерусского окающего говора с. Белогорное Вольского района Саратовской области, подкорпус среднерусского акающего говора с. Земляные Хутора Аткарского района саратовской области). Иной принцип наполнения диалектного текстового корпуса применяется в диалектном подкорпусе Национального корпуса русского языка, в состав которого включаются текстовые фрагменты разных говоров, объединяемые на основе территориально-административного («регион записи»: Архангельская область, Брянская область, Владимирская область и т. д.), а не собственного диалектного членения. Корпус, построенный на таких основаниях, носит иллюстративный характер, знакомит пользователя с территориальной неоднородностью русского национального языка, но не обладает признаками репрезентативности в качестве источника изучения русской диалектной речи, репрезентированной существенно различающимися полнофункциональными языковыми микросистемами.

3. Единицы хранения и выдачи в диалектном корпусе

Членение речевого потока в диалектном корпусе должно, на наш взгляд, отвечать принципу максимального приближения создаваемой в корпусе модели к объекту — естественной коммуникации на диалекте. Текст как единицу хранения в диалектном корпусе целесообразно выделять по формальному критерию соответствия единовременной аудио- или видеозаписи непрерывному участку говорения (диалогического или монологического типа). Любая намеренная сегментация речевого потока (по жанровому, тематическому или иному критерию) существенно снижает эвристическую ценность корпусного материала. Необходимость хранения в базе корпуса целостного, специально не препарированного речевого фрагмента обусловлена существенными отличиями диалектных текстов от текстов «стандартного языка» [Крючкова, Гольдин 2011], вследствие чего диалектный корпус должен быть одновременно и диалектной библиотекой (или архивным собранием материалов). Такое представление диалектной речи в корпусе не только сохраняет невоспроизводимый уникальный речевой материал, но и существенно расширяет возможности использования корпуса, дает материал не только для структурного анализа диалектной речи, но и для ее изучения в коммуникативном аспекте (позволяет, например, изучать жанрово-тематическую структуру диалектной коммуникации, ее идиолектную, лингво-культурную специфику). Это положение, по-видимому, актуально не только для диалектного корпуса, но также для корпусов спонтанной устной речи в целом.

Специфика диалектного материала требует контекстов большей протяженности, чем в корпусе «стандартных» текстов, и возможности получения целого текста. Именно этими положениями определяются параметры выдачи по запросу в СарДК. В СарДК минимальной выдачей является абзац, т. е. структурно-семантическое целое, а максимальной — целый текст как политематическое и полижанровое единство обычно значительной протяженности. Возможности выдачи текстовых фрагментов по тематическому или жанровому критериям (создание пользователем тематических или жанровых подкорпусов) обеспечивается реализуемой в СарДК тематической и жанровой разметкой каждого выделенного на формальной основе диалектного текста.

4. Заключение

Репрезентативность диалектного текстового корпуса связана не только с решением вопросов оптимального для целей корпуса отбора текстового материала, форматов его хранения и выдачи. Большое значение для диалектного корпуса, создаваемого в качестве полноценного научно-образовательного ресурса, имеют также стандарты, регламентирующие способы расшифровки записей диалектной речи, аннотирования его текстовой базы [Крючкова, Гольдин 2015], проблема мультимедийности диалектного корпуса, его справочной и культурологической составляющих.

Литература

1. *Крючкова О. Ю., Гольдин В. Е.* (2008), Текстовый диалектологический корпус как модель традиционной сельской коммуникации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008», с. 268–273.
2. *Крючкова О. Ю., Гольдин В. Е.* (2011), Корпус русской диалектной речи: концепция и параметры оценки // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2011», с. 359–367.
3. *Крючкова О. Ю., Гольдин В. Е.* (2015), Параметры обработки текстов для русского диалектного корпуса // Труды международной конференции «Корпусная лингвистика — 2015», с. 307–314.

References

4. *Kryuchkova O. Ju., Goldin V. E.* (2008), Tekstovyj dialektologicheskij korpus kak model' tradicionnoj sel'skoj kommunikacii [Textual dialect corpus as a model of traditional rural communication]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii "Dialog–2008" [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2008"], pp. 268–273.
5. *Kryuchkova O. Ju., Goldin V. E.* (2011), Korpus russkoj dialektnoj rechi: koncepcija i parametry ocenki [Corpus of Russian dialect speech: concept and parameters of evaluation]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii "Dialog–2011" [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2011"], pp. 359–367.
6. *Kryuchkova O. Ju., Goldin V. E.* (2015), Parametry obrabotki tekstov dlja russkogo dialektnogo korpusa [The parameters of text processing for the Russian dialect corpus]. In: Trudy mezhdunarodnoj konferencii "Korpusnaja lingvistika — 2015" [Proceedings of the international conference "Corpus linguistics — 2015"], pp. 307–314.

Крючкова Ольга Юрьевна
Kryuchkova Olga
E-mail: vpks@rambler.ru

Гольдин Валентин Евсеевич
Goldin Valentin
E-mail: goldinve@yandex.ru

Саратовский государственный университет им. Н. Г. Чернышевского (Россия)
Saratov State University (Russia)

М. А. Кустова
M. A. Kustova

АВТОМАТИЧЕСКАЯ РАЗРАБОТКА УЧЕБНЫХ ТЕСТОВ ПО АНГЛИЙСКОМУ ЯЗЫКУ НА ОСНОВЕ КОРПУСА[1]

DESIGN OF CORPUS-GENERATED EFL PLACEMENT AND PROGRESS TESTS FOR UNIVERSITY STUDENTS

Аннотация. В работе представлен инструмент, позволяющий автоматически создавать вопросы для тестов по английскому языку на корпусной основе. В качестве источника для вопросов используется REALEC (Russian Error-Annotated Learner English Corpus) – корпус работ студентов ВШЭ, изучающих английский язык как иностранный. Все работы в корпусе проверены преподавателями английского ВШЭ, для каждой ошибки ими указан её тип, предполагаемый источник, степень её влияния на понимание текста, а также предложен вариант исправления. На основе этих данных автоматически генерируется тест, состоящий из предложений с ошибками, которые студенту необходимо исправить. Рабочее название инструмента – RETM – REALEC English Test Maker.
Ключевые слова. Корпусная лингвистика, учебный корпус, английский язык как иностранный.

Abstract. The paper introduces a tool which automatically creates questions for EFL placement and progress tests on the basis of a learner corpus. The learner corpus in question is called REALEC (Russian Error-Annotated Learner English Corpus). REALEC is a corpus of HSE students' papers written in English. All the papers in the corpus are graded by English teachers at HSE, who highlight all the mistakes there. Every mistake is annotated by its type, supposed source, and how strong it affects the meaning of the whole sentence. Also, for each mistake the expert provides a variant of correction. On the basis of all these data a test is generated, which consists of faulty sentences that a student needs to correct. The working title of the instrument is RETM – REALEC English Test Maker.
Keywords.

1. Введение

Создание вступительных и контрольных тестов по английскому языку как иностранному сопряжено с определёнными трудностями. При составлении вопросов необходимо учитывать особенности программы учебной дисциплины, а также выбрать такой тип вопроса, который бы должным образом задействовал языковые навыки студента. Это делает составление теста сложным и времязатратным процессом.

В настоящей работе представлен инструмент, автоматически выполняющий часть работы по составлению проверочных и вступительных тестов по английскому языку. Программа генерирует вопросы на исправление ошибок в предложениях, взятых из студенческих

[1] https://ling.hse.ru/realec/

письменных работ. В качестве источника предложений для вопросов программа использует REALEC (Russian Error-Annotated Learner English Corpus) — корпус англоязычных письменных работ студентов ВШЭ. Рабочее название инструмента — RETM (REALEC English Test Maker).

2. Принцип работы

2.1. Brat и схема аннотации REALEC

Прежде чем описывать принципы работы самой RETM, представляется необходимым дать краткое описание корпуса, используемого в качестве источника для вопросов. В этом разделе будет кратко описано внутреннее устройство, а также схема аннотации корпуса REALEC.

Для разметки ошибок в REALEC используется brat rapid annotation tool — онлайн-инструмент для аннотирования текста. Формат brat позволяет определить собственную схему аннотации для корпуса. Для каждого документа, загруженного и размеченного в системе, аннотация хранится в отдельном текстовом файле.

Все работы, загруженные в REALEC, проверяются экспертами, в роли которых обычно выступают преподаватели английского языка в НИУ ВШЭ. Эксперты выделяют в работах ошибки и в соответствии со схемой аннотации корпуса размечают их на трёх уровнях:

- тип ошибки (*wrong choice of article*, *lexical choice* и т. д.; всего более 150 типов);
- предполагаемая причина ошибки (напр. влияние родного языка);
- уровень влияния ошибки на понимание всего текста (небольшой, средний или критический).

Кроме того, для каждой ошибки эксперты предлагают вариант исправления.

2.2. Создание тестовых вопросов

Для создания теста при помощи RETM необходимо прежде всего скачать с сайта REALEC некоторое количество работ и аннотаций к ним. Операции, описанные ниже, RETM проводит на паре файлов «текст-аннотация».

Схема аннотации REALEC включает иерархически выстроенное множество классов и подклассов ошибок. Как правило, для теста не-

обходимы лишь некоторые из них. Преподаватель сообщает RETM типы ошибок, согласующиеся с желаемым типом теста. После этого программа фильтрует все аннотации, содержащиеся в файле аннотаций, и оставляет только те, которые соответствуют выбранным типам.

На следующем шаге программа разбивает исходный текст на предложения. Для каждого из предложений выясняется, содержит ли оно ошибку одного из выбранных типов. Если предложение содержит хотя бы одну такую ошибку, оно послужит материалом для тестового вопроса.

Вопросы генерируются следующим образом. Сначала программа исправляет все ошибки, кроме той, которая принадлежит одному из выбранных преподавателем типов и которую впоследствии должен будет исправить студент. Получившееся предложение, в котором исправлены все ошибки, кроме одной — то, как впоследствии будет выглядеть вопрос теста. В качестве ожидаемого ответа на этот вопрос используется исправление этой ошибки, предложенное экспертом при проверке. В случае, если предложение содержит более одной ошибки, относящейся к выбранным типам, отдельный вопрос по описанному выше алгоритму создаётся для каждой из них.

После того как вышеописанные действия выполнены для всех предложений во всех работах, переданных RETM в качестве источника для тестов, получившиеся вопросы перекодируются в формат XML. Формат файла задан платформой Moodle, на которой работает веб-сайт, используемый для тестирования студентов.

Важно отметить, что RETM не привязана к схеме разметки, принятой в REALEC. Это значит, что теоретически она после минимальных доработок может быть использована для создания тестов на основе других корпусов, использующих brat.

2.3. Оценка преподавателя

По окончании работы RETM преподаватель получает некоторый пул вопросов; в зависимости от количества текстов, поданных на вход, и выбранных типов ошибок, количество вопросов может доходить до нескольких тысяч. Поскольку вопросы были сгенерированы автоматически, некоторые из них могут оказаться неподходящими для реального теста. На этом этапе преподавателю необходимо вручную просмотреть все получившиеся вопросы и убрать из пула те из них, которые не подходят для теста.

Помимо фильтрации вопросов, преподавателю необходимо также приписать каждому из них уровень сложности: низкий, средний или высокий. Это нужно для того, чтобы вопросы в пределах одного теста варьировались по сложности: за каждым верным ответом следовал вопрос более высокого (или равного, если самый высокий уже достигнут) уровня сложности, и наоборот при неверном ответе. Количество баллов за правильный ответ прямо пропорционально уровню сложности вопроса.

3. Преимущества и недостатки RETM

Тесты по английскому языку должны удовлетворять ряду требований. Так, важно, чтобы проверочный тест содержал вопросы, которые бы позволяли студенту отработать в том числе наиболее слабые языковые навыки. В то же время необходимо, чтобы тест соответствовал предполагаемому уровню его владения языком.

Тесты, созданные при помощи RETM, удовлетворяют описанным требованиям. Поскольку источником для вопросов являются ошибки, допущенные такими же студентами, можно с определённой долей уверенности утверждать, что эти ошибки являются типичными. Иначе говоря, такие вопросы направлены именно на отработку наиболее слабых языковых компетенций студентов.

Ещё одно следствие из выбора источника для вопросов заключается в том, что уровень языка в вопросах примерно соответствует навыкам владения английским у студентов, проходящих тест. На это направлена и система выставления уровней сложности — тест как бы дополнительно подстраивается под уровень студента, выставляя ему соответствующие баллы.

Нельзя не отметить, однако, некоторую ограниченность применения RETM. Так, на данный момент инструмент позволяет создавать лишь вопросы на исправление ошибок, где ответ представляет собой короткую строку, которую студент должен самостоятельно ввести с клавиатуры. Такие вопросы зачастую не учитывают всех возможных вариантов ответа, что приводит к возможности несправедливого оценивания теста. Кроме того, объёмные тесты, состоящие исключительно из вопросов одного типа, могут быть утомительными для студентов. На данный момент в разработке находится инструмент, который позволит на основе тех же источников создавать вопросы на выбор верного варианта из списка.

Важно также, что ни RETM, ни разметка в REALEC пока не являются безупречными, а потому для составления финального варианта теста всё равно требуются усилия эксперта. Тем не менее, этот процесс всё равно требует меньше времени и сил, чем создание теста с нуля. Поэтому RETM позиционируется не как полноценная замена живым составителям тестов, но как инструмент, позволяющий облегчить их работу.

References

1. *Bachman, L. F.* (1991), What does language testing have to offer? In: Tesol Quarterly, 25(4).
2. *Granger, S.* (2007), Learner Corpora in Foreign Language Education. In: Encyclopedia of Language and Education. Volume 4. Second and Foreign Language Education, Springer.
3. *Hovy, E.* (2015), Corpus Annotation. The Oxford Handbook of Computational Linguistics 2nd edition, Oxford University Press.
4. Available at: http://realec.org/about/ (1.03.2017)
5. Available at: http://brat.nlplab.org/ (1.03.2017)

Кустова Марина Александровна

Национальный исследовательский университет «Высшая школа экономики» (Россия)

Kustova Marina

National Research University «Higher School of Economics» (Russia)

E-mail: marinakoustova@gmail.com

Г.Я.Мартыненко
G.Y.Martynenko

ЭМПИРИЧЕСКОЕ КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ КАК МЕТОД РАЗЫСКАНИЯ ГРАНИЦЫ МЕЖДУ ЯДРОМ И ПЕРИФЕРИЕЙ В ЧАСТОТНЫХ СЛОВАРЯХ

EMPIRICAL CORRELATION RATIO AS A TOOL FOR IDENTIFYING THE BOUNDARY BETWEEN THE CORE AND THE PERIPHERY IN FREQUENCY DICTIONARIES

Аннотация. Предложен способ определения границы между ядром и периферией в частотных словарях, основанный на последовательном измерении эмпирического корреляционного отношения через определенный интервал. Абсцисса максимума этого отношения принимается за границу между ядром и периферией.
Ключевые слова. Математическая лингвистика, квантитатативная лингвистика, частотный словарь, ядро, периферия, эмпирическое корреляционное отношение, метод скольжения.

Abstract. The paper proposes a method for identifying the boundary between the core and the periphery in frequency dictionaries based on a step-by-step calculation of the empirical correlation ratio over a certain period. The abscissa of the maximum of this ratio is taken as the boundary between the core and the periphery.
Keywords. Mathematical linguistics, quantitative linguistics, frequency dictionary, core, periphery, empirical correlation ratio, method of moving correlation ratio.

1. Введение

В современной науке понятия ядра и периферии занимают важное место. В сущности это междисциплинарные понятия. Они активно используется в социологии, психологии, политологии, экономике, географии и др. науках. В лингвистике эти понятия рассматриваются как одна из универсалий языка, как один из вариантов проявления асимметрии его знаковой системы. Чаще всего оба понятия связываются со множествами лингвистических (фонетических, грамматических, лексических и др.) единиц. При этом в ядро включаются наиболее типичные, употребительные формы, а в периферию включают второстепенные, окказиональные элементы. Но вместе ядро и периферия образуют единую систему в определенном соотношении. Это соотношение всегда асимметрично в пользу периферии.

Противопоставление ядра и периферии характерно не только для лингвистики, но и для литературоведения. Впервые его использовал

Ю. Н. Тынянов в своей книге «Архаисты и новаторы» в 20-е гг. XX в. [Тынянов 1929] в рамках своей теории литературно-художественных систем, среди которых он выделял синхронические и диахронические системы. К синхроническим он относил множества авторов (и их произведений), пишущих в данную литературную эпоху. При этом в каждой синхронической системе Тынянов выделял ядро (множество типичных для данной эпохи писателей) и периферию, в которую включались авторы второстепенные, малозначительные.

Позднее возникли новые интерпретации ядра и периферии: уже в рамках теории сообществ, теории ценозов, в теории статистической совокупности, а позднее — в теории систем [Мартыненко 2009] с отраслевыми реализациями в биологии, науковедении, технетике, информатике и др. дисциплинах. Стали исследоваться неоднородные совокупности собирательного типа самой произвольной природы. При этом возникла задача разделения таких совокупностей на ядро и периферию, в том числе с помощью различного рода математических методов [Горькова 1969; Мартыненко 1988 и др.]. В центре нашего внимания будет один из таких методов: метод максимизации скользящего эмпирического корреляционного отношения.

2. Метод максимума эмпирического корреляционного отношения

Известно, что в неоднородной совокупности дисперсия согласно правилу сложения дисперсий распадается на две части: среднюю групповых дисперсий $\overline{\sigma}^2$ и межгрупповую дисперсию (дисперсию групповых средних около общей средней) δ^2, т. е. $\sigma^2 = \delta^2 + \overline{\sigma}^2$.

Корень квадратный из отношения межгрупповой дисперсии к общей дисперсии называется эмпирическим корреляционным отношением:

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}}\ .$$

Этот показатель характеризует неоднородность совокупности при том или ином способе ее расслоения. Чем ближе η^2 к единице, тем совокупность неоднородней. Воспользуемся этим полезным с точки зрения решения нашей задачи свойством. Будем искать максимум этого показателя на множестве членений исходного распределения с использованием приема скольжении. При движении вниз по спектровому распределению на каждом шаге скольжения будем вычислять значение эмпирического корреляционного отношения между верхней

и нижней частями таблицы. Максимальное значение этого показателя будем считать оптимальной границей между двумя фрагментами неоднородной совокупности.

Приступая к анализу эмпирического материала, следует иметь в виду, что среди совокупностей, привлеченных нами в качестве иллюстраций, следует различать малые совокупности, содержащие не более нескольких сотен разноименных элементов, и большие совокупности, состоящие из сотен тысяч элементов.

Начнем изложения с «коротких» статистических рядов.

Ниже приведены расчетные данные для распределения поэтов по числу написанных на их стихи романсов [Мартыненко; 2006] — см. табл. 1. Соответствующий график приведен на рис. 1, а на рис. 2 приведен еще один аналогичный график.

Таблица 1. Скользящее **корреляционное отношение для распределения поэтов по числу написанных на их стихи романсов**

Число романсов	Число поэтов	Корреляционное отношение η	Число романсов	Число поэтов	Корреляционное отношение
1	119	0,539	11	1	0,876
2	27	0,636	12	1	0,875
3	15	0,766	13	2	0,851
4	5	0,800	16	1	0,828
5	6	0,839	17	2	0,774
6	2	0.850	19	1	0,735
7	4	0,870	22	3	0,539
8	2	0,876	24	1	0,431
9	1	0,879	31	1	
10	2	0,878			

Весьма интересен тот факт, что абсцисса максимума эмпирического корреляционного отношения попадает на интервал, в котором находится локальный минимум множества спектровых распределений. Эта согласованность, на наш взгляд, является дополнительным свидетельством в пользу того, что исследованные и им подобные, но пока не исследованные распределения качественно неоднородны. Ордината максимума варьирует в нашем материале варьирует от 0,7 до 0,9, что

говорит о довольно сильной дифференциации между ядром и периферией.

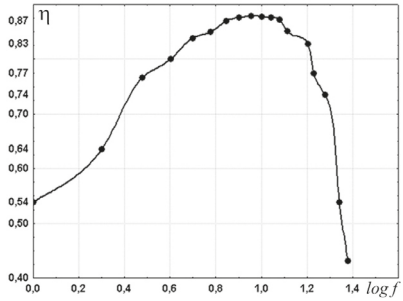

Рис. 1. **График скользящего эмпирического корреляционного отношения η для распределения поэтов по логарифму числа написанных на их стихи романсов (*log f*)**

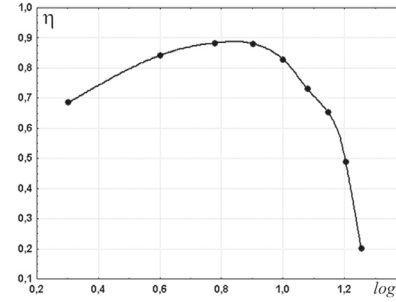

Рис. 2. **Зависимость η для распределения персонажей по логарифму числа их упоминаний в романе Михаила Булгакова «Мастер и Маргарита» (*log f*)**

Теперь рассмотрим «длинные» статистические ряды. Таковыми являются, например канонические частотные словари, насчитывающие десятки и сотни тысяч разноименных элементов.

3. Эмпирическое корреляционное отношение в больших частотных словарях

Что касается частотных словарей канонического типа, то траектория эмпирического корреляционного отношения здесь имеет принципиально иной характер. При построении данной зависимости воспользуемся данными: [Частотный словарь А. П. Чехова 1999] и [Частотный словарь Л. Н. Андреева 2003], а также материалами частотного словаря Ветхого Завета [Алексеев 2004]. Соответствующие графики показаны на рис. 3–5.

График скользящего корреляционного отношения в рассказах А. П. Чехова (рис. 4) полностью идентичен графикам, приведенным выше (рис. 1–2). Этого нельзя сказать об аналогичной зависимости для частотных словарей Л. Н. Андреева и Ветхого Завета (рис. 5–6). В двух последних случаях в кривой отсутствует максимум. При этом она довольно откровенно распадается на две части: левую — до точки *A* и правую — после точки *A*. Причина такого поведения кривой отчасти заключается в том, что частота самого частого слова настолько

велика, что она переводит траекторию кривой с убывающего хода на возрастающий. Значительный вклад в такое поведение кривой вносят и другие частые элементы. Такое их поведение находит отражение в величине дисперсии. Так, в частотном словаре Ветхого Завета при исключении только одного самого частого слова (союза «и») дисперсия падает до 0,379 от реально наблюдаемой, при исключении 10 самых частых слов — до 0,159, а при 50 — уже до 0,025 от наблюдаемой. Это означает, что в частотных словарях строевая лексика вносит решающий вклад в общую вариацию, например дисперсию. Поэтому следует, по-видимому, говорить не о бинарном членении совокупности, а о трех зонах словаря: зоне строевых элементов, зоне высокоактивных семантически полных элементов с широкой семантикой и зоне редких элементов с узкой семантикой. Этот вопрос заслуживает специального и пространного обсуждения, поэтому мы его выносим за рамки данной статьи.

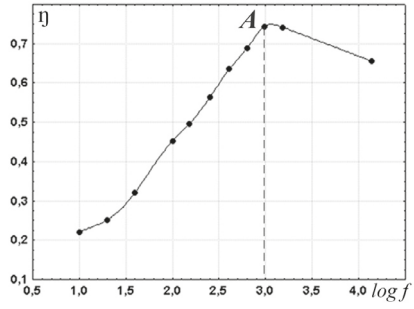

Рис. 3. **График скользящего ЭКО η для спектрового распределения лексических единиц в рассказах А. П. Чехова**

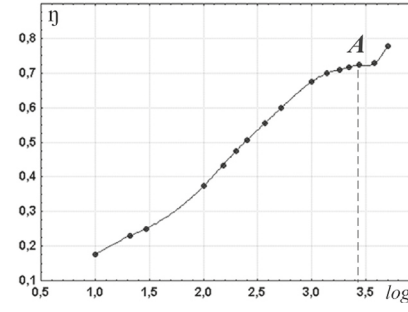

Рис. 4. **График ЭКО η для спектрового распределения лексических единиц в рассказах Л. Н. Андреева**

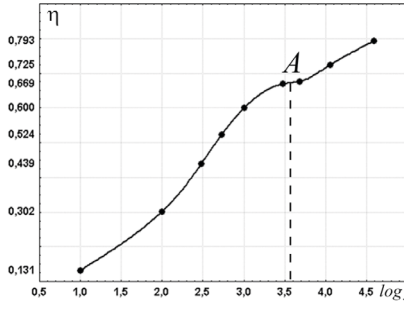

Рис. 5. **График скользящего ЭКО η для спектрового распределения лексических единиц в частотном словаре Ветхого Завета [Алексеев 2004]**

4. Заключение

Предложен метод определения границы между ядром и периферией в частотных словарях разнообразной природы, основанный на пошаговом перемещении границы между ядерным и периферийными подмножествами и измерении на каждом шаге эмпирического корреляционного отношения. Абсцисса максимума этого отношения принимается за границу между ядром и периферией. Особый интерес представляют большие словари, содержащие сотни тысяч и более словоупотреблений. В некоторых из этих словарей присутствует недосформированный максимум. Это вызвано тем, что в них могут доминировать слова с патологической, резко выделяющейся частотой, которые образуют собственное подмножество. Таким словом является союз «и», формирующий жанровую картину в религиозных текстах и художественных текстах, напоминающих библейские (Л. Н. Андреев).

Литература

1. *Алексеев П. М.* (2004), Частотный словарь Ветхого Завета // Структурная и прикладная лингвистика. Вып. 6. СПб., с. 223–237.
2. *Горькова В. И.* (1969), Ранговое распределение на множествах научно-технической информации // Научно-техническая информация. Сер. 2, 1969, № 5, с. 5–11.
3. *Мартыненко Г. Я.* (1988), Основы стилеметрии. Л.: ЛГУ.
4. *Мартыненко Г. Я.* (2006), Семантика корпуса русского романса // Материалы Международной конференции «Корпусная лингвистика 2006». СПб., с. 11–15.
5. *Мартыненко Г. Я.* (2009), Введение в теорию числовой гармонии текста. СПб.: СПбГУ.
6. *Тынянов Ю. Н.* (1929), Архаисты и новаторы Л.: Прибой.
7. *Частотный словарь рассказов А. П. Чехова* (1999), Под ред. Г. Я. Мартыненко. Автор-составитель А. О. Гребенников. СПбГУ.
8. *Частотный словарь Л. Н. Андреева* (2003), Под ред. Г. Я. Мартыненко. Автор-составитель А. О. Гребенников. СПбГУ.

References

1. *Alekseev P. M.* (2004), Chastotnyj slovar' Vetkhogo Zaveta [The frequency dictionary of the Old Testament]. In: Strukturnaja i prikladnaja lingvistika [Structural and applied linguistics], Iss. 6. SPb, pp. 223–237.
2. *Gor'kova V. I.* (1969), Rangovoe raspredelenie na mnozhestvakh nauchno-tekhnicheskoj informacii [Rank distribution on the sets of scientific and technical information]. In: Nauchno-tehnicheskaja informacija [Scientific and technical information]. Ser. 2, 1969, no. 5, pp. 5–11.

3. *Martynenko G. Ja.* (1988), Osnovy stilemetrii [The Foundations of Stylemetrics]. Leningrad, Leningrad State University.
4. *Martynenko G. Ja.* (2006), Semantika korpusa russkogo romansa [Semantics of the Corpus of Russian Romances]. In: Materialy Mezhdunarodnoj konferencii "Korpusnaja lingvistika 2006" [Proc. of the Int. Conf. "Corpus Linguistics 2006"]. St Petersburg, pp. 11–15.
5. *Martynenko G. Ja.* (2009), Vvedenie v teoriju chislovoj garmonii teksta [Introduction to the theory of numerical harmony of text]. St Petersburg, St Petersburg State University.
6. *Tynjanov Ju. N.* (1929), Arkhaisty i novatory [Archaists and innovators]. Leningrad, Priboj.
7. *Chastotnyj slovar' rasskazov A. P. Chekhova* [Frequency Dictionary of short stories by Anton Chekhov] (1999), Martynenko G. Ya (ed.), Grebennikov A. O. (compiler), St. Petersburg, St. Petersburg State University.
8. *Chastotnyj slovar' L. N. Andreeva* [Frequency Dictionary of stories by Leonid Andreev] (2003) Martynenko G. Ya (ed.), Grebennikov A. O. (compiler), St Petersburg St Petersburg State University.

Мартыненко Григорий Яковлевич
Санкт-Петербургский государственный университет (Россия)
Gregory Martynenko
St Petersburg State University (Russia)
E-mail: g.martynenko@gmail.com

А. Ц. Масевич, В. П. Захаров
A. Ts. Masevich, V. P. Zakharov

ДИАХРОНИЧЕСКОЕ ИССЛЕДОВАНИЕ ЛЕКСИКО-СЕМАНТИЧЕСКОГО ПОЛЯ «ВРАГИ»

DIACHRONIC RESEARCH OF THE SEMANTIC FIELD «ENEMIES»

Аннотация. В статье представлены результаты диахронического исследования семантического поля с доминирующим словом «враги». Исследование проводилось на базе русского корпуса Google Books с использованием программного инструмента Google Books N-gram Viewer. Описаны две графические модели частотного поведения слов и словосочетаний и показана их корреляция с историческим процессом.
Ключевые слова. Корпусы текстов, диахронические исследования, частотность лексических единиц, модели частотного поведения, семантическое поле.

Abstract. The paper presents the results of a diachronic research of the semantic field with "enemies" as dominating word, based upon the Russian corpus of the Google Books using the Google Books N-gram Viewer. Two graphic models of the words' behaviour and their correlation with the historical process are described.
Keywords. Corpora of texts, diachronic studies, frequency of lexial units, models of words behavior, semantic field.

Введение

Частота употребляемости лексических единиц меняется в ходе исторического процесса, и её изменения могут оказаться весьма ценными количественными индикаторами, полезными для исторических и культурологических исследований. Наличие представительных корпусов, оснащенных соответствующим программным инструментарием, позволяет получить такие показатели.

Цель нашей работы — кратко описать функционирование лексико-семантического поля «враги» на протяжении длительного промежутка времени.

Исследование проводилось на основе русского корпуса Google Books (более 67 млрд токенов) и системы Google Ngram Viewer [Michel et al. 2011], детально описанной нами в ряде публикаций (напр., [Захаров, Масевич 2014]). Она позволяет строить графики встречаемости слов и коллокаций за выбранный временной период.

Используемый инструментарий позволил выявить наиболее употребительные грамматические формы — чаще всего это родительный падеж единственного и множественного числа,- на основе которых и строятся графики. Система Google Ngram Viewer позволяет с по-

мощью подстановочных знаков отобрать наиболее частотные коллокации с данной словоформой, как в правой, так и в левой позиции. При этом существует возможность задавать части речи другого слова коллокации.

1. Коллокации со словом «враг»

Частота встречаемости биграммы «классового врага» достигает наиболее высоких значений в середине 1930-х гг., затем отмечается её снижение и некоторый рост в начале 1950-х гг.

«Классовый враг» — распространённое идеологическое клише, имеющее очевидную связь с репрессиями. Поэтому частота употребления этой биграммы возрастает во время сопровождающих репрессии пропагандистских кампаний. Во время войны 1941–1945 г. употребление её резко снижается, зато значительно растёт частота других коллокаций со словоформой «врага» (рис. 1).

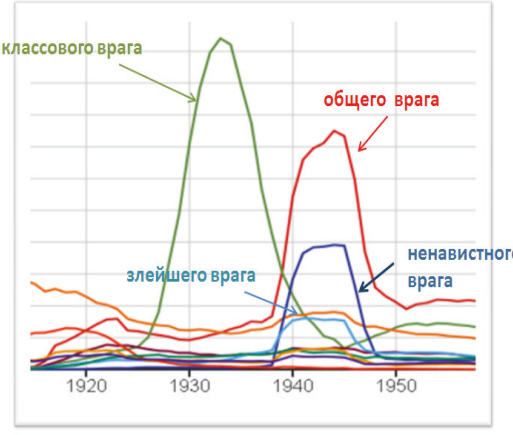

Рис. 1. **Фрагменты наиболее характерных графиков частотного поведения биграмм со словоформой «врага»**

При сопоставлении биграмм «врагов народа» и «классового врага» видно, что в целом тенденция поведения этих биграмм сходна: рост употребляемости в 1920-х при практическом совпадении кривых до середины 1920-х, пики обеих кривых (с некоторыми различиями в высоте и времени пиков) в 1930-е гг., снижением частотности в военные годы, затем некоторый рост в начале 1950-х. Такую модель частотного

поведения лексических единиц назовем «моделью 1»: высокая частота употребления, потом спад и новый, но уже меньший подъем в силу реальных исторических причин.

Рост частот в начале 1950-х, очевидно, соответствует послевоенной волне репрессий. В 1960-е гг. частота встречаемости выражения «врагов народа» ниже, чем выражения «классового врага», хотя разница невелика, а тенденция к плавному снижению соблюдается в обоих случаях. Существенные различия в поведении кривых возникают во второй половине 1980-х и до конца 1990-х. В этот период частота встречаемости биграммы «врагов народа» заметно растет. Это, по-видимому, отражает появление большого количество текстов, содержащих критику сталинизма. Такую модель назовем «моделью 2», когда новый подъем обусловлен, скажем так, «эхом» реальных исторических событий. То же наблюдается с лексикой времен Великой Отечественной войны, когда частота ее употребления снова заметно возрастает в 1960–1970-е, что объясняется печатанием большого количества исторических монографий и военных мемуаров.

2. Синонимы слова «враг»

Следующая анализируемая лексическая единица — «кулаки». «Кулак» — лексема, имеющая много значений (сжатая кисть руки, группировка военных сил, часть механизма, скупой человек и др.). Однако, в исследуемый исторический период словоформа «кулаков» чаще всего употреблялась в «сталинском» значении: богатые крестьяне, использующие наемный труд.

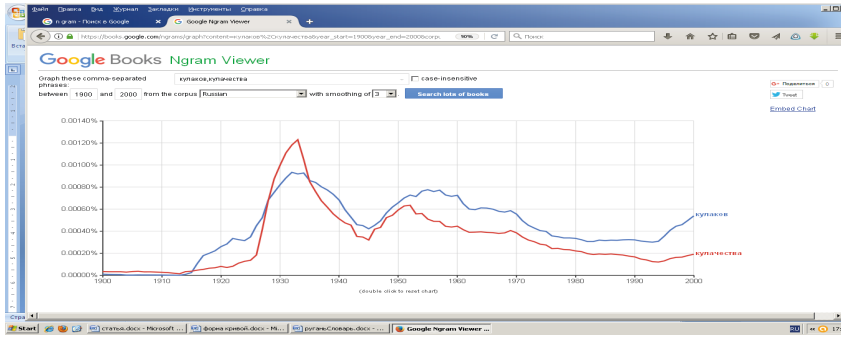

Рис. 2. **Сопоставление графиков частотного поведения словоформ «кулаков» и «кулачества»**

Это подтверждается сопоставлением частотного поведения словоформ «кулаков» и «кулачества» (рис. 2). Лексема «кулачество» представляет собой однозначный политический термин. Графики поведения кривых, тем не менее, сходны.

Рассмотрим поведение экспрессивных существительных пейоративного характера. Различными способами были отобраны следующие существительные: *бандиты, выродки, гады, двурушники, захватчики, изверги, изменники, клеветники, лакеи, лишенцы, мерзавцы, мироеды, мразь, нечисть, отщепенцы, палачи, паразиты, предатели, преступники, приспешники, прихвостни, прихлебатели, трусы, убийцы, шпионы*.

Для анализа в данной статье отберем два слова «выродки» и «бандиты».

Рис. 3. **Графики частотного поведения биграмм со словоформой «выродков» в правой позиции**

Из набора биграмм со словоформой «выродков» можно выделить две основных категории. Те биграммы, что в левой части имеют прилагательные «бухаринских» и «троцкистских», можно отнести к категории, обозначаемой как «враги народа», их пики приходятся на вторую половину 1930-х гг. Вторая категория — биграммы с прилагательными «фашистские» и «гитлеровские», которые относятся к нацистской

Германии. Однако обращает на себя внимание, что пик встречаемости биграммы «фашистских выродков» выпадает на 1939 г., т. е. данное выражение относится еще не к военному противнику, а употребляется в бранном значении применительно к «врагам народа» внутри страны. Пик встречаемости биграммы «гитлеровских выродков» приходится на 1943 г. Рост употребляемости этой биграммы начинается чуть позднее, чем биграммы «фашистских выродков». Вероятно, это отражает тот факт, что накануне войны фигура Гитлера не представлялась в советских печатных документах как однозначное воплощение зла.

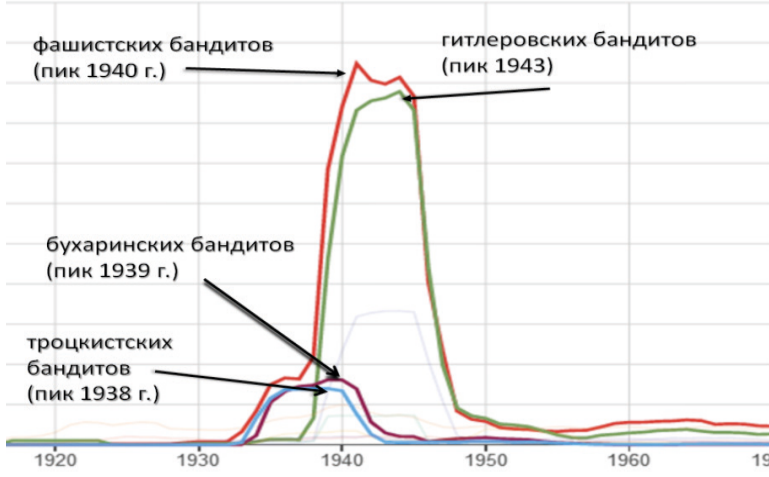

Рис. 4. **Графики частотного поведения биграмм со словоформой «бандитов» в правой позиции**

Если сравнить самые частотные биграммы со словоформами «выродков» и «бандитов» (рис. 3 и 4) и прилагательными, определяющими «адресатов» брани, то можно отметить совпадение годов их пиков. Совершенно очевидно, что это объясняется тем, что эти словосочетания употреблялись исключительно в «политическом» значении.

На рис. 5 показано поведение одной биграммы «злейший враг», но составляющие её слова взяты в разных грамматических формах. Видно, что частотное поведение этих биграмм следует разным моделям. Выражение «злейших врагов», по-видимому, синонимично выражению «врагов народа», и отсюда повторный пик в начале 1950-х как отражение внутриполитической ситуации в СССР, а «злейшего

врага» — выражениям «общего врага», «гитлеровских захватчиков» и другим, относящимся к нацистской Германии. Можно сказать, что здесь мы наблюдаем явление семантизации грамматической формы.

Рис. 5. **Частотное поведение биграмм «злейшего врага» и «злейших врагов»**

Заключение

Графики, отражающие частотное поведение лексических единиц (словоформ и коллокаций), позволяют выявить группы единиц, частотное поведение которых во времени имеет определенные черты сходства. На наш взгляд, можно говорить о прототипах частотного поведения лексических единиц в данном языке и в данный исторический период.

Частотное поведение лексических единиц, безусловно, обусловлено действием многих социально-психологических, политических и других факторов. Эти факторы и их корреляция с полученными данными требуют профессионального описания. Однако и чисто лингвистический взгляд добавляет некоторые аспекты к осмыслению исторических процессов.

Литература

1. *Michel, J-B. et al.* (2011), Quantitative Analysis of Culture Using Millions of Digitized Books science. Science 331, 176 (2011); DOI 1126/Science. 1199644 / J-B. Michel et al. URL: http://www.sciencemag.org/content/331/6014/176.full.html (дата обращения 30.01.2017).
2. *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., с. 303–327.

References

1. *Michel, J-B. et al*. (2011), Quantitative Analysis of Culture Using Millions of Digitized Books science, Science 331, 176 (2011); DOI 1126/Science. 1199644. Available at: http://www.sciencemag.org/content/331/6014/176.full.html (last access 30.01.2017).
2. *Zakharov V. P., Masevich A. C.* (2014), Diakhronicheskie issledovanija na osnove korpusa russkikh tekstov Google Books Ngram Viewer [Diachronic research on the base of the Russian corpus of the Google Books Ngram Viewer]. In: Strukturnaja i prikladnaja lingvistika [Structural and Applied Linguistics], Vol. 10, Saint-Petersburg, pp. 303–327.

Масевич Андрей Цезаревич
Masevich Andrei
E-mail: andmasev@mail.ru

Захаров Виктор Павлович
Zakharov Victor
E-mail: v.zakharov@spbu.ru

Санкт-Петербургский государственный институт культуры (Россия)
St Petersburg State Institute of Culture (Russia)

Ю. С. Масленникова, В. В. Бочкарев, В. Д. Соловьев
Yu. S. Maslennikova, V. V. Bochkarev, V. D. Solovyev

ВЕРОЯТНОСТНАЯ МОДЕЛЬ ДЛЯ ОЦЕНКИ ОБЪЕМА ЛЕКСИКОНА ПО ДАННЫМ КОРПУСА GOOGLE BOOKS NGRAM[1]

PROBABILITY MODEL FOR A VOCABULARY SIZE ESTIMATION USING THE GOOGLE BOOKS NGRAM CORPUS

Аннотация. Разработан и протестирован метод уточнения скорость образования новых слов на основе вероятностных оценок объема лексикона по корпусным данным. Метод использует оценки прогнозирования «в прошлое» частоты употребления редких словоформ на основе линейной или нейросетевой моделей, обучаемых по критерию максимального правдоподобия. При таком подходе выигрыш в точности оценки частот употребления в ранние годы тем выше, чем меньшую частоту имеет слово в анализируемый период. Используя апостериорные оценки вероятности частоты появления словоформ, был уточнен объем лексикона для разных лет и, соответственно, скорость образования новых слов. В рамках рассмотренной вероятностной модели было показано, что в более ранних работах скорость появления новых слов завышена как минимум в 2 раза.
Ключевые слова. Частоты слов, объем лексикона, прогнозирование, Google Books Ngram.

Abstract. A method for determining the birth rate of new words is developed and tested, the method is based on probabilistic estimates of the vocabulary size using a corpus of books. Back predicted frequencies of rare words are estimated using linear and neural networks models that are optimized by the maximum likelihood criteria. This approach provides the gain in the accuracy of frequencies estimation for the early years, the benefit is higher than lower the frequency of the word during analyzed period. Using a posteriori estimates for the probability of words births, the vocabulary size for different years and the birth rate of new words were clarified. According the proposed probabilistic model, it was shown that in earlier papers the birth rate of new words is inflated at least 2 times.
Keywords. Word frequencies, vocabulary size, prediction, Google Books Ngram.

1. Введение

Несмотря на длительную историю изучения языков, мы до сих пор не знаем, даже приблизительно, сколько же слов содержит конкретный язык. Возьмем для примера английский язык. Наиболее полный из изданных до сих пор словарей английского — Oxford English Dictionary [OED Online 2017] — содержит в настоящее время более 600 000 слов. Однако, очевидно, что это не все слова языка. Например, они не содержат чрезвычайно редкие слова (встречающиеся реже чем 1 на миллиард словоупотреблений). Надежда на получение до-

[1] РФФИ № 15-06-07402

статочно полного списка слов языка появилась после создания Google корпуса Google Books Ngram, содержащего более 500 млрд. слов английского языка. В [Michel et al. 2011] предпринята попытка оценить число слов по этому корпусу. Оценки получены только в трех точках. В 1900 г. язык содержал, по их оценке, 544 тыс. слов, в 1950 — 597 тыс., в 2000 — 1022 тыс. слов. В [Michel et al. 2011] приведены графики числа слов, полученные линейной экстраполяцией на остальные годы 20-го века. Подход, примененный в [Michel et al. 2011], не учитывает в полной мере очень редкие слова. Такие слова могут существовать в языке, но не появиться в некотором году, в силу ограниченного объема текстов в данный год. В настоящей работе предлагается математическая модель, позволяющая учесть этот фактор и более точно оценить реальное число слов, например, в ранние года. В основе метода лежат прогностические оценки «в прошлое» частоты употребления редких словоформ, которые потом используются для расчета более правдоподобных оценок вероятности появления словоформы в лексиконе с учетом объема корпуса.

2. Метод и результаты

В основе предлагаемого подхода лежит идея использовать для уточнения оценки частоты употребления редких слов в ранние периоды времени результаты их прогнозирования. Как правило, оценки частот для более позднего времени более достоверны ввиду большего объема корпуса. Следовательно, на основе поздних данных можно спрогнозировать ожидаемые значения частоты для более ранних периодов. Такой способ называется прогнозирование «в прошлое». В случае линейной авторегрессионной модели порядка m прогнозируемые «в прошлое» отсчеты x_t^b (буква b означает слово «back» — «назад») могут быть записаны, как $\hat{x}_t^b = -\sum_{k=1}^{m} a_k^b x_{t+k}$.

При поиске коэффициентов a_k^b данной модели следует учесть закон распределения значений исследуемого ряда. В нашем случае речь идет об оценке числа употреблений редких слов, а, следовательно, можно ожидать, что такой ряд распределен согласно закону Пуассона, который задается функцией вероятности:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Дисперсия и математическое ожидание случайной величины — числа употреблений словоформы, распределённой в соответствии с законом Пуассона, равны параметру распределения: $M(X) = D(X) = \lambda$. Наиболее точные оценки параметра λ можно получить с использованием метода максимального правдоподобия [Jackson 1989]. В этом случае максимизируется логарифмическая функция правдоподобия вида:

$$logL(\hat{X}, \lambda) = \sum_i \log(f(X_i)) - \sum_i f(X_i) - \sum_i \hat{X}!$$

Аналогично можно построить также и нелинейную модель прогнозирования с использованием аппарат искусственных нейронных сетей с обучением по методу максимального правдоподобия. Более подробно данный подход представлен в статье [Maslennikova et al. 2014].

Алгоритмы прогнозирования были протестированы на временных рядах частот, полученных из корпуса Google Books Ngram по 20 тыс. английских словоформ, которые редко встречались в 1800-х годах. При использовании указанного выше подхода выигрыш в точности оценки частоты употребления оказывается тем больше, чем меньшую частоту имеет слово в анализируемый период. Для примера, при частоте 0.5 употреблений в год среднеквадратическое отклонение оценки уменьшается в 2 раза по сравнению с обычной оценкой по среднему значению эмпирической частоты. Прогнозирование частот употребления на разные горизонты времени показало, что с увеличением горизонта прогноза СКО увеличивается.

Для примера на рис. 1 показаны результаты прогнозирования «в прошлое» частоты употребления двух редких словоформ ‘shiftlessness’ и ‘tunnelling’ простой моделью линейного предсказания первого порядка (данная модель приводит к экспоненциальной зависимости частоты от времени). График представлен в логарифмическом масштабе по оси ординат.

На примере прогнозирования частот употребления всех 20 тысяч редких английских словоформ было показано, что ошибка прогнозирования в случае линейного и нейросетевого подходов распределена приблизительно по логнормальному закону (рис. 2). На рис. 2 показана плотность распределения натурального логарифма ошибки прогнозирования частоты употребления словоформ, а также ее аппроксимация методом ядерных оценок и нормальным законом распределения (m = 0,34, σ = 1,86). Из рисунка видно качественное соответствие распределения ошибки логнормальному закону.

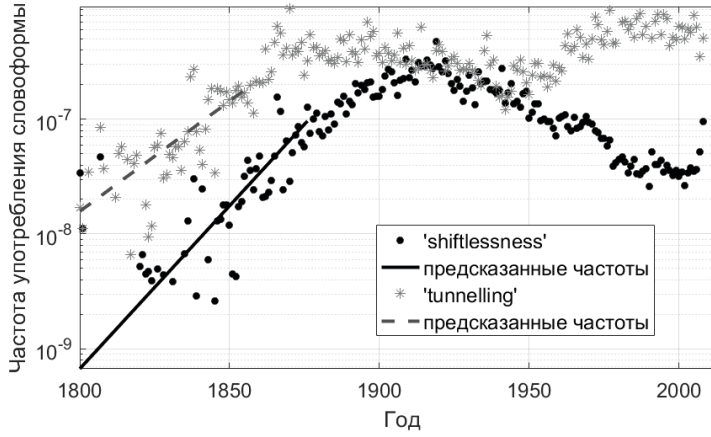

Рис. 1. **Прогнозирование частоты употребления «в прошлое» на примере двух словоформ: 'shiftlessness' и 'tunnelling'**

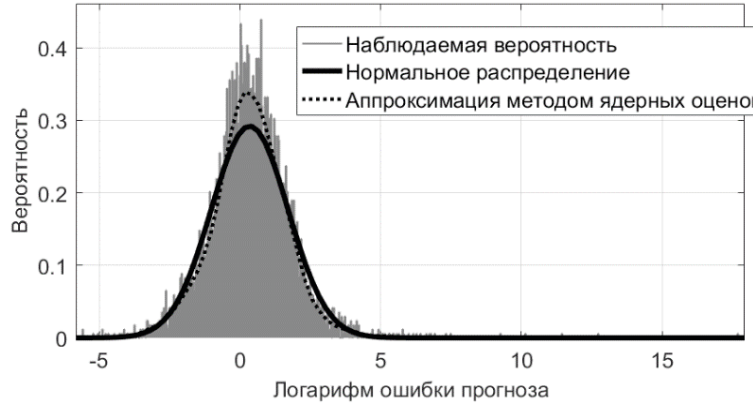

Рис. 2. **Плотность распределения логарифма ошибок прогнозирования частоты употребления словоформ на 1 год назад**

Знание закона распределения ошибок дает возможность при наличии прогноза на интересующий нас интервал времени оценить частоты употребления по критерию максимума апостериорной вероятности. Данный критерий дает существенное преимущество по сравнению с оценками по среднему значению эмпирической частоты.

Наличие хороших оценок частоты употребления словоформ для раннего периода дает возможность использовать эту информацию для уточнения действительного объема лексикона и скорости словообразования. Пусть в тот или иной год слово впервые фиксируется в корпусе. Это может означать, что ранее уже присутствовавшее в живом языке слово попало в корпус из-за увеличения его объема. С другой стороны, это может быть действительно новое слово. Экстраполируя частоту на предшествовавшие годы, мы можем рассчитать, какова вероятность того, что слово с такой частотой не будет зафиксировано в корпусе известного объема. Выполнив такие подсчеты для каждого слова, получаем возможность уточнить объем лексикона (и соответственно, скорость появления новых слов) для разных лет. Пусть f_t это относительная частота употребления слова за определенный год *t*, N_t — объем корпуса в словах за этот год. Тогда указанная вероятность может быть оценена, как $P(X = 0) = \prod_t (1 - f_t)^{N_t}$. То есть, по сути, решается задача проверки статистической гипотезы о том, что словоформа присутствовало в языке и ранее того момента, когда в первый раз было зафиксировано в корпусе. Например, для словоформы *'shiftlessness'* вероятность того, что ранее 1800 г. словоформа не попала в базу из-за малого объема корпуса составила 0,75, а для словоформы *'tunnelling'* — 0,56. Видно, что данная вероятность достаточно высокая.

Эффективность предложенного метода была проверена посредством статистического моделирования. Для этого были отобраны случайным образом ряд слов и построены временные ряды их частот. Далее значения частот занижались в нужное число раз, и генерировались случайные временные ряды с пуассоновским распределением, соответствующие данной временной зависимости частоты. После этого выполнялась оценка частоты для ранних периодов и определялась ошибка. Таким образом были получены значения среднеквадратической ошибки оценки при различном уровне средней частоты и определен выигрыш, получаемый по сравнению со стандартным подходом.

3. Заключение

Используя оценки частот для 20 тысяч редких словоформ английского языка, был уточнен объем лексикона для разных лет и, соответственно, скорость образования новых слов. В рамках рассмотренной вероятностной модели было показано, что в более раннее опубликованных работах скорость появления новых слов завышена как минимум в 2 раза.

References

1. OED Online. March 2017. Oxford University Press. Available at: http://www.oed.com/viewdictionaryentry/Entry/11125 (accessed April 05, 2017).
2. *Michel J., Shen Y., Aiden A., Veres A., Gray M., The Google Books Team, Pickett J., Hoiberg D., Clancy D., Norvig P., Orwang J., Pinker S., Nowak M., Aiden E.* (2011), Quantitative Analysis of Culture Using Millions of Digitized Books. In: Science. 331 (6014), pp. 176–182.
3. *Jackson L. B.* (1989), Digital Filters and Signal Processing. 2nd Edition. Boston, Kluwer Academic Publishers, pp. 255–257.
4. *Maslennikova Y. S., Bochkarev V. V., Voloskov D. S.* (2014), Modelling of word usage frequency dynamics using artificial neural network. In: Journal of Physics: Conference Series, 490, 012180.

Масленникова Юлия Сергеевна
Maslennikova Yulia
E-mail: yuliamsl@gmail.com

Бочкарев Владимир Владимирович
Bochkarev Vladimir
E-mail: vbochkarev@mail.ru

Соловьев Валерий Дмитриевич
Solovyev Valery
E-mail: maki.solovyev@mail.ru

Казанский федеральный университет (Россия)
Kazan Federal University (Russia)

Ю. И. Морозова, Е. Б. Козеренко
Yu. I. Morozova, E. B. Kozerenko

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ТЕРМИНОВ ПРЕДМЕТНОЙ ОБЛАСТИ «ВИРУСОЛОГИЯ»

AUTOMATIC RECOGNITION OF TERMS OF THE SUBJECT AREA «VIRUSOLOGY»

Аннотация. Доклад посвящен проблеме автоматического распознавания терминов узких предметных областей. Описан эксперимент по применению статистического метода выделения терминов на материале русскоязычных текстов предметной области «Вирусология». Сравниваются два варианта этого метода, дается оценка их применимости.
Ключевые слова. Автоматическое распознавание терминов; предметная область; корпус текстов; терминология.

Abstract. The report deals with the problem of automatic recognition of terms of narrow subject areas. The authors describe an experiment of applying the statistical method of term recognition. A corpus of texts of the subject field «Virusology» in Russian is used for the experiment. Two variants of the method are compared from the point of view of their applicability.
Keywords. Automatic term recognition; subject area; text corpus; terminology.

1. Методы автоматического распознавания терминов

Доклад посвящен актуальной теме автоматического распознавания терминов в текстах научных статей.Исследователями были предложены различные методы автоматического распознавания терминов, основанные на словарях, правилах, машинном обучении, статистическом анализе.

Словарный метод предполагает наличие готового словаря изучаемой предметной области, с которым сопоставляются лексические последовательности, встречающиеся в тексте. Данный метод позволяет выделять термины в текстах с высокой точностью, однако составление и обновление словарей различных предметных областей является трудоемкой задачей. Кроме того, в некоторых науках существуют такие классы терминов, которые содержат миллионы наименований и пополняются новыми наименованиями с очень большой скоростью, что делает составление и пополнение словарей в ручном режиме невозможным. В связи с этим актуальной является задача автоматизации процесса составления словарей различных предметных областей, для решения которой было предложено несколько методов.

Метод, основанный на правилах, использует шаблоны построения терминов определенных семантических классов, задаваемые экспертами. В описании шаблонов используются орфографические, лексические, морфологические и синтаксические признаки. Например, в работе [Ananiadou 1994] предложен метод морфологического анализа слов английского языка, позволяющий распознавать медицинские термины.

В рамках метода, основанного на машинном обучении, используются корпусы текстов с лингвистической разметкой. Цель машинного обучения — создать систему, способную распознавать в тексте термины определенных семантических классов. Для обучения системы используются такие признаки как графематическое оформление, часть речи, морфемный состав, окружающие слова и др. Например, в работе [Raychaudhuri 2002] описано решение задачи распознавания названий генов и белков с помощью модели максимальной энтропии.

Статистический метод позволяет найти в коллекции текстов предметной области наиболее вероятные термины, без учета их семантического класса. Из текстовой коллекции извлекаются все слова и словосочетания, соответствующие синтаксическим шаблонам, и на основе частотных характеристик производится их ранжирование ([Frantzi 2000], [Захаров 2014], [Лукашевич 2010], [Браславский 2007]). Слова и словосочетания, получившие ранг выше определенного порогового значения, выдаются системой в качестве потенциальных терминов.

Дистрибутивно-семантический метод предполагает использование информации о контекстах, в которых встречаются потенциальные термины ([Frantzi 2000], [Лукашевич 2010]).

2. Эксперимент по применению статистического метода

Целью данной работы было сравнительное исследование различных вариантов статистического метода на материале текстовой коллекции, относящейся к предметной области «Вирусология» (объем коллекции — 403 362 словоупотребления). После лемматизации текстовой коллекции средствами системы Sketch Engine [Система управления корпусами текстов Sketch Engine], из нее были извлечены все леммы существительных, встретившихся в коллекции более одного раза. Затем эти слова были ранжированы тремя различными способами.

В качестве первого способа ранжирования была рассмотрена частота встречаемости слов в коллекции исследуемых текстов.

Второй и третий способы основываются на сравнении частот встречаемости слов в двух коллекциях (исследуемых и референтных текстов), в соответствии с формулой, предложенной в [Statistics used in the Sketch Engine]:

$$\frac{f_{focus} + n}{f_{ref} + n} \quad (1)$$

где f_{focus} — нормированная частота встречаемости слова в коллекции исследуемых текстов, f_{ref} — нормированная частота встречаемости слова в коллекции референтных текстов, n — варьируемый параметр.

Нормированная частота встречаемости слова в коллекции текстов вычисляется по формуле:

$$\frac{f \cdot 1000000}{N} \quad (2)$$

где f — частота встречаемости слова в коллекции текстов, N — объем коллекции текстов.

Данная формула помогает снижать ранг высокочастотных общеупотребительных слов и тем самым улучшает качество извлечения терминов. Одна из референтных коллекций, используемых в эксперименте, это корпус текстов RuTenTen, содержащий тексты из Интернет по различным темам и доступный в системе Sketch Engine (объем корпуса — 18 280 486 876 словоупотреблений). Можно предположить, что, если использовать не любые тексты из Интернет, а только научные тексты, то качество извлечения терминов увеличится. Для проверки этой гипотезы авторами был создан корпус текстов научных статей по различным отраслям науки (объем корпуса — 10 249 443 словоупотребления).

Для оценивания качества распознавания терминов в предыдущих работах использовалось либо сравнение со словарем, либо привлечение экспертов. Авторами данного доклада был проведен эксперимент по экспертной оценке ранжированных списков с участием двух экспертов предметной области. Была использована методика работы с экспертами, описанная в [Браславский 2007]. Эксперт последовательно для каждого слова отвечает на вопрос: «Является ли данное слово термином предметной области «Вирусология»?» Варианты ответа эксперта: «да» или «нет». Терминами считаются те слова, которые оба эксперта признали терминами. Степень согласия экспертов

между собой, вычисленная как доля совпадающих ответов, оказалась невысокой (66 %), в связи с чем было принято решение использовать в качестве эталона толковый словарь «Молекулярная биология и генетика» [Толковый словарь «Молекулярная биология и генетика»]. Качество распознавания терминов вычислялось как число пересечений между первыми 100 словами ранжированного списка и словарем.

На первом этапе эксперимента изучалась зависимость между значением варьируемого параметра n и качеством распознавания терминов при применении метода ранжирования, основанного на формуле 1. Было выяснено, что для двух описанных референтных корпусов текстов наилучшие результаты достигаются при $n = 100$.

На втором этапе эксперимента было проведено сравнение трех способов между собой. Коллекция исследуемых текстов была разбита на 10 приблизительно равных по объему частей, и оценка качества проводилась отдельно для каждой части. Таким образом были получены три выборки, средние значения которых были сравнены между собой с помощью теста Стьюдента. В табл. 1 приводятся средние значения показателей качества трех методов распознавания терминов. Для второго и третьего способов указываются значения, полученные при $n = 100$. Средние значения различаются между собой статистически значимым образом (тест Стьюдента, $p < 0,05$).

Таблица 1. **Сравнение трех вариантов статистического метода автоматического распознавания терминов**

Вариант метода	Качество извлечения терминов (средние значения)
частота встречаемости слова в коллекции исследуемых текстов	13,3
соотношение частот встречаемости слова в коллекции исследуемых текстов и в коллекции научных статей	17,9
соотношение частот встречаемости слова в коллекции исследуемых текстов и в корпусе RuTenTen	19,5

Как видно из табл. 1, гипотеза о возможности улучшения качества распознавания терминов при применении коллекции референтных текстов, состоящей из текстов научных статей, не оправдалась.

3. Заключение

В докладе описан эксперимент по сравнению различных вариантов статистического метода распознавания терминов на материале текстовой коллекции, относящейся к предметной области «Вирусология», в ходе которого были сделаны следующие выводы:

- при оценке качества работы методов привлекать экспертов нецелесообразно, так как степень согласия между ними является низкой (66 %), необходимо либо использовать словарный метод оценки, либо конкретизировать задачу, стоящую перед экспертами, ограничившись лишь терминами определенных семантических классов;
- применение коллекции референтных текстов, состоящей из текстов научных статей, не позволяет добиться улучшения результата по сравнению с применением корпуса RuTenTen в качестве референтной коллекции;
- наилучшее качество распознавания терминов было достигнуто при применении формулы 1, вычисляющей соотношение частот встречаемости слова в коллекции исследуемых текстов и в корпусе RuTenTen, при значении варьируемого параметра *n* равном 100.

Литература

1. *Браславский П. И., Соколов Е. А.* (2007), Автоматическое извлечение терминологии с использованием поисковых машин Интернета // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог 2007», с. 89–94.
2. *Захаров В. П., Хохлова М. В.* (2014), Автоматическое выявление терминологических словосочетаний // Структурная и прикладная лингвистика, 10, с. 182–200.
3. *Лукашевич Н. В., Логачев Ю. М.* (2010), Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование, 11. URL: http://num-meth.srcc.msu.ru/zhurnal/tom_2010/pdf/v11r211.pdf (09.03.2017).
4. Система управления корпусами текстов Sketch Engine. URL: https://www.sketchengine.co.uk (09.03.2017).
5. Толковый словарь «Молекулярная биология и генетика»ю URL: http://enc-dic.com/genetics (09.03.2017).
6. *Ananiadou S.* (1994), A Methodology for Automatic Term Recognition // Proceedings of COLING-94, pp. 1034–1038.
7. *Frantzi K., Ananiadou S., Mima H.* (2000), Automatic Recognition of Multi-Word Terms // International Journal of Digital Libraries, 3 (2), pp. 117–132.

8. *Raychaudhuri S., Chang J. T., Sutphin P. D., Altman R. B.* (2002), Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature // Genome research, 12 (1), pp. 203–214.
9. Statistics used in the Sketch Engine. URL: https://www.sketchengine.co.uk/wp-content/uploads/ske-statistics.pdf (09.03.2017).

References

1. *Ananiadou S.* (1994), A Methodology for Automatic Term Recognition. In: Proceedings of COLING-94, pp. 1034–1038.
2. *Braslavskij P. I., Sokolov E. A.* (2007), Avtomaticheskoe izvlechenie terminologii s ispol'zovaniem poiskovyh mashin Interneta [Automatic term extraction using Internet search engines]. In: Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy Mezhdunarodnoj konferencii "Dialog 2007" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog–2007"], pp. 89–94.
3. *Frantzi K., Ananiadou S., Mima H.* (2000), Automatic Recognition of Multi-Word Terms. In: International Journal of Digital Libraries, 3 (2), pp. 117–132.
4. *Lukashevich N. V., Logachev Ju. M.* (2010), Kombinirovanie priznakov dlja avtomaticheskogo izvlechenija terminov [Combining features for automatic term recognition]. In: Vychislitel'nye metody i programmirovanie [Computational methods and programming], 11. Available at: http://num-meth.srcc.msu.ru/zhurnal/tom_2010/pdf/v11r211.pdf (09.03.2017).
5. *Raychaudhuri S., Chang J. T., Sutphin P. D., Altman R. B.* (2002), Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature. In: Genome research, 12 (1), pp. 203–214.
6. Statistics used in the Sketch Engine. Available at: https://www.sketchengine.co.uk/wp-content/uploads/ske-statistics.pdf (09.03.2017).
7. The Sketch Engine corpus management system. Available at: https://www.sketchengine.co.uk (09.03.2017).
8. Tolkovyj slovar' "Molekuljarnaja biologija i genetika" [Explanatory dictionary "Molecular biology and genetics"]. Available at: http://enc-dic.com/genetics (09.03.2017).
9. *Zaharov V. P., Hohlova M. V.* (2014), Avtomaticheskoe vyjavlenie terminologicheskih slovosochetanij [Automatic recognition of terminological word combinations]. In: Strukturnaja i prikladnaja lingvistika [Structural and applied linguistics], 10, pp. 182–200.

Морозова Юлия Игоревна
Morozova Yuliya
E-mail: yulia-ipi@yandex.ru

Козеренко Елена Борисовна
Kozerenko Elena
E-mail: kozerenko@mail.ru

Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук (Россия), Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский педагогический государственный университет» (Россия)
Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (Russia), Moscow State Pedagogical University (Russia)

А. Д. Москвина, О. А. Митрофанова,
А. Р. Ерофеева, Я. К. Харабет

A. D. Moskvina, O. A. Mitrofanova,
A. R. Erofeeva, Ja. K. Charabet

АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ И СЛОВОСОЧЕТАНИЙ ИЗ РУССКОЯЗЫЧНЫХ КОРПУСОВ ТЕКСТОВ С ПОМОЩЬЮ АЛГОРИТМА RAKE[1]

AUTOMATIC EXTRACTION OF KEY WORDS AND PHRASES FROM RUSSIAN TEXT CORPORA BY MEANS OF RAKE ALGORITHM

Аннотация. В докладе представлены результаты работы по модификации алгоритма RAKE, используемого для быстрого извлечения ключевых слов и словосочетаний. В качестве источника информации о границах синтаксических групп в RAKE предлагаются правила грамматики синтаксического анализатора для русского языка на основе NLTK (NLTK4RUSSIAN). Для определения эффективности работы алгоритма с русскоязычными текстами были проведены эксперименты на материале представительных корпусов.
Ключевые слова. Автоматическое выделение ключевых слов и словосочетаний, RAKE, русскоязычные корпусы текстов.

Abstract. RAKE algorithm is used for effective and rapid automatic extraction of keywords and phrases. We present modification of RAKE designed for processing Russian corpora. RAKE uses information about the boundaries of syntactic groups. In our study RAKE addresses the rules of the Russian syntactic parser based on NLTK (NLTK4RUSSIAN). Experiments carried out on representative Russian text corpora prove the reliability of RAKE modification.
Keywords. Automatic extraction of key words and phrases, RAKE, Russian text corpora.

Одной из актуальных задач автоматической обработки текста является извлечение ключевых слов и словосочетаний. Существуют различные подходы к ее решению. С точки зрения результатов анализа принято противопоставлять методы, направленные на выделение а) ключевых слов, б) ключевых словосочетаний, в) ключевых слов и словосочетаний. Также есть возможность выбора между методами, требующими внешние источники информации (словари или фоновые корпусы), и не требующими таковых. С точки зрения задейство-

[1] Исследование выполняется при поддержке гранта РФФИ по проекту № 16-06-00529 «Разработка лингвистического комплекса для автоматического семантического анализа русскоязычных корпусов текстов с применением статистических методов».

ванного теоретического аппарата различаются статистические (мера TF×IDF, логарифмическая функция правдоподобия Log-Likelihood, совмещение модели совместной встречаемости и критерия X^2 и др.), лингвистические (основанные на синтаксических алгоритмах и словарных данных) и гибридные (C-Value, KEA, RAKE, графовые модели TextRank, DegExt и др.) методы. Выбор алгоритма определяется особенностями решаемой задачи: см., например, исследование [Красавина, Мирзагитова 2015].

В фокусе нашего исследования находится алгоритм RAKE (Rapid automatic keyword extraction), запатентованный авторами в 2009 году [Rose et al.]. Алгоритм RAKE основан на предположении о том, что ключевые выражения часто представляют собой не только отдельные слова, но и фразы. Такие фразы могут состоять из нескольких слов, однако не включают в себя знаки пунктуации, служебные слова и слова, не несущие ярко выраженного лексического значения. На первом этапе обработки текста необходимо выделить фразы-кандидаты в ключевые словосочетания. Для этого текст разбивается на отрывки по знакам препинания и словарю, содержащему так называемые стоп-слова (артикли, местоимения, вводные слова, и тд). Полученные цепочки слов являются кандидатами на роль ключевых словосочетаний. Для каждого слова на основе общей частоты слова и средней длины фразы, в которую оно входит, рассчитывается вес, в то время как вес фразы-кандидата, в свою очередь, рассчитывается как сумма весов входящих в неё слов.

Алгоритм RAKE для задач обработки естественного языка был реализован на языке Python. Позже он был адаптирован для работы с библиотеками NLTK (URL: http://www.nltk.org/) [Bird et al. 2009]. В исходном виде алгоритм пригоден для автоматической обработки англоязычных корпусов текстов. Нами были предприняты шаги, направленные на модификацию RAKE для работы с русскоязычным материалом.

Было принято решение об использовании результатов поверхностного синтаксического анализа для того, чтобы алгоритм RAKE корректного выделял отдельные слова и цепочки слов, являющиеся потенциальными ключевыми выражениями. В то же время, в силу богатства морфологии русского языка, для определения синтаксических групп и их границ требуется предварительный морфологический анализ. Основная идея заключается в том, чтобы, не затрачивая слишком много ресурсов, провести границы между теми словами, которые оче-

видно не могут входить в одно словосочетание, и сохранять данную информацию при обработке текста оригинальным алгоритмом RAKE.

Предобработка текста включает в себя разбиение текста на условные слова (по пробелам) и проставление границ условных синтаксических групп (используется знак-разделитель "|"). На этапе морфологического анализа входного текста используется морфоанализатор Pymorphy2 (URL: http://pymorphy2.readthedocs.io/en/latest/). Обработанный текст сохраняется в отельный файл, который впоследствии обрабатывается при помощи RAKE. Синтаксический анализ опирается на небольшой набор правил. Группы выделяются на основе линейной последовательности слов с определенными грамматическими характеристиками. Выделяются, в том числе, группы ADJ+N, N+Adj+N (при наличии согласования), одиночные леммы. Также используются правила, проводящие границу между двумя словоформами с определенными характеристиками. Правила составлялись на основе грамматики синтаксического парсера NLTK4RUSSIAN [Москвина и др. 2016] и проверялись на эмпирических данных в ходе экспериментов.

Правила основываются на морфологических характеристиках слов, линейно линейно следующих друг за другом, при этом используется возможность а) обозначить конец некоторой группы, объединив таким образом слова перед проведенной границей, и б) провести границу между обрабатываемыми на текущем шаге словами. Наречия, краткие формы прилагательных и причастий, компаратив, глаголы и различные глагольные формы выделяются в отдельные самостоятельные группы, т. е. встретив в тексте такое слово, алгоритм «окружает» его границами с обеих сторон. Существительное, стоящее в генитиве или аккузативе, считается завершающим синтаксическую группу — потенциальное ключевое словосочетание, и граница проводится с правой стороны. Если за существительным следует существительное или прилагательное в другом падеже, за исключением генитива, такие слова считаются принадлежащим разным синтаксическим конструкциям, и граница проводится между ними. Именные группы, состоящие из согласованных существительного и прилагательного, выделяются с помощью проставления границы справа от существительного, идущего после прилагательного и стоящего в том же падеже. Правила представляют собой условный цикл, и, функционируя в определенном порядке, позволяют сохранить некоторые относительно длинные вложенные конструкции, как, например, *производство пилотируемых кораблей*, или *трансформация непилотируемых космических аппара-*

тов. Данные группы при этом не являются результатом глубокого синтаксического анализа, в приведенных примерах первые и последние слова никак формально не связаны, а правила работают только на стоящих рядом словах. Пример обработанного предложения:

Наиболее | совершенными герметизирующими материалами, обеспечивающими | надежную | и | устойчивую герметичность соединений, | являются | самовулканизирующиеся пасты.

Здесь видно, как в отдельные группы выделены наречие и глагол, остались нетронутыми согласованные именные группы *совершенными герметизирующими материалами* и *самовулканизирующиеся пасты*, а также генитивная группа *устойчивую герметичность соединений*.

В некоторых случаях возможным недостатком является, выделение слишком длинных конструкций, которые формально имеют ту же синтаксическую структуру, что и словосочетания, состоящие из двух-трех слов. Например, такой конструкцией является следующая фраза из списка полученных ключевых словосочетаний на основе корпуса научных текстов: *собственную программу ускоренного научно-технического развития*. Формально она имеет структуру генитивной именной группы, т. е. такую же как у словосочетания *затрата мощности*, однако второе получает меньший вес, поскольку состоит из меньшего количества слов. Такие длинные конструкции вряд ли разумно считать кандидатами в ключевые словосочетания. Правила составлялись на основе эмпирических данных экспериментов. Правила имеют достаточно наглядный вид, их легко корректировать или отключать.

На следующем этапе обработанный текст принимается программой, реализующей алгоритм RAKE с использованием NLTK. Программа производит токенизацию текста, замену знаков пунктуации и стоп-слов на символ-разделитель, выделение последовательностей между ними, расчет их веса. Ключевые выражения в ранжированном виде сохраняются в отдельный файл.

Ниже приведен фрагмент выдачи:

пониженными концентрациями тяжелых стабильных изотопов 13.669047619
устойчивую синхронизацию циркадианных ритмов 13.6666666667
контрольными «наземными» партиями геля 13.6666666667
осушитель отработанных субстратных вкладышей 13.6666666667
….

капитальные затраты 4.11213517665
использование технологии 4.11210706245
системой самонаведения 4.11209150327
эффект памяти 4.11204481793
кабельной сети 4.11197916667
восполнением сети 4.11197916667
физических процессов 4.11193218865
уменьшение активности 4.11184792219
стартовой орбиты 4.11181646763

Эксперименты проводились на материале четырех корпусов русскоязычных текстов, представляющих научный, публицистический, официально-деловой и художественный функциональные стили. Хотя тексты различаются стилистическими параметрами, их объединяет тематика: ракетостроение и аэрокосмические исследования. Объем каждого из корпусов составляет примерно 500 тыс. с/у, суммарный объем обработанных текстов, тем самым, оценивается в 2 млн с/у. Для каждого из корпусов были получены списки ключевых выражений. Некоторые примеры триграмм, биграмм и униграмм приведены ниже.

Научный корпус

триграммы: далекие звездные миры, рассеивание баллистических ракет, аппаратура космических ракет, использование шлюзовых камер, создание экспериментальной установки и т. д.

биграммы: наветренная поверхность, медицинское обеспечение, детальное изучение, система управления, точные данные, межорбитальный буксир-стыковщик, мягкая посадка и т. д.

униграммы: наземный, топливный, теоретический, стальной, геологический, флот, термоэлектрический, инфракрасный и т. д.

Публицистический корпус

триграммы: сборник научных трудов, законодательная увязка программ, орбитальная станция «Мир», международный астрономический союз, средства массовой информации и т. д.

биграммы: краткий перечень, научный центр, космический комплекс, федеральная служба, наземный сегмент, пусковая установка и т. д.

униграммы: наблюдатель, конструкционный, дом-музей, кризис, европейский, рынок, беспилотный, главный, дешевый и т. д.

Официально-деловой корпус

триграммы: высшее учебное заведение, московский авиационный институт, торцевой плазменный двигатель, устная публичная речь, современная программная среда, реализация компетентностного подхода и т. д.

биграммы: техническая эксплуатация, государственный язык, программный продукт, потенциальный клиент, Верховный Совет, математическое моделирование, правовой документ, квалифицированный специалист и т. д.

униграммы: площадка, комплексный, конструктивный, ракета, выбор, развитие, поток, авиационный, инструктаж, транспортный, двигатель, неофициальный и т. д.

Художественный корпус

триграммы: ажурные чугунные столбики, пучки свежих ландышей, вымирание архаических млекопитающих, непоправимый этический надлом и т. д.

биграммы: темная глыба, вихревые потоки, яркая вспышка, сотрудник центра, мощный прожектор, теневая сторона и т. д.

униграммы: прекрасный, мощный, лобызания, обстоятельство, новый, мышление, кусок и т. д.

Была проведена оценка результатов применения алгоритма RAKE при работе с экспериментальными корпусами. Мы исходили из предположения о том, что ключевые слова и словосочетания, составляющие семантическое ядро корпуса, могут иметь соответствие в тематической модели корпуса (т. е. компоненты n-грамм должны быть представлены в составе кластеров, отражающих распределение слов по темам и тем по документам корпуса). При построении тематических моделей корпусов использовался алгоритм LDA (Latent Dirichlet Allocation) в пакете GenSim для Python [Митрофанова 2015]. В каждой тематической модели отбирались 200 статистически значимых лемм (по 10 из 20 тем), далее фиксировалось их наличие/отсутствие в списках ключевых выражений. За единичными исключениями все леммы в составе тем обнаружены в верхней трети списка ключевых выражений, которая оценивается как наиболее информативная.

Полученные данные дают основания считать результаты работы алгоритма RAKE приемлемыми, а сам алгоритм RAKE в русскоязыч-

ной модификации пригодным для использования в лингвистических исследованиях. В дальнейшем планируется сравнение RAKE с другими алгоритмами и экспертиза результатов с участием информантов.

Литература

1. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing.
2. *Rose S. J., Cowley W. E., Crow V. L., Cramer N. O.* (2009), Rapid Automatic Keyword Extraction for Information Retrieval and Analysis. URL: http://www.google.co.ve/patents/US8131735
3. *Красавина В. Д., Мирзагитова А. Р.* (2015), Оптимизация поиска в системе LeadScanner с помощью автоматического выделения ключевых слов и словосочетаний // Труды международной конференции «Корпусная лингвистика–2015». СПб.
4. *Митрофанова О. А.* (2015), Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика–2015». СПб.
5. *Москвина А. Д., Орлова Д., Паничева П. В., Митрофанова О. А.* (2016), Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Компьютерная лингвистика и вычислительные онтологии. Труды XIX Международной объединенной научной конференции «Интернет и современное общество», Санкт-Петербург, 22–24 июня 2016 г.

References

1. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing.
2. *Rose S. J., Cowley W. E., Crow V. L., Cramer N. O.* (2009), Rapid Automatic Keyword Extraction for Information Retrieval and Analysis. Available at: http://www.google.co.ve/patents/US8131735
3. *Mitrofanova O. A.* (2015), Probabilistic Topic Modelling of the Russian Text Corpora by Means of GenSim Toolkit. In: Proceedings of the International Conference "Corpus Linguistics — 2015". Saint–Petersburg.
4. *Krasavina V. D., Mirzagitova A. R.* (2015), Optimization for LeadScanner system by means of automatic keyword and keyphrase extraction. In: Proceedings of the International Conference "Corpus Linguistics — 2015". Saint Petersburg.
5. *Moskvina A., Orlova D., Panicheva P., Mitrofanova O.* (2016), Development of the Core for Syntactic Parser for Russian based on NLTK libraries. In: Computational Linguistics and Digital Ontologies. IMS-2016 Proceedings. Saint Petersburg.

Москвина Анна Денисовна
Moskvina Anna
E-mail: moskvina.anya@gmail.com

Митрофанова Ольга Александровна
Mitrofanova Olga
E-mail: o.mitrofanova@spbu.ru

Ерофеева Алия Ришатовна
Erofeeva Aliia
E-mail: amirzagitova@gmail.com

Харабет Якуб Константинович
Charabet Jakub
E-mail: jkharabet@gmail.com

Санкт-Петербургский государственный университет (Россия)
Saint Petersburg State University (Russia)

И. С. Николаев
I. S. Nikolaev

КОРПУС ТЕКСТОВ НА ИЖОРСКОМ ЯЗЫКЕ КАК ОСНОВА ЛИНГВИСТИЧЕСКОЙ ЭКСПЕРТНОЙ СИСТЕМЫ

Аннотация. Одним из эффективных способов сделать материалы по ижорскому языку доступными для широкого круга специалистов и заинтересованных членов языкового сообщества может стать лингвистическая экспертная система, которая интегрирует данные текстовых и звуковых корпусов. Задачей такой экспертной системы является не хранение материалов, а их поиск и интерпретация в соответствии с запросами пользователей. Ядром экспертной системы является грамматическая база данных по ижорскому языку, а также связанная с ней лексикографическая база данных. В статье освещены вопросы интеграции и адаптации лингвистической экспертной системы с внешними корпусами. Обсуждаются некоторые перспективы создания научно-образовательного веб-ресурса по ижорскому языку.
Ключевые слова. Корпус текстов, звуковой корпус, экспертная система, база данных, ижорский язык.

Annotation. A linguistic expert system, integrating text and sound corpora data, might be one of effective means to make the Izhorian language data available to a wide range of professionals and interested members of the language community. Main aim of the expert system is not storage of data, but search and interpretation according to the user's queries. The core of the expert system is an Izhorian grammar database and an integrated lexicographic database. The article deals with the problems of integration and adaptation of the linguistic expert system with external corpora. Some perspectives of an Izhorian scientific and educational web-resource are discussed.
Keywords. Text corpora, sound corpora, expert system, database, Izhorian language.

При изучении языков, находящихся под угрозой исчезновения, одной из первоочередных задач является создание звуковых и текстовых корпусов. Такие корпусы часто создаются отдельными исследователями или небольшими группами на основе полевых материалов. Другим видом корпусов являются коллекции оцифрованных текстов, которые были собраны предыдущими поколениями ученых в другие исторические периоды развития конкретного языка или его диалектов. Особый вид материалов представляют собой словари и грамматики, а также научные статьи и монографии, посвященные отдельным лингвистическим аспектам языка. В нашем распоряжении имеются также полевые дневники, мемуары, отчеты и разнообразные документальные материалы, связанные с историей изучения данного языкового сообщества. Кроме того, в последнее время появились документальные фильмы и другие видеоматериалы, посвященные истории и современной повседневной жизни языкового сообщества, включая видеоинтервью с носителями исчезающих языков.

Все эти материалы представляют огромную ценность как для науки, так и для каждого языкового сообщества. Однако часто они хранятся в разрозненном виде, в научных учреждениях и архивах разных стран, поэтому любой новый исследователь вынужден заново искать и интерпретировать эти данные. Так обстоит дело и с материалами по ижорскому языку, относящемуся к прибалтийско-финской группе уральских языков, на котором говорят несколько десятков человек на территории Ленинградской области.

Одним из эффективных способов сделать данные по ижорскому языку доступными для широкого круга специалистов и заинтересованных членов языкового сообщества может стать лингвистическая экспертная система, которая интегрирует максимально большое количество материалов из разных источников. Задачей такой экспертной системы является не хранение данных, а их поиск и интерпретация в соответствии с запросами пользователей. При необходимости экспертная система может использоваться и для обучения языку на основе текстов и других материалов, которые в ней хранятся. Таким образом речь идет не о полноценной экспертной системе, а об ограниченной интеллектуальной системе, основанной на знаниях [Giarratano et al. 2005].

В основе экспертной системы находятся текстовые и звуковые корпусы, собранные разными исследователями в разные эпохи, размещенные, прежде всего, в Фонограммархиве Карельского научного центра РАН и частично публикованные в сборнике «Народные песни Ингерманландии» [Конкка 1974], в Архиве исчезающих языков Университета Лондона и в Архиве Общества финской литературы. Кроме того, в распоряжении автора имеется собственный звуковой и текстовый корпус современных текстов на ижорском языке на бытовые темы, полученный в результате экспедиций в Ленинградскую область в 1996–2001 годах. Информация обо всех этих корпусах приведена в табл. 1.

Многие современные научные архивы, доступны через сеть Интернет, что дает возможность пользоваться ими удаленно. Однако каждый из архивов имеет свой индивидуальный интерфейс и особые условия доступа к материалам, в некоторых случаях предусматривающие регистрацию пользователя и ограничение на скачивание материалов. Поэтому для использования корпусов текстов из этих архивов необходимо разработать и создать структуру ссылок на данные. Так было сделано, например, для Архива исчезающих языков Университе-

та Лондона. Пользователь экспертной системы может получить ссылку на определенный звуковой файл из этого архива, но прослушать его он сможет только, если сам зарегистрируется на сайте архива.

Если нет ограничений на копирование материалов, можно параллельно со ссылками на текстовые файлы корпуса создать образ корпуса и собственный интерфейс для работы с ним. Таким образом организована работы с корпусом народных песен Ингерманландии в Архиве Общества финской литературы. Копии текстов из этого корпуса хранятся в базе данных, некоторые из них морфологически аннотированы.

Таблица 1. **Ижорские корпусы и их доступность**

№	Название корпуса	Веб-адрес	Доступ
1.	Фонограммархив Карельского научного центра РАН	phonogr.krc.karelia.ru	свободный
2.	Архив исчезающих языков Университета Лондона	elar.soas.ac.uk	регистрация
3.	Архив Общества финской литературы	www.finlit.fi	свободный
4.	Корпус современных ижорских текстов	локальный ресурс	ограничен
5.	Ингерманландские народные песни	локальный ресурс	ограничен

Ядром экспертной системы является грамматическая база данных по ижорскому языку, которая создана на основе предшествующих исследований автора доклада и некоторых других грамматических описаний [Николаев 2010]. При помощи базы данных можно проводить грамматический анализ и интерпретацию текстов из архивов, а результаты исследования могут сохраняться в базе данных и, при необходимости, иметь ссылки на соответствующие тексты и звуковые файлы из корпусов. Грамматическая база данных связана с грамматическим описанием ижорского языка в электронном виде, которое сделано на основе краткой онтологии грамматических терминов и понятий.

Работа с грамматическим разделом экспертной системы может происходить, например, следующим образом.

1. Если мы слушаем звуковой файл эпической песни «Iso tammi» («Большой дуб») из корпуса Архива Общества финской литературы, то можно просмотреть его текстовую расшифровку из корпуса текстов Ингерманландских народных песен. Для этого текста можно увидеть грамматическую аннотацию, хранящуюся в грамматической базе данных. Для любой словоформы из текста можно увидеть ссылку на соответствующую парадигму словоизменения из грамматического описания. Для грамматического описания можно получить библиографическую ссылку на раздел грамматики А. Лаанеста [Лаанест 1966] или В. Поркка [Porkka 1885].

2. Если для данного звукового файла нет текстовой расшифровки, то мы можем ввести название песни или строку из текста в поле поиска и проверить нет ли в корпусах аналогичных расшифрованных и аннотированных текстов. Предусмотрен неточный поиск.

3. Если мы производим грамматический анализ текста из корпуса современных текстов на ижорском языке, то для каждой словоформы мы можем найти ссылку на похожие словоформы и их контексты в других корпусах. Такая система подсказок может также использоваться для расшифровки звуковых файлов. Все такие подсказки могут быть подкреплены ссылками на грамматическое описание.

Составной частью грамматической базы данный является лексикографическая база данных, которая представляет собой электронный словарь ижорского языка со ссылками на контексты из внешних корпусов и на грамматическую информацию. Электронный словарь составлен на основе корпуса текстов, которые были проанализированы грамматически. В словаре имеются ссылки на словари ижорского языка Р. Нирви [Nirvi 1971] и А. Лаанеста [Laanest 1997]. Лексикографическая база данных позволяет также создавать пользовательские словари, например, частотные, или словари, отражающие лексику разговорного жанра или жанра эпической песни, и включать в них оговоренное количество слов и определенную лингвистическую информацию.

Кроме того, в экспертной системе имеется библиографический раздел, содержащий ссылки на все известные и доступные научные работы по ижорскому языку, и на другую литературу, посвященную истории, культуре и языку ижорской народности, а также соседних и родственных этносов. Многие лингвистические и этнографические работы являются библиографической редкостью, поэтому известны очень узкому кругу специалистов. Даже многие новейшие статьи по

ижорской проблематике малодоступны, поэтому их включение в библиографический раздел было бы весьма желательным.

К настоящему времени разработано ядро лингвистической экспертной системы по ижорскому языку — грамматическая база данных — и производится ее интеграция с внешними корпусами текстов и звуковых файлов. Создание связей между корпусами и управляющим ядром экспертной системы — наиболее сложный и важный аспект проекта. После того, как такие связи будут созданы, возможности эвристического поиска в корпусах значительно возрастут.

В перспективе предполагается создание информационного ресурса в сети Интернет, посвященного Ингерманландии, ижорцам и ижорскому языку, а также соседним народам и их языкам. Основой этого ресурса будет рассмотренная экспертная система с набором звуковых и текстовых корпусов, грамматическая и лексикографическая база данных и библиографический раздел.

Литература

1. *Конкка У.* (ред.) (1974), Народные песни Ингерманландии. Ленинград.
2. *Лаанест А. Х.* (1966), Ижорский язык: Языки народов СССР. Том 3. М., с. 102–117.
3. *Николаев И. С.* (2010), Исследовательская база данных по морфологии ижорских эпических песен: терминология, модели и реализация // Структурная и прикладная лингвистика, № 8, с. 233–242.
4. *Giarratano J. C., Riley G. D.* (2005), Expert Systems: Principles and Programming, Boston, MA.
5. *Laanest A.* (1997), Isuri keele Hevaha murde sonastik. Tallinn.
6. *Nirvi R. E.* (1971), Inkeroismurteiden sanakirja. Helsinki.
7. *Porkka V.* (1885) Ueber den ingrischen Dialekt mit Beruecksichtigung der uebrigen ignermanlaendischen Dialekte. Helsingfors.

References

1. *Giarratano J. C., Riley G. D.* (2005), Expert Systems: Principles and Programming, Boston, MA.
2. *Konkka U.* (ed.) (1974), Narodnyje pesni Ingermanlandii [Folk songs of Ingermanland]. Leningrad.
3. *Laanest A. H.* (1966), Izhorskij jazyk [Izhorian language]. In: Jazyki narodov SSSR [Languages of Peoples of USSR]. Vol. 3. M., pp. 102–117.
4. *Laanest A.* (1997), Isuri keele Hevaha murde sonastik [Dictionary of Hevaha dialect of Izhorian Language]. Tallinn.
5. *Nikolaev I. S.* (2010), Issledovatel'skaja baza dannyh po morfologii izhorskih epicheskih pesen: terminologija, modeli i realizacija [Research Database on Izhorian Epic Songs

Morphology: Terminology, Models and Realization]. In: Strukturnaja i prikladnaja lingvistika [Structural and Applied Linguistics], no. 8, pp. 233–242.
6. *Nirvi R. E.* (1971), Inkeroismurteiden sanakirja [Dictionary of Ingrian Dialects]. Helsinki.
7. *Porkka V.* (1885), Ueber den ingrischen Dialekt mit Beruecksichtigung der uebrigen ignermanlaendischen Dialekte [About the Ingrian Dialect Compared with other Ingermanland Dialects]. Helsingfors.

Николаев Илья Сергеевич
Санкт-Петербургский государственный университет (Россия)
Nikolaev Ilya
St Petersburg State University (Russia)
E-mail: i.s.nikolaevi@spbu.ru

А. Ч. Пиперски
A. Ch. Piperski

СРАВНЕНИЕ КОРПУСОВ МЕРОЙ χ^2: СИМВОЛЫ, СЛОВА, ЛЕММЫ ИЛИ ЧАСТЕРЕЧНЫЕ ПОМЕТЫ?

COMPARING CORPORA USING χ^2: CHARACTERS, WORDS, LEMMATA, OR POS TAGS?

Аннотация. В докладе обсуждается, на уровне каких единиц лучше всего производить сравнение корпусов при помощи меры χ^2. На материале подкорпусов Британского национального корпуса сравнивается качество этой меры при анализе 1-, 2- и 3-грамм символов, слов, лемм и частеречных помет различных типов и демонстрируется, что эта мера лучше всего работает на уровне символьных 2-грамм.
Ключевые слова. Сравнение корпусов, Known-Similarity Corpora, χ^2, символьные n-граммы.

Abstract. This paper discusses what kind of linguistic units is best suited for comparing corpora using χ^2. Based on subcorpora from the British National Corpus, I compare the performance of this measure when using 1-, 2-, and 3-ngrams of characters, words, lemmata, and POS tags of different types. The experiment shows that this measure fares best when using character 2-grams.
Keywords. Corpus comparison, Known-Similarity Corpora, χ^2, character ngrams.

Проблема сравнения корпусов активно обсуждается в корпусной лингвистике около 20 лет [Kilgarriff 1997, 2001; Kilgarriff, Rose 1998; Шайкевич 2015; Fothergill et al. 2016]. В работе [Kilgarriff 2001] показано, что из обсуждавшихся к моменту написания статьи мер наиболее успешно работает сравнение частотных списков с помощью χ^2; сама процедура этого сравнения подробно описана в [Schäfer, Bildhauer 2013: 94–95].

Также в статье [Kilgarriff 2001] предлагается способ оценки мер сравнения корпусов при помощи создания так называемых Known-Similarity Corpora (KSC), если взять два заметно различных корпуса C_1 и C_2, можно составить, например, 11 корпусов, первый из которых полностью состоит из фрагментов C_1, второй — на 90 % из фрагментов C_1 и на 10 % из фрагментов C_2, третий — на 80 % из фрагментов C_1 и на 20 % из фрагментов C_2, …, одиннадцатый — полностью из фрагментов C_2. Для такого набора из 11 корпусов можно сделать 660 суждений вида «корпуса 3 и 5 должны быть более похожи друг на друга, чем корпуса 1 и 6» (при этом первая пара сравниваемых корпусов должна находиться внутри отрезка, ограниченного корпусами второй пары). Качество меры сравнения корпусов можно оценить как процент верных суждений такого рода, даваемых этой мерой.

В работах, посвящённых сравнению корпусов, обычно анализируются частотные списки слов (т. е. 1-грамм на уровне слов). В настоящем докладе проверяется, можно ли улучшить качество сравнения, взяв для анализа частотные списки, составленные по другим типам единиц: по 1-, 2- и 3-граммам на уровне символов, слов, лемм и частеречных помет. В качестве источников для построения KSC были взяты подкорпуса Британского национального корпуса, а именно издания Accountancy (acc), The Art Newspaper (art), British Medical Journal (bmj), The Environment Digest (env), The Guardian (gua), Today (tod), использованные ещё в работе [Kilgarriff 2001]. В таблице 1 приводится процент верных суждений для наборов KSC, построенных на всех 15 парах источников, с помощью меры χ^2 по 400 верхним элементам усреднённого частотного списка на различных типах единиц. В заголовках столбцов char обозначает уровень символов, word — уровень слов, lemma — уровень лемм, pos — базовые частеречные пометы (глагол, существительное и т. п.), claws — более дробные частеречные пометы в системе разметки CLAWS5; цифры после этих помет обозначают 1-, 2- и 3-граммы единиц соответствующего уровня.

Таблица 1. **Процент верных суждений для наборов KSC на различных типах единиц**

	char1	char2	char3	word1	word2	word3
acc art	90,15	97,12	98,03	95,91	97,42	89,24
acc bmj	96,21	97,73	96,67	95,15	93,48	96,06
acc env	97,42	99,09	96,21	95,61	98,48	94,55
acc gua	91,36	92,12	98,03	92,42	87,12	83,18
acc tod	98,79	96,67	96,52	97,88	92,12	91,82
art bmj	97,88	96,97	98,64	98,03	96,67	86,36
art env	95,91	98,64	98,64	99,39	97,73	97,27
art gua	88,94	98,64	95,76	94,85	96,97	96,06
art tod	98,33	99,39	99,09	99,55	97,42	90,76
bmj env	97,73	99,39	97,58	98,18	99,7	96,82
bmj gua	96,97	99,55	98,33	96,52	94,7	89,85
bmj tod	99,7	100	98,48	97,58	95,15	88,79
env gua	98,33	99,24	98,79	99,24	99,55	97,27

	char1	char2	char3	word1	word2	word3
env tod	99,55	99,7	99,55	99,55	99,24	96,06
gua tod	91,52	89,09	92,42	89,24	89,7	75,61
Медиана	**97,42**	**98,64**	**98,03**	**97,58**	**96,97**	**91,82**
Ср. арифм.	**95,92**	**97,56**	**97,52**	**96,61**	**95,7**	**91,31**
	lemma1	**lemma2**	**lemma3**	**pos1**	**pos2**	**pos3**
acc art	98,94	93,33	91,67	77,88	90,91	93,33
acc bmj	95,45	96,06	81,97	79,85	83,64	91,97
acc env	99,09	98,03	93,18	92,12	93,03	98,94
acc gua	97,27	92,27	81,52	78,03	90	82,12
acc tod	96,06	88,79	88,33	89,39	91,36	95,76
art bmj	98,79	98,79	88,33	93,94	93,03	96,97
art env	99,85	98,48	95,61	95,15	95,91	95
art gua	98,48	97,42	90,91	90,45	92,27	93,48
art tod	97,73	97,42	94,24	90	97,12	97,42
bmj env	99,7	98,79	96,82	94,55	98,33	98,33
bmj gua	98,03	96,97	87,73	99,39	96,97	93,18
bmj tod	98,03	95,45	94,85	93,48	97,58	96,67
env gua	98,94	99,55	97,73	99,85	99,24	98,18
env tod	99,39	98,03	98,18	100	100	99,85
gua tod	83,64	88,79	84,39	76,82	97,42	93,03
Медиана	**98,48**	**97,42**	**91,67**	**92,12**	**95,91**	**95,76**
Ср. арифм.	**97,29**	**95,88**	**91,03**	**90,06**	**94,45**	**94,95**
	claws1	**claws2**	**claws3**			
acc art	97,12	95,61	92,12			
acc bmj	85,15	85,45	91,06			
acc env	98,33	98,18	98,48			
acc gua	94,09	93,18	91,67			
acc tod	93,33	95,91	96,67			
art bmj	95,3	98,48	98,33			
art env	91,52	97,12	98,33			

	claws1	claws2	claws3			
art gua	90,91	94,55	86,36			
art tod	95,45	95	98,03			
bmj env	98,79	99,55	98,79			
bmj gua	97,27	98,48	97,58			
bmj tod	98,48	98,79	99,55			
env gua	99,55	98,64	99,7			
env tod	99,24	98,18	99,55			
gua tod	90,91	90,91	93,64			
Медиана	**95,45**	**97,12**	**98,03**			
Ср. арифм.	**95,03**	**95,87**	**95,99**			

Значения медианы и среднего арифметического для различных типов единиц показывают, что наиболее успешно сравнение корпусов осуществляется не на уровне слов, как это обычно принято делать, а на уровне символьных 2-грамм. В таблице 2 представлены типы единиц, дающие наиболее качественные результаты сравнения (таблица отсортирована по медиане, при равенстве медиан — по среднему арифметическому).

Таблица 2. **Качество сравнения по различным типам единиц**

Место	Единицы	*n*-граммы	Медиана	Ср. арифм.
1	символы	2	98,64	97,56
2	леммы	1	98,48	97,29
3	символы	3	98,03	97,52
4	частеречные пометы (CLAWS5)	3	98,03	95,99
5	слова	1	97,58	96,61

Таким образом, хотя сравнение корпусов с помощью символьных *n*-грамм и может показаться менее лингвистически содержательным, чем с помощью единиц более высокого уровня, именно оно позволяет получать наиболее качественные результаты сравнения. В будущем планируется проверить полученные выводе на материале языков, от-

личных от английского, а также с использованием других мер сходства корпусов, кроме χ^2.

Литература

1. *Fothergill R., Cook P., Baldwin T.* (2016), Evaluating a topic modelling approach to measuring corpus similarity // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pp. 273–279.
2. *Kilgarriff A.* (1997), Using word frequency lists to measure corpus homogeneity and similarity between corpora, ACL W97–0122.
3. *Kilgarriff A.* (2001), Comparing corpora // International Journal of Corpus Linguistics 6 (1), pp. 97–133.
4. *Kilgarriff A., Rose T.* (1998), Measures for corpus similarity and homogeneity. ACL W98–1506.
5. *Schäfer R., Bildhauer F.* (2013), Web corpus construction. San Rafael: Morgan & Claypool.
6. *Шайкевич А. Я.* (2015), Меры лексического сходства частотных словарей // Труды международной конференции «Корпусная лингвистика–2015». СПб.: СПбГУ, с. 434–442.

References

1. *Fothergill R., Cook P., Baldwin T.* (2016), Evaluating a topic modelling approach to measuring corpus similarity. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pp. 273–279.
2. *Kilgarriff A.* (1997), Using word frequency lists to measure corpus homogeneity and similarity between corpora, ACL W97–0122.
3. *Kilgarriff A.* (2001), Comparing corpora. In: International Journal of Corpus Linguistics 6 (1), pp. 97–133.
4. *Kilgarriff A., Rose T.* (1998), Measures for corpus similarity and homogeneity. ACL W98–1506.
5. *Schäfer R., Bildhauer F.* (2013), Web corpus construction. San Rafael: Morgan & Claypool.
6. *Shaikevich A. Ya.* (2015), Mery leksicheskogo skhodstva chastotnykh slovarej [Measures of lexical similarity for frequency dictionaries]. In: Proceedings of the International Conference "Corpus Linguistics-2015", Saint Petersburg, Russia, pp. 434–442.

Пиперски Александр Чедович
Российский государственный гуманитарный университет
/ НИУ «Высшая школа экономики» (Россия)
Piperski Alexander
Russian State University for the Humanities / National Research University Higher School of Economics (Russia)
E-mail: apiperski@gmail.com

В. И. Подлесская
V. I. Podlesskaya

СТРАТЕГИИ ПЕРЕДАЧИ ЧУЖОЙ РЕЧИ В УСТНОМ ДИСКУРСЕ В СРАВНЕНИИ С ПИСЬМЕННЫМ: ОПЫТ КОРПУСНОГО ИССЛЕДОВАНИЯ[1]

DIRECT AND INDIRECT SPEECH REPORTING STRATEGIES IN SPOKEN AND WRITTEN DISCOURSE: A COMPARATIVE CORPUS STUDY

Аннотация. Качественный и количественный анализ эпизодов передачи чужой речи в спонтанном монологическом дискурсе произведен на материале экспериментального корпуса «Веселые истории из жизни» из коллекции Spokencorpora 2013, включающего 40 рассказов, каждый в двух версиях – устной и письменной. Выделено четыре способа отражения речевого события в тексте, условно упорядоченных по степени экспликации этого события: цитационная конструкция с прямым дейксисом > цитационная конструкция с косвенным дейксисом > описание речевого события без введения чужой речи > полное отсутствие речевого события в тексте. Были размечены способы отражения одного и того же речевого события в устной и письменной версиях рассказа и обнаружена следующая закономерность: при переходе от устной версии к письменной степень экспликации события либо остается той же, либо понижается.
Ключевые слова. Прямая речь, косвенная речь, устный дискурс, корпус.

Abstract. The paper investigates qualitatively and quantitatively speech-reporting strategies in "Funny life-stories" – a subcorpus of the Prosodically Annotated Corpus of Spoken Russian (Spokencorpora 2013) that contains 40 narratives, in spoken and written version each. Four main patterns of integrating a speech event into a narrative were indicated reflecting a decreasing degree of explicating the event: citation in the form of direct speech > citation in the form of indirect speech > describing a speech event without reporting speech > no speech event reflected. Comparing the distribution of these patterns in spoken and written versions of the collected narratives allowed to demonstrate the following tendency: the same speech event in the written version tends to be equally or less explicated than in the spoken version.
Keywords. Direct speech, indirect speech, spoken discourse, corpus.

1. Постановка вопроса

В работе предпринята попытка понять, как канал передачи информации — устный vs. письменный — может влиять на выбор дискурсивной стратегии в живой речи. Полигоном являются способы передачи чужой речи в составе нарратива. Я использую материал корпуса «Веселые истории из жизни» электронной коллекции «Рассказы о сновидениях и другие корпуса звучащей речи» (Spokencorpora 2013). Корпус содержит 40 устных монологов (аудиофайлы с синхронизиро-

[1] РГГУ/РАНХиГС. Работа выполнена при поддержке РНФ, грант 17-18-01184 .

ванными просодически размеченными транскриптами, респонденты от 18 до 60 лет, около 10 000 словоупотреблений), плюс письменные версии этих же рассказов, самостоятельно записанные авторами спустя несколько дней после записи устной версии (около 7000 словоупотреблений). Исследуемый корпус предоставляет уникальный материал для сравнения дискурсивных стратегий — в том числе, стратегий цитации — в устной и письменной речи, поскольку включает устные и письменные версии рассказов одного и того же говорящего, основанные на одном и том же сюжете. Были проанализированы все эпизоды отражения речевого события в составе нарратива, обнаруженные в корпусе: 162 эпизода в устном подкорпусе (далее в тексте — примеры с индексом FS-sp) и 94 эпизода в письменном подкорпусе (примеры с индексом FS-wr). Дальнейшее изложение будет строиться следующим образом: в следующем разделе я кратко обрисую используемый в работе подход к пониманию противопоставления прямой и косвенной речи. В разделе 3 будет представлен сравнительный качественный анализ способов передачи чужой речи в устном и в письменном подкорпусе; в разделе 4 приводятся количественные данные и формулируются выводы.

2. Прямая и косвенная речь как основные стратегии цитации

Главное прагматическое свойство прямой речи состоит в том, что цитация формирует самостоятельную иллокуцию. Вследствие этого, все дейктические элементы цитации ориентированы на внутреннюю коммуникативную ситуацию, а не на внешнюю ситуацию нарратива, в который интегрирована чужая речь (я буду использовать рабочий термин «прямой» дейксис). Другим важным следствием иллокутивной самостоятельности цитации является отсутствие ограничений на тип иллокуции: в прототипических случаях прямой речи возможны типы иллокуций, отличные от сообщения (вопросы, обращения, восклицания, директивы). В прототипических случаях прямой речи в составе цитации становятся возможными так называемые main-clause phenomena «явления главного предложения» (Падучева 1996а:299). Ср. обнаруженные в исследованном корпусе конструкции, непереводимые в режим косвенной речи: вопросы, вводимые дискурсивным маркером а (*А где здесь сервис-центры?; Ну а вы не можете сказать?*); восклицания, вводимые междометиями (*Ой какие попугайчики!*); вопросы с клитизировнной частицей *-то* (*Чего вообще случилось-то?*);

звуковые имитации (*Бээ*); интродуктивные формулы (*Так вот: …*), дискурсивные повторы (*Очень много значит. Очень.*) и др.

Главное прагматическое свойство косвенной речи, напротив, состоит в том, что авторская ремарка и собственно цитация образуют единую иллокуцию. Вследствие этого, все дейктические элементы цитации (личные, временные, пространственные) ориентированы не на внутреннюю коммуникативную ситуацию, а на внешнюю ситуацию нарратива, в который интегрирована чужая речь (для таких случаев я буду использовать рабочий термин «косвенный» дейксис). Другое важное следствие иллокутивной несамостоятельности цитации состоит в том, что в составе цитации затруднены типы иллокуции, отличные от сообщения. Цитация в формате косвенной речи грамматически строится как сентенциальный актант вводящего предиката.

В живой речи реальный формат цитационной конструкции может отклоняться от прототипа. Например, цитация с очевидными признаками прямого личного и временного дейксиса может вводиться подчинительным союзом *что*:

(1) FS28-sp[2]
16. И собственно когда мы отдали наши /биле̲ты,
17. нам \сообщи̲ли:
18. что «\Де̲вушки@,
19. ·а ваш рейс улетел \вчера̲!»

При анализе цитационных конструкций в корпусе было принято следующее рабочее решение: независимо от того, имеются ли в конструкции те или иные отклонения от прототипа, ей приписывается статус «конструкция с прямым дейксисом», если в составе цитации личные местоимения и личные формы глагола употреблены «прямо» (без индексального сдвига), а в отсутствие свидетельств личного дейксиса (или в дополнение к нему) — где в составе цитации имеются main-clause phenomena. В противном случае цитационной конструкции приписывается статус «конструкция с косвенным или неидентифицированным дейксисом.

3. Стратегии цитации в устном и письменном подкорпусах

Для систематического сравнения способов передачи чужой речи в подкорпусах, я выделила все упоминания о речевых событиях в тек-

[2] Индекс примера содержит отсылку к ярлыку текста в составе корпуса. Номера строк соответствуют нумерации строк в рассказе. Об используемой системе дискурсивной транскрипции см. Кибрик, Подлесская (ред.) 2009, Spokencorpora 2013.

стах. Эти речевые события могут быть описаны либо с помощью цитационной конструкции, либо описательно, без введения чужой речи, ср. описание речевого события *Попросил сметаны* и цитационную конструкцию с введением чужой речи *Попросил: «Положите, пожалуйста, сметаны!»*. Письменные и устные версии рассказов были выровнены, с тем, чтобы выявить пары текстовых фрагментов (устный/письменный), отсылающих к одному и тому же речевому событию в сюжетной канве рассказа. Разумеется, были выявлены и ситуации, когда в одном из подкорпусов речевое событие отражено (описательно или с цитацией) а в другом оно вообще не упоминается. Ниже я покажу, в каких форматах может быть представлено одно и то же событие в устной и письменной версиях.

Самая частая ситуация в корпусе (подробное обсуждение количественных данных отложим до следующего раздела) — когда и в устной, и в письменной версии рассказа событие представлено цитационной конструкцией с прямым дейксисом (в том числе, и с лексикограмматичекими и просодическими отклонениями от прототипа), ср. (2а) и (2б) ниже:

(2а) FS03-sp
41. ··· А /она уже^{h} ··· вот ··\тоже на всю \столовую:
42. ·· «Да /\несвежая у нас \сметана^{h}!»

(2б) FS03-wr
а вместо этого интеллигентно прокричала мне в ответ: «Сметана у нас сегодня не свежая!».

Конструкциям с прямым дейксисом в устном подкорпусе могут соответствовать конструкции с косвенным дейксисом в письменном:

(3а) FS18-sp
139. ··· и вдруг ··· Брентон мне \говорит:
140. ··· «Ты значит иди /вперёд¡,
141. ·· я тебя сейчас \догоню.»

(3б) FS18-wr
Брентон сказал мне, чтобы я шла дальше.

Следующий, представленный в корпусе, класс пар: в устном подкорпусе — цитационная конструкция с прямым дейксисом, в письменном — описание речевого события без эксплицитного введения чужой речи:

(4а) FS18-sp
97. А Брентон говорит «Нет-нет-/\нет,

98. будем /\стоя-ать,
99. придёт \автобус.»

(4б) FS18-wr
но Брентон никак не соглашался идти пешком

Очень частотны (см. раздел 4) пары, в которых устная версия представлена цитационной конструкцией с прямым дейксисом, а в письменной версии соответствующее речевое событие вообще отсутствует. Так, в следующем примере в устной версии (5а) в строках 53–56 цитационная конструкция описывает речевое событие — обращение к собеседнице с предложением налить ей вина. В письменной же версии этого эпизода, см. (5б) есть лишь описание сопутствующих действий (сходил, сел, пытался наполнить стаканчик), но собственно речевого события — обращения к собеседнице с предложением — нет:

(5а) FS10-sp
47. \я:
48. «/\Опа!»,
49. ·· но абсолютно не \растерялся!,
50. \подкачиваю к ней,
51. –посмотреть,
52. –что там у неё в –бокале,
53. говорю «/\Детка@,
54. ·· да \что так у тебя мало –налито?,
55. давай я тебе /\побольше налью¡!
56. А то у меня много в \стакане.»,

(5б) FS10-wr
Но в тот момент я не растерялся, сходил еще за алкоголем, занял место рядом с моей очень симпатичной начальницей, вроде как я участвую в общем разговоре, и стал пытаться наполнить ее пластиковый стаканчик, который она держала в руках.

Весьма немногочисленные в устном подкорпусе цитационные конструкции с косвенным дейксисом в подавляющем большинстве случаев в письменной версии также представлены цитационными конструкциями с косвенным дейксисом, и лишь в единичных случаях соответствующие речевые события даны описательно или вообще не упоминаются.

Рассмотренные нами способы отражения речевого события в тексте можно условно упорядочить по степени экспликации этого события. Максимально эксплицированным следует признать представление речевого события в виде цитационной конструкции с прямым

дейксисом. Менее эксплицированными являются конструкции с косвенным дейксисом, затем следуют описания события без введения чужой речи, и, наконец, полное отсутствие речевого события в тексте. Посмотрим теперь, как количественно распределены в устном и письменном корпусах разные степени экспликации речевого события.

4. Количественные данные и выводы

Ниже в таблице 1 представлены сводные данные об частотности способов представления чужой речи в обследованном материале. Эти данные демонстрируют несколько важных тенденций.

Первое. Внедрение чужой речи в нарратив больше свойственно устной речи, чем письменной: в устном подкорпусе обнаруживается почти в два раза больше упоминаний о речевом событии, чем в письменном — 162 против 94.

Второе. Для устной речи прямое цитирование более характерно, чем для письменной. В устном подкорпусе доля конструкций с прямым дейксисом от общего числа упоминаний о речевом событии существенно выше: в устном подкорпусе их 143 из 162 (88.3 %), а в письменном — 57 из 94 (60.6 %).

Третье. При переходе от устной версии к письменной степень экспликации события либо остается той же, либо понижается.

Для наглядности в таблице 1 затемнены клетки там, где уровень экспликации речевого события сохраняется или снижается при переходе от устной версии к письменной.

Как хорошо видно из таблицы, если некоторое речевое событие передается в устной версии рассказа в виде конструкции с прямым дейксисом, то в письменной версии оно, как правило, тоже передается как прямая цитация (52 случая из 143) или вообще не имеет соответствия (46 из 143), реже оно передается в виде косвенной цитации (23 из 143) или описательно (22 из 143). Если в устной версии мы имели косвенную цитацию, то в письменной преимущественно — тоже косвенную (13 из 19).

Нарушения данной закономерности крайне редки и они, как правило связаны с нарушением прототипа прямого или косвенного цитирования. Приведу в пример случай, когда в устной версии упоминания речевого события нет, а в письменной обнаруживается непрототипическое прямое цитирование: письменная версия (6б) содержит отсутствующую в устной версии (6а) мотивировку — анонимное об-

щее мнение, поданное как чужая речь без указания на конкретного говорящего. Обращает на себя внимание намеренное употребление пишущим кавычек для оформления прямой цитаты:

(6а) FS02-sp
9. /Мы решили заехать на /территорию /\университетаw.
10. Но /там висит ↑\–кирпи-ич.
11. ···· \Во-от-т.
12. Но /так как я всё время /нарушаю,
13. то /мы решили
14. что под /кирпич \можно прое-ехать.

(6б) FS02-wr
Мы решили поехать на территорию университета. Там весит знак кирпич! Решили поехать, так как «туда все ездят… почему бы не поехать».

Таким образом корпусные данные позволяют заключить, что устная речь в большей степени ориентирована на непосредственное отображение речевых событий в монологическом дискурсе: они отражаются в устной версии рассказов чаще, чем в письменной, и с большей степенью экспликации.

Таблица 1. **Экспликация речевого события в устном и письменном подкорпусах**

			Письменный			
			Цитационные конструкции (94)		**Описание речевого события**	**Эпизод отсутствует**
Устный			Прямой дейксис (57)	Косвенный и смешанный дейксис (37)		
	Цитационные конструкции (162)	Прямой (143)	**52**	**23**	**22**	**46**
		Косвенный и смешанный дейксис (19)	1	**13**	**1**	**4**
	Описание речевого события		2			
	Эпизод отсутствует		2	1		

References

1. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009), Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow, Jazyki Slavjanskix Kul'tur.
2. *Paducheva E. V.* (1996a), Subjectivnaja modal'nost': illokutivnye pokazateli i vvodnye slova [Subjective modality: illocutive markers and parentheticals]. In: Semantičeskie issledovanija. Semantika vremeni i vida v Russkon jazyke. Semantika narrativa [Semantic investigations Semantics of tense and aspect in Russian. Semantics of narrative.] Jazyki russkoj kul'tury, Moscow, pp. 297–320.
3. Spokencorpora (2013), Prosodically Annotated Corpus of Spoken Russian (PrACS-Russ). Pilot version. Available at: http://spokencorpora.ru

---

Подлесская Вера Исааковна
Российский государственный гуманитарный университет /
Российская академия народного хозяйства и государственной службы (Россия)
Podlesskaya Vera
Russian State University for the Humanities /
Russian Academy of National Economy and Public Administration (Russia)
E-mail: vi_podlesskaya@il-rggu.ru

А. Е. Поляков
A. E. Polyakov

ГРАММАТИЧЕСКИЙ СЛОВАРЬ ЦЕРКОВНОСЛАВЯНСКОГО ЯЗЫКА (ПО МАТЕРИАЛАМ КОРПУСА)[1]

GRAMMATICAL DICTIONARY OF CHURCH SLAVONIC (CORPUS-BASED)

Аннотация. Грамматический словарь церковнославянского (ЦС) языка создан на основе корпуса ЦС текстов, который является частью Национального корпуса русского языка. Словарь включает около 30 000 лексем, охватывающих около 150 000 уникальных словоформ. Каждая лексема и словоформа снабжена грамматической информацией и имеет связь с корпусом. Словарь и грамматическая модель ЦС изменения созданы эмпирически в результате анализа сходных словоформ и грамматических моделей.
Ключевые слова. Корпусная лингвистика, церковнославянский язык, грамматический словарь.

Annotation. A grammatical dictionary of Church Slavonic (CS) is based of the corpus of CS texts within the Russian National Corpus. The dictionary contains about 30 000 lexemes covering about 150 000 different wordforms. Each lexeme and wordform is provided with grammatical information and has links to the corpus. The dictionary and the grammatical model of CS infexion are built empirically as a result of the analysis of similar wordforms and grammatical patterns.
Keywords. Corpus linguistics, Church Slavonic, grammatical dictionary.

Современный церковнославянский (ЦС) язык, используемый в богослужебной практике, до сих пор не имеет адекватного научного описания. Существующие грамматики и словари дают идеализированную картину, которая часто не соответствует фактическому состоянию языка, отраженному в текстах. Грамматические описания ЦС языка нередко копируются со старославянского и не учитывают особенностей его функционирования в среде живых славянских языков. Практически единственное современное описание [Гаманович 1964] является слишком кратким и не затрагивает многих сложных вопросов словоизменения. В результате порой невозможно понять, как реально изменяется то или иное слово, если в грамматике оно не описано, а в словаре для него даны только избранные примеры.

Существующие словари ЦС языка обычно преследуют не описательные, а нормативные цели и отражают лишь некоторое подмножество лексики, например устаревшие и непонятные слова, церковные термины, собственные имена — [Алексеев 1815], [Дьяченко 1900] и др.

[1] Работа выполнена при поддержке РГНФ (проект 17-04-12064 «Разработка модулей НКРЯ для автоматической разметки и словарной поддержки старорусских и церковнославянских текстов»)

Множество ЦС слов включены в академические словари русского языка [САР 1789; СЦРЯ 1847] с пометой «церковное», поскольку в русской языковой среде ЦС воспринимался как особый «высокий» стиль, а не как отдельный язык.

Создаваемый грамматический словарь ЦС языка является не нормативным, а дескриптивным и строится на основе обширного корпуса ЦС текстов, снабженных специальной разметкой. Данный корпус является частью проекта «Национальный корпус русского языка» (http://ruscorpora.ru/search-orthlib.html). Корпус включает около 1200 текстов, которые охватывают все основные типы и жанры ЦС литературы (богослужебные, святоотеческие, писание и др.), и имеет объем более 4 млн словоупотреблений. Корпус такого объема вполне репрезентативен с точки зрения охвата лексики и различных жанрово-тематических групп текстов. Подробнее о ЦС корпусе см. [Поляков 2014; Добрушина, Поляков 2013; Добрушина, Кравецкий, Поляков 2015].

Словарь представляет собой список лексем и словоформ с приписанной им грамматической информацией (лемма, часть речи, грамматические признаки, номер парадигмы и др.). Словарь создается индуктивно, на основе анализа реальных словоформ, представленных в корпусе. Вначале из корпуса генерируется список словоформ и делается его первичная проверка и чистка. Затем делается ручная лемматизация для частотных слов, определяются типичные шаблоны и строится первичная модель словоизменения. Затем на основе этой модели специальная программа генерирует разборы для остальных слов, затем эти разборы проверяются вручную, снова уточняется грамматическая модель, пока не будут проанализированы все слова.

Параллельно с созданием словаря строится грамматическая модель ЦС словоизменения: грамматические таблицы, список парадигм, состав грамматических признаков, правила порождения словоформы по лемме, список чередований и др. Таблицы парадигм записываются в формальном виде, пригодном для компьютерного анализа. Все тонкости словоизменения указываются в эксплицитной форме, в виде формальных правил.

В отличие от традиционных грамматик, парадигмы не задаются априорно, а выводятся эмпирически на основе анализа множества словоформ, имеющих однотипные наборы флексий. В результате номенклатура парадигм получается значительно более детальной и может существенно отличаться от традиционных грамматик. Так, традиционное второе склонение (рабъ) распадается на более чем 10 типов в за-

висимости от конечной согласной, наличия беглой гласной и других особенностей; глаголы имеют около 40 словоизменительных типов.

Грамматический словарь включает 150 тыс. словоформ, которые группируются в 30 тыс. лексем. Лексемы в словаре даны только в тех формах, которые реально представлены в корпусе. Для лексем, относящихся к частотным грамматическим типам, можно породить недостающие формы, но это ненадежно. Многие лексемы в корпусе представлены единичными примерами, так что для них нельзя точно установить исходную форму и грамматические признаки.

Грамматический словарь доступен в Интернете по адресу: http://dic.feb-web.ru/slavonic/dicgram/.

Словарная статья в словаре имеет следующую структуру:

1) лемма (словарная форма) и ее варианты (если есть);
2) грамматические признаки лексемы (часть речи, род, одушевленность, вид, переходность);
3) код парадигмы;
4) список всех словоформ с указанием их грамматических признаков (число, падеж, время, лицо и др.), а также их частоты в корпусе.

Для устаревших и малопонятных слов может даваться краткое толкование или перевод, который не претендует на полноту, а служит просто для информации.

Словарь связан с корпусом через гипертекстовые ссылки: при нажатии на лемму или словоформу вы можете увидеть все примеры ее употребления в корпусе. При нажатии на код парадигмы вы попадаете на список парадигм. В настоящее время словарь используется для лексико-грамматического поиска в корпусе, в дальнейшем предполагается его более глубокая интеграция в корпус.

Словарь, построенный на базе и интегрированный в корпус, позволяет устранить неполноту и неточность грамматических описаний и дает надежную лексическую и грамматическую базу для создания новых словарей и грамматик ЦС языка. Так, он может использоваться как грамматическая база для нового словаря ЦС языка, который создается в Церковнославянском центре Института русского языка.

Литература

1. *Алексеев П.* (1815), Церковный словарь. Ч. I–IV. М., 1815–1816.
2. *Алипий* (*Гаманович*) (1964), Грамматика церковно-славянскаго языка. Jordanville (N. Y.).

3. *Добрушина Е. Р., Поляков А. Е.* (2013), Корпус церковнославянского языка: возможности, методы создания, перспективы // Вестник ПСТГУ. Серия III: Филология. Вып. 1 (31), с. 32–44.
4. *Добрушина Е. Р., Кравецкий А. Г., Поляков А. Е.* (2015), Корпус и частотный грамматический корпусный словарь церковнославянского языка в составе НКРЯ // Национальный корпус русского языка: 10 лет проекту. Труды Института русского языка им. В. В. Виноградова. Вып. 6. М., с. 116–141.
5. *Дьяченко Г.* (1900), Полный церковно-славянский словарь. М.
6. СЦРЯ (1847), Словарь церковно-славянского и русского языка, составленный Вторым отделением Императорской Академии наук. Т. I–IV. СПб.
7. САР (1789), Словарь Академии Российской. Ч. I–VI. СПб., 1789–1794.
8. *Поляков А. Е.* (2014), Корпус церковнославянских текстов: проблемы орфографии и грамматики // Przeglad wschodnioeuropejski. V. 1, s. 245–254.

References

1. *Alekseev P.* (1815), Cerkovnyj slovar' [Church Slavonic dictionary]. Ch. I–IV. Moscow, 1815–1816.
2. *Alipij* (*Gamanovich*) (1964), Grammatika cerkovno-slavjanskago jazyka [Grammar of Church Slavonic]. Jordanville (N. Y.).
3. *Dobrushina E. R., Poljakov A. E.* (2013), Korpus cerkovnoslavjanskogo jazyka: vozmozhnosti, metody sozdanija, perspektivy [Church Slavonic corpus: facilities, methods of creation. perspectives]. In: Vestnik PSTGU. Serija III: Filologija, vyp. 1 (31), pp. 32–44.
4. *Dobrushina E. R., Kraveckij A. G., Poljakov A. E.* (2015), Korpus i chastotnyj grammaticheskij korpusnyj slovar' cerkovnoslavjanskogo jazyka v sostave NKRJa [Corpus and frequency grammatical dictionary of Church Slavonic]. In: Nacional'nyj korpus russkogo jazyka: 10 let proektu // Trudy Instituta russkogo jazyka im. V. V. Vinogradova. Vyp. 6. Moscow, pp. 116–141.
5. *D'jachenko G.* (1900), Polnyj cerkovno-slavjanskij slovar' [Complete Church Slavonic dictionary]. Moscow.
6. SCRJa (1847), Slovar' cerkovno-slavjanskogo i russkogo jazyka [Dictionary of Church Slavonic and Russian]. T. I–IV. SPb.
7. SAR (1789), Slovar' Akademii Rossijskoj [Dictionary of the Russian Academy]. Ch. I–VI. SPb., 1789–1794.
8. *Poljakov A. E.* (2014), Korpus cerkovnoslavjanskix tekstov: problemy orfografii i grammatiki [Corpus of Church Slavonic: problems of spelling and grammar]. In: Przeglad wschodnioeuropejski, vol. 1, pp. 245–254.

Поляков Алексей Евгеньевич
ООО «МАРС» (Россия)
Polyakov Alexey
ООО «MARS» (Russia)
E-mail: pollex@mail.ru

Е. В. Рахилина, В. П. Фесенко
E. V. Rakhilina, V. P. Fesenko

ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ КОРПУСА ТЕКСТОВ XIX ВЕКА В ЛИНГВИСТИЧЕСКОМ ИССЛЕДОВАНИИ[1]

OPPORTUNITIES FOR CORPUS OF XIX CENTURY TEXTS USAGE IN LINGUISTIC RESEARCH

Аннотация. Доклад посвящен Корпусу текстов XIX века, созданному в Школе лингвистики НИУ ВШЭ. Корпус специальным образом размечен и содержит информацию о лингвистических единицах, которые изменили значение или форму в период с XIX до XXI век. Рассматривается возможность применения данных корпуса для микродиахронических исследований в области синтаксиса (в частности, проблемы выбора родительного/ винительного падежа объекта при переходном глаголе с отрицанием), а также небольших лексикосемантических сдвигов (ср. *не уметь, нездоров*) и др.
Ключевые слова. Корпус, XIX век, отрицание, изменение семантики, изменение сочетаемости, историческая грамматика, русский язык.

Abstract. The paper describes the Corpus of XIX-th century texts (School of Linguistics, RSU HSE). The Corpus has a sophisticated system of annotation and contains information about linguistic items that have changed their meaning and/or form since XIX-th century. The paper discusses how the Corpus can be used as an additional linguistic tool for microdiachronicre- search on syntactic problems (like genitive of negation) and minor lexical semantic shifts (cf. *ne umet'* 'to be hopeless at doing smth', *nezdorov* 'healthless').
Keywords. Corpus, XIX century, negation, semantic change, historical grammar, Russian.

Проект «Корпус текстов XIX века» стартовал в сентябре 2016 года на базе Школы лингвистики НИУ ВШЭ под руководством Е. В. Рахилиной. Объектом разметки Корпуса текстов XIX века являются морфологические, лексические и синтаксические конструкции, обнаруживающие различия между нормами русского языка XIX и XXI веков. Цель создания корпуса — составление базы данных текстов XIX века и разметка в них языковых единиц разного рода, отличающих языковые нормы XIX и XXI веков. Данные корпуса могут быть применены для поиска и описания этих единиц, включая конструкции. Ближайшим результатом работы над корпусом может стать составление списка устаревших конструкций XIX века, который в дальнейшем может лечь в основу соответствующего словаря. Информация из корпуса мо-

[1] При поддержке гранта РНФ 16-18-02071 «Пограничный русский: оценка сложности восприятия русского текста в теоретическом, экспериментальном и статистическом аспектах».

жет использоваться как для научного описания нормы русского языка в диахронии, так и как помощь в понимании текстов, написанных два века назад.

Составление базы началось с разметки текста романа М. Ю. Лермонтова «Герой нашего времени» (объем — 43 566 слов). На сегодняшний день в корпусе размечены или находятся на стадии разработки роман И. А. Гончарова «Обыкновенная история» (97 868), повесть И. С. Тургенева «Ася» (13 931) и рассказ «Мой сосед Радилов» (2 289). Помимо художественных, база корпуса будет включать тексты других стилей.

С помощью корпуса возможен поиск по типам конструкций, изменившихся с XIX века, а также по конкретным словам и конструкциям. Такой поиск позволяет находить примеры для исследования языковых конструкций, а также отмечать новые лингвистические сюжеты для изучения истории языка.

Иллюстративным материалом к докладу послужит использование Корпуса текстов XIX века для решения комплекса вопросов, связанных с отрицанием:

— проблема объектного генитива при отрицании;
— изменение функционирования сочетания *не умеет;*
— изменение семантики краткого прилагательного *нездоров.*

1. Выбор падежа объекта при отрицании

Наиболее известная проблема русистики, связанная с отрицанием, — выбор падежа объекта при переходном глаголе с отрицанием. Об отклонениях от существовавшей с XVIII века нормы замены винительного падежа на родительный при отрицании с научной точки зрения первым начать говорить А. И. Томсон в начале XX века [Томсон 1902]. После этого на протяжении XX века и до сего дня она привлекает внимание теоретиков лингвистики (см., например, [Борщев, Парти 1998; Объектный генитив 2008]). Однако о неполном соответствии грамматической нормы выбору падежа в реальной речи писал еще А. С. Пушкин в первой половине XIX века (см. [Винокур 1959]). Корпус текстов XIX века с разметкой конструкций, отличающихся от принятых в современном русском языке, позволяет по-новому подойти к исследованию этой проблемы.

Поиск в Корпусе текстов XIX века по помете *genneg* дает конструкции с родительным падежом при отрицании, в которых современный

носитель языка выбрал бы винительный падеж. Например: *Она покраснела и не хотела назвать дня, вспомнив свою милую выходку* («Княжна Мери»). В современных исследованиях объект *день* рассматривается как референтный и требует винительного падежа. В докладе подробно, с привлечением материала НКРЯ, обсуждается процесс изменения именных ограничений на эту отрицательную конструкцию, связанных с референциальным статусом существительного.

2. Сочетаемость выражения *не уметь*

Помимо классической проблемы падежа при отрицании, интересную динамику показывает конструкция *не уметь*. Поиск в Корпусе текстов XIX века дает следующие контексты, отличающиеся от современных: *Ведь этакой народ! — сказал он, —* и *хлеба по-русски назвать не умеет, а выучил: «офицер, дай на водку!»* («Бэла»); *Так иногда отличный анатомик не умеет вылечить от лихорадки; Он изучал все живые струны сердца человеческого, как изучают жилы трупа, но никогда не умел он воспользоваться своим знанием; — Вы также переменились, — отвечала она, бросив на него быстрый взгляд, в котором он не умел разобрать тайной насмешки* («Княжна Мери»). Можно заметить, что все зависящие от *не умеет* глаголы — совершенного вида. Дальнейший анализ примеров из НКРЯ показывает, что в XIX веке доля примеров с сочетанием *не уметь + уСВ* значительно выше, чем в XXI веке: в среднем 43 % из всех сочетаний *не уметь + V*. В XXI веке этот средний показатель равен 14 %. Притом эти 14 % примеров принадлежат авторам, родившимся не позднее 1969 года, воспитанным на литературе более раннего периода.

Любопытно отметить, что в XIX веке достаточно частотно использование в качестве зависимого от *не уметь* слова существительного *грамота* в дательном падеже, что сегодня соответствует сочетанию *не обучен*: *он не умеет грамоте;* есть также маргинальный контекст с предложным управлением: *не умеет в грамоте*. Кроме того, в XIX веке встречаются контексты с зависимым от *не уметь* придаточным изъяснительным предложением, равным сегодняшнему *не знать*: *не умел с чего начать; не умею, как вас назвать; что сказать, он не умеет* и пр.

Примечательно также, что в XIX веке было возможно опущение (с точки зрения современной нормы) глагола *говорить* при *не уметь* в конструкции со значением владения иностранным *языком:* не умеет по-немецки, по-барски не умею, не умел по-русски ни слова.

Обсуждается, как в дальнейшем общее значение слова *уметь* распределилось между глаголами *мочь, знать* и *быть обученным.*

3. Семантика краткого прилагательного *нездоров*

Интересным объектом микроисторического исследования может быть краткое прилагательное *нездоров.* В Корпусе XIX века встречаем следующий пример: *Печорин был долго нездоров, исхудал, бедняжка; только никогда с этих пор мы не говорили о Бэле: я видел, что это ему будет неприятно, так зачем же?* («Бэла»). В этом примере *был долго нездоров* «в переводе» на современный русский язык означает *долго болел.*

Анализ примеров НКРЯ показывает, что в XIX веке *нездоров* было полным синонимом слова *болен.* Так можно было сказать и о человеке, у которого разболелась голова или начался насморк, но и о человеке, страдающем оспой, лежащем на смертном одре или раненом. Допустимым было сочетание *нездоров* с наречиями *очень, сильно.* Любопытно, что слова *нездоров* и *болен* не различались по своей семантико-синтаксической сочетаемости с существительным: *болен простудой* и *нездоров флюсом, нездоров сильным кашлем.*

В XXI веке *нездоров* встречается в двух основных контекстах. Прежде всего, это небольшое недомогание: *Но я нынче нездоров, Мне что-то тяжело, пойду засну.* [Музыкальные паузы режиссера Васильева (2003) // «Театральная жизнь», 2003.05.26]; *Элмар был нездоров и хотел отдохнуть…* [Армен Медведев. Территория кино (1999–2001)]; *Он слегка нездоров и расстроен…* [Александр Проханов. Господин Гексоген (2001)].

Второй контекст, в котором сегодня используется *нездоров,* — психиатрический: *психически нездоров, нездоров душевно.*

Другими словами, у краткого прилагательного *нездоров* ушло значение тяжелой физической болезни, раны и под.

Как видно из приведенных примеров, разметка Корпуса текстов XIX века может быть очень полезной для выявления и дополнительного исследования «проблемных точек» в языке XIX века — тех единиц и конструкций, семантика или функционирование которых в той или иной степени отличаются от современных. Дальнейший, более глубокий анализ этих отличий может быть проведен с помощью обращения к более широкому материалу Национального корпуса русского языка.

Литература

1. *Борщев В. Б., Парти Б.* (1998), Бытийные предложения и отрицание в русском языке: семантика и коммуникативная структура // Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям (под ред. А. С. Нариньяни). Казань, Хетер, с. 173–182.
2. *Винокур Г. О.* (1959) Пушкин и русский язык (Сб. «А. С. Пушкин. 1837–1937», М., 1937) // Избранные работы по русскому языку. М.: Учпедгиз, с. 189–206.
3. Объектный генитив при отрицании в русском языке (2008) [Ред. кол.: А. Б. Летучий, Е. В. Рахилина, Т. И. Резникова; Сост. Е. В. Рахилина]. М.: Пробел — 2000. 176 с. (Исследования по теории грамматики; Вып. 5).
4. *Томсон А. И.* (1902), Винительный падеж прямого дополнения в отрицательных предложениях в русском языке. Отд. отт. из «Русского Филологического Вестника». Варшава: Типография Варшавского Учебного Округа. 43 с.

References

1. *Borshchev V. B., Partee B.* (1998), Bytijnye predlozhenija i otricanie v russkom jazyke: semantika i kommunikativnaja struktura. [Existential sentences and negation in Russian: semantics and communicative structure]. In: Trudy Mezhdunarodnogo seminara Dialog'98 po komp'juternoj lingvistike i ee prilozhenijam (pod red. A. S. Narin'jani). [Incoll. Proceedings of the International seminar on computer linguistics and its applications Dialogue'98 (ed. A. S. Narin'jani)]. Kazan', Heter, pp. 173–182.
2. *Vinokur G. O.* (1959) Pushkin i russkij jazyk (Sb. «A. S. Pushkin. 1837–1937», M., 1937) [Pushkin and Russian language (Coll. "A. S. Pushkin. 1837–1937", Moscow, 1937)]. In: Izbrannye raboty po russkomu jazyku [Selected works on Russian language]. Moscow: Uchpedgiz, pp. 189–206.
3. *Ob ektnyj genitive pri otricanii v russkom jazyke* (2008) [Object-Genitive under Negation in Russian (2008)]. Red. kol.: A. B. Letuchij, E. V. Rahilina, T. I. Reznikova; Sost. E. V. Rahilina [Ed. staff: A. B. Letuchij, E. V. Rahilina, T. I. Reznikova; Comp. E. V. Rahilina]. Moscow: Probel — 2000, 176 p. (Issledovanija po teorii grammatiki; Vyp. 5). [Research on theory of grammar; edition 5].
4. *Tomson A. I.* (1902), Vinitel'nyj padezh prjamogo dopolnenija v otricatel'nyh predlozhenijah v russkom jazyke [Accusative case of direct object in Russian]. Otd. ott. iz «Russkogo Filologicheskogo Vestnika» [Separate print from "Russian Filological Herald"]. Varshava: Tipografija Varshavskogo Uchebnogo Okruga. 43 p.

Рахилина Екатерина Владимировна
Rakhilina Ekaterina
E-mail: rakhilina@gmail.com

Фесенко Вера Павловна
Fesenko Vera
E-mail: verun4ik_18@mail.ru

Национальный исследовательский университет
«Высшая школа экономики» (Россия)
National Research University Higher School of Economy (Russia)

Е. И. Риехакайнен
E. I. Riekhakaynen

КОРПУС ТРАСКРИБИРОВАННЫХ РУССКИХ УСТНЫХ ТЕКСТОВ: ТЕКУЩИЕ ВОЗМОЖНОСТИ И ПЕРСПЕКТИВЫ

CORPUS OF TRANSCRIBED RUSSIAN SPEECH: CURRENT OPTIONS AND CHALLENGES

Аннотация. В статье на примере исследования, посвященного оценке сохранности фонетического облика словоформ в начале межпаузальных интервалов в русской спонтанной речи, демонстрируются возможности Корпуса транскрибированных русских устных текстов и намечаются перспективы его дальнейшего развития, наиболее актуальные из которых (помимо расширения Корпуса) – это создание новых уровней аннотирования (в частности, уровня идеальной транскрипции) и добавление новых поисковых возможностей.
Ключевые слова. Корпусы устной речи, русская устная речь, редукция, восприятие речи.

Abstract. The article describes a phonetic study of word forms at the beginning of interpausal intervals in spontaneous Russian that allowed to demonstrate how the Corpus of Transcribed Russian Speech can currently be used and developed. For the moment, new levels of annotation (such as ideal transcription) and new search options are the most required innovations.
Keywords. Spoken corpora, Russian speech, reduction, spoken word recognition.

1. Корпус транскрибированных русских устных текстов

Статья продолжает серию публикаций, посвященных принципам создания и использования Корпуса транскрибированных русских устных текстов (далее — Корпус), разрабатываемого сотрудниками Санкт-Петербургского государственного университета. В предыдущих статьях ([Венцов и др. 2013; Венцов и др. 2015] и др.) мы использовали различные варианты названия нашего корпуса — Корпус русской устной речи, Речевой корпус, Корпус русских спонтанных текстов. Однако, как представляется, именно формулировка «Корпус транскрибированных русских устных текстов» наилучшим образом отражает как его содержание (в нем в настоящее время представлены расшифровки не только спонтанной речи, но и, например, радиосводок Ю. Б. Левитана), так и особенности аннотирования (Корпус до сих пор остается единственным известным нам общедоступным корпусом русской речи, снабженным полной фонетической транскрипцией)[1].

Основная цель создания Корпуса — его последующее применение для моделирования восприятия естественной звучащей речи. Именно

[1] Более подробное сопоставление Корпуса с другими корпусами русской устной речи представлено в [Венцов и др. 2013: 224].

этим обусловлены те принципы аннотирования (включая обязательное наличие сплошной фонетической расшифровки), которые приняты в Корпусе (о необходимости столь подробного аннотирования см., например, в [Tucker et al. 2016]). Далее в статье будут упомянуты только те принципы аннотирования, которые являются первостепенными для исследования, о котором пойдет речь в Разделе 2.

2. Исследование сохранности фонетического облика словоформ в начале межпаузальных интервалов

2.1. Гипотеза

Экспериментальные свидетельства в пользу того, что контекст является ведущим фактором при распознавании редуцированных словоформ в естественной русской речи (см. [Риехакайнен 2016]), а также тот факт, что в момент восприятия конкретного фрагмента устной речи слушающему доступен только левый контекст, позволяют предположить, что в начальных фрагментах речевой цепи редукция должна встречаться реже, поскольку отсутствует непосредственный левый контекст, за счет которого могли бы быть восстановлены редуцированные элементы.

2.2. Методика исследования

Сформулированную выше гипотезу решено было проверить на материале одной из записей, входящих в Корпус, а именно на материале расшифровок радиопередачи «Утренний гость» (при создании конкорданса при поисковом запросе в онлайн-версии Корпуса фрагменты этого текста обозначаются guest01.wav).

Во всех текстах Корпуса на данный момент осуществлена разметка на межпаузальные интервалы, при этом различаются паузы вдоха (inh), паузы вздоха (sigh), придыхание (aspir), гортанная смычка (gst(e)) и собственно паузы (paus(e) или p). Поиск по словам (в орфографии и в транскрипции) и по типам пауз доступен по адресу: http://narusco.ru/search/trn-search.php.

Из Корпуса были отобраны и проанализированы все фонетические слова, которые были употреблены непосредственно после любой из пауз. В распоряжении исследователей[2] при этом была расшифровка

[2] Отбор и первичную обработку результатов осуществлял студент СПбГУ Александр Сергеевич Смирнов под руководством автора статьи.

в виде связного текста, но поставленную задачу можно решить и с помощью онлайн-поиска, задавая последовательно все виды пауз и отбирая контексты из записей, обозначенных guest01.wav.

В выборку вошли 579 контекстов. Дальнейшее исследование заключалось в анализе сохранности фонетического облика начальных фонетических слов в каждом из случаев.

2.3. Результаты

В целом полученные результаты свидетельствуют в пользу подтверждения сформулированной гипотезы. Только в 169 (29,2 %) из 579 контекстов, попавших в выборку, было зафиксировано выпадение или изменение звуков в фонетическом слове в начале межпаузального интервала, т. е. сразу после паузы. При этом в 142 из них редукции подвергались не более двух звуков. Подавляющее большинство из этих примеров — это словоформы с редукцией на конце слов или в заударной флексии. Такие реализации могут быть однозначно восстановлены до полных или благодаря сохранности большей части словоформы (например, *которую*[3]*, украденную* и др.), или благодаря ближайшему правому контексту, в котором содержится информация, указывающая на морфологические признаки словоформы (например, *вчерашний ребёнок и* др.), т. е. левый контекст не играет в распознавании подобных реализаций ключевой роли.

Уже на данном этапе описания результатов возникают методологические вопросы: что считать идеальными произнесением и как оценить полученный процент редуцированных реализаций в начале межпаузальных интервалов? Первый вопрос в рамках данного исследования был решен самым простым способом — конкретные реализации сопоставлялись с максимально полным произнесением, однако очевидно, что в ряде случаев даже прескриптивная орфоэпическая норма будет отличаться от такого варианта произнесения (см. подробнее об этом в [Риехакайнен 2016: 72]). Чтобы каждому конкретному исследователю в дальнейшем не приходилось сталкиваться с этой проблемой, предполагается в будущем дополнить аннотирование Корпуса уровнем идеальной транскрипции, которая будет создана автоматически на основе орфографической расшифровки текста. Что касается второго вопроса, то для того, чтобы определить, что сло-

[3] Жирным шрифтом здесь и далее в примерах выделены звуки, подвергшиеся полной количественной редукции.

воформы в начале межпаузальных интервалов действительно менее склонны к редукции, чем в других позициях, необходимо знать статистику количества редуцированных словоформ по всему Корпусу. На настоящее время эти подсчеты можно провести вручную, однако создание уровня идеальной транскрипции существенно упросит решение этого вопроса.

Подробный анализ тех 27 случаев, в которых выпадению или качественному изменению подвергались более двух элементов, представлен в [Смирнов 2017], в рамках же данной статьи будет упомянута только одна группа из них, а именно те случаи, когда для надежного распознавания редуцированной словоформы требовалось обращение к левому контексту, который предшествовал паузе (например, *работ**ают***, *Влади**мирович***). Такие примеры в очередной раз демонстрируют, что паузы в спонтанной речи могут возникать не только на границе клауз (синтагм), но и внутри них. В результате поиска по всем типам пауз в выборку попали оба типа случаев. Сформулированная же выше гипотеза должна быть подтверждена прежде всего на примерах первого типа (т. е. на границах клауз), поскольку в случае т.н. «расчлененных» синтагм единство элементов, их образующих, может достигаться вопреки паузам — за счет грамматических средств. Благодаря проведенным ранее исследованиям и наличию Базы «расчлененных» дискурсивных единиц (http://narusco.ru/EDU_BASE) мы имели данные о границах клауз в анализируемом тексте. В дальнейшем подобная информация может быть также интегрирована в основную разметку Корпуса.

Поскольку во многих из проанализированных случаев редукция затронула окончания (в первую очередь прилагательных и глаголов), следующим шагом в исследовании распознавания восприятия речи должно стать моделирование того, каким образом слушающий восстанавливает редуцированную морфологическую информацию. Однако для исследований подобного рода необходимо иметь возможность поиска по различным морфологическим параметрам или хотя бы по частям речи, для чего необходимо аннотирование Корпуса на морфологическом уровне.

Таким образом, проведенное исследование показало, каким образом можно использовать уже имеющиеся возможности Корпуса, а также позволило наметить наиболее актуальные пути развития Корпуса, которые, как мы надеемся, будут воплощены в жизнь в ближайшее время.

Литература

1. *Венцов А. В., Нигматулина Ю. О., Раева О. В., Риехакайнен Е. И., Слепокурова Н. А.* (2013), Корпус русских спонтанных текстов: структура и единицы // Корпусная лингвистика — 2013: Труды международной научной конференции. СПб., с. 223–230.
2. *Венцов А. В., Нигматулина Ю. О., Раева О. В., Риехакайнен Е. И., Слепокурова Н. А.* (2015), От корпуса устной речи к базе «расчлененных» дискурсивных единиц // Корпусная лингвистика — 2015: Труды международной научной конференции. СПб., с. 154–161.
3. *Риехакайнен Е. И.* (2016), Восприятие русской устной речи: контекст + частотность. СПб.
4. *Смирнов А. С.* (2017), Без контекста: редукция в начале межпаузальных интервалов // Проблемы порождения и восприятия речи. Материалы XIV выездной школы-семинара. Череповец, в печати.
5. *Tucker B., Ernestus M.* (2016), Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon // The Mental Lexicon, 11 (3), pp. 375–400.

References

1. *Ventsov A. V., Nigmatulina Yu.O., Raeva O. V., Riekhakaynen E. I., Slepokurova N. A.* (2013), Korpus russkih spontannyh tekstov: struktura i edinitsy [Corpus of Russian spontaneous texts: structure and items]. In: Korpusnaja lingvistika — 2013: Trudy mezhdunarodnoj nauchnoj konferentsii. [Corpus linguistics — 2013: Proceedings of the international scientific conference]. St Petersburg, pp. 223–230.
2. *Ventsov A. V., Nigmatulina Yu.O., Raeva O. V., Riekhakaynen E. I., Slepokurova N. A.* (2015), Ot korpusa ustnoj rechi k baze "raschlenennyh" diskursivnyh edinits [From a speech corpus to a database of "broken" discourse units]. In: Korpusnaja lingvistika — 2015: Trudy mezhdunarodnoj nauchnoj konferentsii. [Corpus linguistics — 2015: Proceedings of the international scientific conference]. St Petersburg, pp. 154–161.
3. *Riekhakaynen E. I.* (2016), Vosprijatie russkoj ustnoj rechi: kontekst + chastotnost' [Recognition of Russian speech: context + frequency]. St Petersburg.
4. *Smirnov A. S.* (2017), Bez konteksta: reduktsija v nachale mezhpauzalnyh intervalov [Without context: reduction in the beginning of interpausal intervals]. In: Problemy porozhdenija i vosprijatija rechi. Materialy XIV vyezdnoj shkoly-seminara [Problems of speech production and recognitions. Proceedings of the 14th workshop]. Cherepovets, to appear.
5. *Tucker B., Ernestus M.* (2016), Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. In: The Mental Lexicon, 11 (3), pp. 375–400.

Риехакайнен Елена Игоревна
Санкт-Петербургский государственный университет (Россия)
Riekhakaynen Elena
Saint Petersburg State University (Russia)
E-mail: e.riehakajnen@spbu.ru

А. Ройтберг, К. Христосова
A. Roitberg, K. Hristosova

РАЗМЕТКА БРИДЖИНГ-АНАФОРЫ НА МАТЕРИАЛЕ РУССКОЯЗЫЧНЫХ ТЕКСТОВ[1]

BRIDGING ANAPHORA ANNOTATING FOR RUSSIAN

Аннотация. В работе описывается первый опыт аннотирования бриджинг-анафоры для русского языка. Нами была проведена разметка 190 коротких новостных текстов. Каждый текст размечался тремя аннотаторами. Как и ожидалось, сложные дискурсивные явления вызывают трудности у разметчиков, возникает много расхождений в разметке. В статье приводятся примеры наиболее сложных для разметки случаев, обсуждаются типичные ошибки аннотаторов.
Ключевые слова. Бриджинг-анафора, разметка анафоры, корпусная лингвистика.

Abstract. We describe the first experience of bridging anaphora annotating for Russian. We have annotated 190 texts. Each texts was annotated by three annotators. As anticipated, it was very difficult to annotate complicate discourse phenomena, so there are a lot of disagreements between different annotations. We try to show most complicated cases and typical mistakes of annotators and discuss the reasons of the discrepancy.
Keywords. Bridging-anaphora, anaphora annotating, corpus linguistics.

1. Введение

Проблема восстановления анафорических отношений возникает при решении практически всех задач прикладной лингвистики, связанных с работой над текстом. В последние годы появляются корпуса, содержащие разметку не только прямой, но и бриджинг-анафоры. См., например, [Hou 2013; Korzen et al; Lüngen 2007; Nedoluzko et al. 2009; Poesio 2000]. Термин «бриджинг-анафора» был предложен в ставшей классической работе [Clark 1975]. Кларк предложил разделить анафору на прямую, при которой анафорически связанные элементы кореферентны, и непрямую, где анафорически связанные элементы не кореферентны. Для непрямой анафоры был предложен термин бриджинг.

(1) Мы подошли к дому. Стены были увиты плющом[2].

[1] РФФИ 5-07-09306 «Стандарты оценки методов автоматического извлечения информации из текстов»

[2] Во всех примерах мы будем выделять бриджинг-элемент полужирным шрифтом, а якорь подчеркивать.

Слова стены и дом анафорически связаны, так как имеются в виду стены именно того дома, который упомянут ранее, при этом данные выражения не кореферентны.

В работах о бриджинге для выражения, к которому идет отсылка, принят термин якорь; для выражения, нуждающегося в толковании через связь с предыдущим текстом, нет общепринятого термина, мы будем использовать термин бриджинг-элемент.

2. Принципы построения русского бриджинг-корпуса

Для русскоязычного бриджинг-корпуса RuGenBridge был применен новый принцип отбора бриджинг-пар, названный генитивным бриджингом. Как бриджинг размечались пары элементов, связанные анафорически и способные образовывать грамматичную генитивную конструкцию.

(2) Теперь я познакомился с отцом. Сын, оказывается, был на него очень похож.

Слова отец и сын связаны анафорически. Кроме того, они могут образовывать генитивную конструкцию отец сына. Gen, подробнее [Roitberg, Nedoluzhko 2016].

В нашем корпусе мы постулируем бриджинг-отношения между бриджинг-элементом и всей кореферентной цепочкой якоря.

(3) Дом весь покосился, было видно, что ремонтом здания давно никто не занимался. Дверь держалась на одной петле.

В данном случае слово дверь анафорически связано со словом здание, которое в свою очередь кореферентно, со словом дом. Мы рассматриваем бриджинг-связь между бриджинг-элементом дверь и якорной цепочкой дом — здание. Для всех якорей были размечены кореферентные цепочки.

Заметим, что аннотаторам давалась инструкция соединять бриджинг-элемент с линейно ближайшем участником якорной цепочки, т. е. в Примере 3 правильная разметка: дверь — здание. Тем не менее, при сравнении разметок двух аннотаторов совпадение пары засчитывалось, если совпали бриджинг-элементы, а якоря кореферентны.

Также отметим, что корпус RuGenBridge создавался как обучающий и тестовый корпус для системы автоматического распознавания бриджинга. Цель создания корпуса определяла те решения, которые принимались относительно выбора стандарта разметки.

Первоначально корпус был размечен одним аннотатором, но летом 2016 года была проведена дополнительная разметка: 178 текстов было размечено двумя, а 190 текстов — тремя аннотаторами.

3. Разметка корпуса несколькими аннотаторами

Разметка явлений верхнего уровня языка всегда трудна, так как требует от аннотатора, с одной стороны, довольно большого объема специальных знаний, а с другой стороны, большой внимательности, так как нужно анализировать не какое-то конкретное слово или словосочетание, а сравнительно большой отрывок текста. Задача особенно усложняется, если для данного явления еще нет четкого определения и общепринятой классификации типов. Думаем, это стало причиной того, что в нашей работе наблюдалось значительное расхождение между аннотаторами. В итоге удалось достигнуть приемлемого уровня согласия между первым и вторым аннотаторами (F-мера = 0,71); согласие между третьим и первым аннотаторами оказалось очень низким (F = 0,23), поэтому разметка третьего аннотатора в дальнейшей работе практически не учитывалась. Для оценки согласия мы использовали F-меру, так как более распространенная метрика — «каппа» не подходит для сравнения разметок редких явлений.

4. Анализ расхождений разметок разных аннотаторов

При разметке бриджинг-анафоры возникают несколько основных типов ошибок: 1) пропуск бриджинг-пары; 2) разметка пары, не являющейся примером бриджинга; 3) неправильный выбор якоря для некоторого бриджинг-элемента. Кроме непосредственно ошибок могут возникать и расхождения между аннотаторами.

4.1. Пропуски бриджинг-пар при разметке

Пропуски правильных пар возникали у всех разметчиков, но есть случаи, в которых пропуски возникают чаще, чем в других.

Так, один аннотатор часто допускал пропуски в ситуации, когда бриджинг-элемент и якорь расположены линейно близко. В таком случае анафорическая связь кажется очевидной, но по формальным критериям бриджинг-отношения должны размечаться:

(4) <…> первый номер списка партии Чалый в видеообра-щении к жителям Севастополя заявил <…>

Кроме того, заметное количество пропусков возникало в ситуации, когда элементы бриджинг-отношений расположены далеко друг от друга.

4.2. Разметка «лишних» пар

Наибольшее количество ошибок вызывала проверка того, могут ли анафорически связанные слова образовывать генитивную конструкцию.

Например, пары, состоящие из названия области/края и т. п., включающего в себя родовое слово и названия страны, обычно могут образовывать генитивную конструкцию: Московская область России. По аналогии разметчики ошибочно проводили связь от названий городов к стране, хотя в этих случаях генитивная конструкция уже невозможна.

(5) Силам безопасности Ирака удалось отразить атаку боевиков на завод в Байджи.

Также вызывали трудности случаи, когда близко друг от друга встречались потенциальные бриджинг-элементы, один из которых подходил под генитивный критерий, а другой — нет:

(6) взорвался автобус, сообщают свидетели происшествия. По предварительным данным, жертв и пострадавших нет

Аннотатор провел две связи: жертвы — происшествия и пострадавших — происшествия, но жертвы происшествия — возможно, а *пострадавшие происшествия — нет.

4.3. Разные, но кореферентные якоря

Как упоминалось выше, согласно принципам разметки нашего корпуса, аннотаторы должны были связывать бриджинг-элемент с линейно ближайшим к нему якорем. Но иногда разметчики пропускали линейно ближайший элемент.

(7) Тольяттиазота, в случае если не соглашусь на условия по продаже предприятия$_1$. <…> их цель рейдерский захват предприятия$_2$. <…> не легче ли было бы выкупить долю миноритариев.

Первый аннотатор провел связь к линейно ближайшему якорю: миноритариев — предприятия$_2$, в то время как второй аннотатор провел связь миноритариев — Тольятиазота

4.4. Разногласия, вызванные ошибками понимания

В небольшом количестве случаев расхождение разметок были вызваны разницей в понимании текста.

(8) Страны G7 готовят план поддержки Украины в случае перебоев с поставками газа этой зимой. <…>. В среду 7 мая страны G7 договорились сократить энергетическую зависимость от России.

В Примере 8 один аннотатор разметил связь энергетическую зависимость — страны G7, в то время, как другой аннотатор разметил связь энергетическую зависимость — Украины.

Интересно, что система автоматического обнаружения бриджинга, обученная на нашем корпусе, предложила для нескольких бриджинг-элементов два различных якоря, каждый из которых отражал один из возможных вариантов понимания текста.

(9) в Instagram Папы Римского появилась фотография понтифика, обнимающего двух девочек, предположительно, с синдромом Дауна, с желто-голубой лентой в руках.

Система разметила два варианта связи: руках — девочек и руках — понтифика. Оба понимания текста являются допустимыми.

5. Заключение

Данная работа показала сложность разметки дискурсивных явлений. Для улучшения согласия аннотаторов необходимо внести небольшие уточнения в инструкцию для аннотаторов, а главное — проводить более серьезную теоретическую подготовку разметчиков. Тем не менее нам удалось добиться приемлемого уровня согласия аннотаторов и сделать обучающий корпус для системы автоматического распознавания бриджинга.

References

1. *Clark H. H. Bridging* (1975), Proceedings of the 1975 workshop on Theoretical issues in natural language processing, Association for Computational Linguistics, pp.169–174.
2. *Hou Y., Market K., Strube M.* (2013), Cascading Collective Classification for Bridging Anaphora Recognition using a Rich Linguistic Feature Set. EMNLP 2013.
3. *Korzen I., Buch-Kromann M.* (2011), Anaphoric relations in the Copenhagen dependency treebanks, S. Dipper and H. Zinsmeister, editors, Corpus-based Investigations of Pragmatic and Discourse Phenomena, Bochumer Linguistische Arbeitsberichte, volume 3.

4. *Lüngen H.* (2008), RRSet-Taxonomy of rhetorical relations in SemDok, Interne Reports der DFG-Forschergruppe.
5. *Nedoluzhko A., Mírovský J., Ocelák R., Pergler J.* (2009), Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank.
6. *O'Reilly T., McNamara D. S.* (2007), Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers, Discourse Processes, Vol. 43(2).
7. *Poesio M.* (2000), Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results, Proc. of the 2nd LREC.
8. *Roitberg A., Nedoluzhko A.* (2016), Bridging Corpus for Russian in comparison with Czech, Coreference Resolution beyond OntoNotes, p. 59.

Ройтберг Анна Михайловна
ИМПБ РАН — филиал ИПМ им. М. В. Келдыша РАН (Россия)
НИУ ВШЭ (Россия)
Roitberg Anna
IMPB RAS — Branch of KIAM RAS (Russia)
NRU HSE (Russia)
E-mail: cvi@yandex.ru

Христосова Ксения
НИУ ВШЭ (Россия)
Hristosova Ksenia
NRU HSE (Russia)

С. О. Савчук
S. O. Savchuk

УСТНАЯ ПУБЛИЧНАЯ РЕЧЬ В МУЛЬТИМЕДИЙНОМ МОДУЛЕ НКРЯ[1]

ORAL PUBLIC SPEECH IN MULTIMODAL MODULE OF THE RUSSIAN NATIONAL CORPUS

Аннотация. В статье представлен опыт создания специализированных мультимедийных корпусов, включающих устную научную и устную политическую речь в разнообразии их жанровых разновидностей и в различных ситуациях произнесения.
Ключевые слова. Устный научный дискурс, устный политический дискурс, Мультимедийный корпус русского языка.

Abstract. The article deals with the project compiling specialized multimodal corpora of oral Russian scientific and political discourse in a variety of genres and in different situations. The composition of the corpus, technology of database constructing, types of annotation and the interface are described, as well as the prospects of its use in studying oral public discourse.
Keywords. Oral scientific discourse, oral political discourse, Multimodal Russian Corpus.

Введение

Устная научная и политическая речь, а также другие, менее изученные в лингвистике виды коммуникации (деловая, правовая, рекламная, церковная) составляют сферу публичной речи, основными признаками которой являются ориентация на публичную аудиторию, социально-значимая тематика, относительная подготовленность (неспонтанность), языковая и стилистическая организованность. Несмотря на разделяемое многими понимание дискурса как речевой деятельности, протекающей в определенном социальном пространстве, объектом изучения в большинстве исследований становятся письменные источники — либо опубликованные тексты, либо стенограммы или транскрипты устных выступлений (см., например, [Лаптева 1985], [Лаптева 2007], [Паршина 2012]). Между тем устное исполнение публичной речи, даже при условии ее предварительной подготовки, нельзя приравнять к печатной публикации или стенограмме, поскольку при письменной фиксации теряются средства актуализации смыслов и эмоционального воздействия на слушателя, которыми облада-

[1] Исследование выполнено при финансовой поддержке РФФИ, проект № 15-06-04334 и программы ОИФН РАН «Евразийское наследие и его современные смыслы».

ет устная речь. Если еще недавно лингвисты не имели инструмента, позволяющего фиксировать и анализировать устную русскую речь в единстве с манерой и обстоятельствами ее произнесения, то в настоящее время таким инструментом стал мультимедийный корпус в составе НКРЯ (МУРКО). Инструментарий корпуса позволяет изучать как фонетические и интонационные особенности тех или иных конструкций, так и жестикуляционное сопровождение устной речи. В настоящее время в составе МУРКО активно формируется коллекция текстов, относящихся к устной научной и политической речи. Состав коллекции и особенности ее представления в корпусе описаны в настоящей статье.

Состав корпусов и источники текстов

Формирование текстового состава мультимедийных корпусов устной научной и политической речи осуществляется на основе отбора наиболее частотных жанров и типовых ситуаций, представляющих разновидности научного и политического дискурса. При этом используются только общедоступные источники, размещенные в интернете, а также собственные записи и материалы, предоставленные партнерами, среди которых Фонотека ИРЯ РАН, видеоархив МГУ им. М. В. Ломоносова, проект «Устная история» и др. Материал представлен в виде аудио- и видеозаписей с текстовыми расшифровками.

Коллекция записей *устной научной речи* (УНР) обширна и разнообразна по составу говорящих (их более 100), областей науки (естественные, гуманитарные, прикладные), временному диапазону, по разнообразию жанрового состава. Ведущими монологическими жанрами, представленными в корпусе, являются *доклад, лекция;* диалогическими — *дискуссия, семинар.* Разновидности научного дискурса — *специальная научная речь*, *учебная научная речь*, *научно-популярная речь* — обусловлены различными ситуациями общения. Для каждой разновидности характерны своя иерархия жанров и жанровые варианты, а также специфические жанры, не представленные в других разновидностях. В частности, для учебной научной речи наиболее частотными жанрами являются лекции, семинары и школьные уроки. Для научно-популярной речи — научно-популярные лекции, беседы и интервью, рассказы, комментарии, прежде всего в СМИ.

В настоящее время УНР представлена в составе Мультимедийного корпуса лекциями («Публичные лекции “Полит.ру”» и проект

Academia на телеканале «Культура»), а также беседами с ведущими учеными (телепередача «Гордон»). Это обширный интересный материал (объемом более 800 тыс. словоупотреблений), на базе которого выполнен ряд исследований (см., в частности, [Гришина 2015]). Текущая работа по формированию корпуса УНР состоит в увеличении жанрового разнообразия текстов и коррекции баланса. С этой целью в состав корпуса включены записи учебных лекций по математике, химии, информатике, экономике, правоведению, психологии, искусствоведению и др., предназначенные для студентов и школьников, записи мастер-классов и семинаров. Для пополнения коллекции научно-популярной разновидности собраны записи телепередач и интернет-проектов научно-просветительского характера: «Идеи, которые изменили мир», «Истории из будущего», «Вопрос науки», «ЕХперименты», интервью на канале «Эксперт-ТВ», лекторий «Прямая речь», «Постнаука» и др. (см. табл. 1).

Таблица 1. **Планируемое пополнение подкорпуса научной речи**

Жанры	Специальная научная речь (тыс. с/у)	Научно-популярная речь (тыс. с/у)	Учебно-научная речь (тыс. с/у)	Кол-во словоуп. (тыс. с/у)
Доклад	51	35	10	96
Лекция	24	10	66	100
Семинар	17		14	31
Дискуссия	8			8
Ток-шоу		20		20
Интервью		55		55
Всего	**100**	**120**	**90**	**310**

Подкорпус *устной политической речи* (УПР) создается впервые. Главная задача состит в том, чтобы при ограниченном объеме корпуса отразить в нем вариативность политического дискурса по нескольким параметрам. Разнообразие по *составу говорящих* достигается включением в корпус выступлений представителей как власти, так и оппозиции, руководителей федерального и местного уровня, представителей разных ветвей власти, общественных деятелей. Для отражения вариативности по типу *ситуаций общения* и *тематики* в корпус включены обсуждения вопросов внутренней и внешней политики в самых раз-

ных ситуациях. Типологически их можно объединить в три группы: 1) выступления профессиональных политиков в профессиональной аудитории, в СМИ, на встречах с избирателями и пр.; 2) выступления политологов, обозревателей и комментаторов в СМИ; 3) выступления на митингах, встречах с представителями власти, с кандидатами в период предвыборных кампаний и пр. Жанровое разнообразие достигается включением в корпус основных монологических и диалогических жанров политического дискурса: *речь, доклад, комментарий, с* одной стороны, *и интервью, пресс-конференция, парламентские слушания, круглый стол, дискуссия, совещание, встречи с избирателями* — с другой (см. табл. 2).

Таблица 2. **Планируемый состав подкорпуса политической речи**

Жанры		**Кол-во (тыс. с/у)**
Монологические	доклад	20
	лекция	20
	речь	38
	комментарий	20
Диалогические	интервью	82
	пресс-конференция	50
	парламентские слушания	25
	круглый стол	25
	дискуссия	30
	совещание	10
	ток-шоу	70
	встречи с избирателями, дебаты	60
Всего		**450**

Технология организации базы данных и аннотация

Технологическая цепочка опирается на опыт создания Мультимедийного корпуса, подробно описана в работах [Гришина 2007], [Гришина 2015], [Савчук, Махова 2017] и сводится к следующим этапам. Каждая аудио-/видеозапись разрезается на небольшие фрагменты (клипы). На соответствующие фрагменты разрезаются расшифровки этих фрагментов. После этого клип и расшифровка выравниваются между собой. Клипы и текстовые фрагменты снабжаются одинаковы-

ми именами, при этом нумерация должна строго совпадать, поскольку на этом основан поиск соответствующих фрагментов. На запрос пользователя фрагменты текста, которые снабжены клипами.

Разметка текстовых фрагментов осуществляется по стандартам НКРЯ, а также в соответствии со стандартами мультимедийного корпуса. Каждый текстовый фрагмент снабжен метатекстовой, морфологической, семантической, акцентологической аннотацией; каждой реплике приписана социологическая аннотация (сведения о говорящем). Орфоэпическая разметка и разметка акцентной структуры слова осуществляются в процессе индексации в автоматическом режиме.

Доступ к текстам устной научной и политической речи осуществляется по адресу http://ruscorpora.ru/search-murco.html. В дальнейшем планируется внести изменения в интерфейс МУРКО, которые сделают возможным отбор для исследовательских целей пользовательских подкорпусов научной и политической речи, сопоставимых по объему, по однородности состава (диалоги или монологи), жанровой и тематической принадлежности.

Создаваемые специализированные корпуса предоставят возможность исследовать 1) естественную устную русскую речь в мультимедийной форме; 2) особенности диалогической и монологической устной речи; 3) жанровое своеобразие и вариативность устного научного и политического дискурса в зависимости от тематики, типа аудитории (выступление на профессиональной конференции, в учебной аудитории, в СМИ), индивидуальной манеры говорящих; 4) своеобразие научного и политического монолога и диалога в сопоставительном аспекте, сравнивая их по разным параметрам, например жанровому (доклад на научной конференции и на заседании правительства, интервью с политиком и с ученым, научная дискуссия и политические дебаты и т. д.).

Литература

1. *Гришина Е. А.* (2015), Мультимодальный модуль в составе Национального корпуса русского языка // Труды Института русского языка им. В. В. Виноградова, № 6, с. 65–88.
2. *Гришина Е. А.* (2015), О русском жестикуляционном отрицании // Труды Института русского языка им. В. В. Виноградова, № 6, с. 556–604.
3. *Гришина Е. А.* (2009) Мультимедийный русский корпус (МУРКО), проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., с. 150–174.

4. *Лаптева О. А.* (2007), Живая русская речь с телеэкрана. Разговорный пласт телевизионной речи в нормативном аспекте. Изд. 6-е. М.: Издательство ЛКИ.
5. *Лаптева О. А.* (ред.). (1985), Современная русская устная научная речь. Общие свойства и фонетические особенности. Т. I. Красноярск: Изд-во Красноярского ун-та.
6. *Паршина О. Н.* (2012), Российская политическая речь: Теория и практика Изд. 3-е. М.: Книжный дом "ЛИБРОКОМ".
7. *Савчук С. О., Махова А. А.* (2017), Мультимедийный модуль в составе Национального корпуса русского языка: направления развития. Анализ разговорной русской речи (AP^{3}–2017) // Труды седьмого междисциплинарного семинара. СПб.: Политехника-принт, с. 83–89.

References

1. *Grishina E. A.* (2015), Mul'timodal'nyy modul' v sostave Natsional'nogo korpusa russkogo yazyka [Multimodal module as part of the Russian National Corpus]. In: Trudy Instituta russkogo yazyka im. V. V. Vinogradova [Proceedings of the V. V. Vinogradov Russian Language Institute], 6, pp. 65–88.
2. *Grishina E. A.* (2015), O russkom zhestikulyatsionnom otritsanii [On gestural negation in the Russian language]. In: Trudy Instituta russkogo yazyka im. V. V. Vinogradova [Proceedings of the V. V. Vinogradov Russian Language Institute], 6, pp. 556–604.
3. *Grishina E. A.* (2009), Mul'timediynyy russkiy korpus (MURKO), problemy annotatsii [Multimodal Russian Corpus (MURCO), Problems of Annotation]. In: Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy [The National Corpus of the Russian Language: 2006–2008. New results and perspectives]. St Petersburg, Nestor-Istoriya Publ., pp. 150–174.
4. *Lapteva O. A.* (2007), Zhivaya russkaya rech' s teleekrana. Razgovornyi plast televizionnoi rechi v normativnom aspekte [Russian speech from a television screen. Conversational layer of television speech in the normative aspect]. 6th ed. Moscow, LKI Publ.
5. *Lapteva O. A.* (ed.) (1985), Sovremennaya russkaya ustnaya nauchnaya rech'. Obshchie svoystva i foneticheskie osobennosti [Modern Russian oral scientific speech. General properties and phonetic features]. Vol. 1. Krasnoyarsk, Krasnoyarsk Univ. Press.
6. *Parshina O. N.* (2011), Rossiyskaya politicheskaya rech': Teoriya i praktika [Russian Political Speech: Theory and Practice]. 3nd ed. Moscow, LIBROKOM Publ.
7. *Savchuk S. O., Makhova A. A.* (2017), Mul'timediynyi modul' v sostave Natsional'nogo korpusa russkogo yazyka: napravleniya razvitiya [Multimedia module in the Russian National Corpus: directions of development]. In: Sed'moi mezhdistsiplinarnyi seminar "Analiz razgovornoy russkoy rechi" (AR^{3}–2017) [7th interdisciplinary seminar "Russian Speech Analysis"]. St Petersburg: Politehnika-print Publ., pp. 83–89.

Савчук Светлана Олеговна
Институт русского языка им. В. В. Виноградова РАН
Savchuk Svetlana
V. V. Vinogradov Russian Language Institute, RAS
E-mail: savsvetlana@mail.ru

С. Ю. Семенова
S. Yu. Semenova

ОБ ИСПОЛЬЗОВАНИИ ДАННЫХ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА ДЛЯ ИЛЛЮСТРИРОВАНИЯ СТАТЕЙ КОМПЬЮТЕРНОГО СЕМАНТИЧЕСКОГО СЛОВАРЯ[1]

RUSSIAN NATIONAL CORPUS AS RESOURCE OF TEXT EXAMPLES FOR THE AUTOMATED SEMANTIC DICTIONARY

Аннотация. Рассматриваются вопросы использования корпусных данных в качестве иллюстративного материала в компьютерной лексикографии. Речь идет о пополнении зоны иллюстраций прикладного формализованного семантического словаря РУСЛАН примерами из НКРЯ. Работа осуществляется в рамках модернизации словаря, сформировавшегося, в основном, на рубеже 1990–2000-х гг. под руководством Н. Н. Леонтьевой при участии автора. Помимо онтологического обновления иллюстраций, предполагается структурирование иллюстративной зоны словарной базы данных (ранее в ней не практиковавшееся) – выделение полей, прицельно иллюстрирующих конкретные составляющие словарного описания (поля, акцентирующие способы насыщения конкретных валентностей, показывающие функционирование несвободных словосочетаний и нек. др.). Осуществляется определенная разметка цитат из корпуса, нацеленная на выделение наиболее информативных фрагментов предложений для иллюстрирования конкретных лексем.
Ключевые слова. Использование корпусных данных в лексикографии, иллюстрирование словарных статей, компьютерный семантический словарь.

Abstract. Using of corpus data in computational lexicography is under consideration. Sentences from Russian National Corpus are being loaded into the illustration zone of the NLP-aimed semantic dictionary of RUSLAN. This automated formal dictionary created on the whole in 1995–2005 is being enlarged and renovated at present. Modernization of the dictionary includes the illustration zone division into a number of fields and ranging of segments of the chosen corpus sentences as more or less typical contexts of a word described.
Keywords. Corpus data in computational lexicography, corpus examples in a dictionary, semantic dictionary for NLP, illustrational zone structure.

Работа посвящается применению корпусных данных в компьютерной лексикографии, а именно, использованию массивов НКРЯ при иллюстрировании семантического словаря для автоматического анализа текста. Имеется в виду русский компьютерный словарь РУСЛАН, создававшийся на рубеже 1990-х—2000-х гг. в НИВЦ МГУ под руководством Н. Н. Леонтьевой при участии автора [Леонтьева, Семенова 2002; Леонтьева, Семенова 2003].

[1] Работа осуществляется при поддержке РФФИ (РГНФ), Проект РГНФ № 17-04-00594 «Автоматический словарь РУСЛАН: обновленная концепция, новая лексика» (2017—2019гг.).

Словарь в значительной мере структурирован и формализован; статьи составлены на символьном метаязыке, разработанном Н. Н. Леонтьевой. Словник, насчитывающий около 12 тыс. единиц, вначале был ориентирован на обработку общественно-политических текстов, затем стала добавляться общелитературная лексика.

В настоящее время назрела потребность модернизации словаря РУСЛАН, при сохранении, в основном, его структуры и метаязыка. В задачи модернизации входят расширение и статистически выверенное выравнивание словника, уточнение представления полисемии и валентностей, добавление формализованных описаний ситуаций, обозначаемых предикатными словами, а также развитие зоны текстовых иллюстраций.

Изначально словарные статьи не иллюстрировались в базе данных словаря (а лишь в бумажных черновиках), затем база стала пополняться примерами, преимущественно модельными, показывающими, в основном, различие лексических значений полисемичных слов и насыщение валентностей [Семенова 2003]. При этом поле текстовых примеров было без внутренней структуры. Такое иллюстрирование имело место в тот период, когда отечественная корпусная лингвистика только начинала формироваться.

Далее предполагается существенное обновление иллюстрирующих контекстов на основе корпусных данных. Главными источниками иллюстрирования являются основной и газетный подкорпусы НКРЯ, хотя не исключается обращение и к другим коллекциям и массивам, в том числе к блогам. Потребность в подборе нового иллюстративного материала обусловлена, в том числе, некоторым онтологическим устареванием дискурса рубежа веков, отраженным в прежней версии словаря (как устаревшие воспринимаются некоторые встречающиеся в примерах политические реалии, количественная фактография, биографические данные и т. п.).

Наряду с обновлением и расширением иллюстраций предусматривается структуризация самой иллюстративной зоны, разделение ее на поля, демонстрирующие разные аспекты поведения слова (лексемы), заполнение валентностей (с указанием основной иллюстрируемой валентности и с демонстрированием основных отраженных в словарном описании семантических классов актантов), показ нерегулярных конструкций, реальное употребление фразеологизмов в контексте, интересные примеры с нечетко выделяемыми лексемами, показ полных и усеченных вариантов цитат. Основная стратегия — выбор кон-

текстного окружения, наиболее типичного для лексемы (с перспективой использования примеров при дизамбигуации). Предполагается определенная экспериментальная работа с корпусным материалом, нацеленная на выработку принципов отбора и разметки выбираемых предложений (предполагается привлечение информантов для оценки релевантности и «прототипичности» корпусных примеров).

Опора на НКРЯ имеет место и при других работах по модернизации данного словаря; в частности, при учете частотности вхождений в корпус при выравнивании словника, при определении преимущественного состава и таксономии актантов. На основе реального корпусного материала вырабатывается прагматически значимое для прикладного семантического словаря дробление слов на лексемы.

Литература

1. *Леонтьева Н. Н., Семенова С. Ю.* (2002), Об отражении полисемии в прикладном семантическом словаре // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог '2002'. Протвино, 6–11 июня 2002 года. М.: Наука. Т. 2, с. 489–496.
2. *Леонтьева Н. Н., Семенова С. Ю.* (2003), Семантический словарь РУСЛАН как инструмент компьютерного понимания // Понимание в Коммуникации. Материалы научно-практической конференции. 5–6 марта 2003 г. М.: МГГИИ, с. 41–46.
3. *Семенова С. Ю.* (2003), Примеры в компьютерном семантическом словаре: некоторые наблюдения над процессом подбора // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003 (Протвино, 11–16 июня 2003 г.) М., с. 593–598.

References

1. *Leonteva N. N., Semenova S. Yu.* (2002), Ob otrazhenii polisemii v prikladnom semanticheskom slovare [On meaning ambiguation representation in the applied semantic dictionary]. In: Kompyuternaya lingvistika i intellektualnye tekhnologii. Trudy mezhdunarodnogo seminara Dialog '2002'. Protvino, 6–11 iyunya 2002 goda. [Computational Linguistics and Intellectual Technologies: Proceedings of International Conference "Dialogue'2002". Protvino, June 6–11.2002]. Moscow, vol. 2, pp. 489–496.
2. *Leonteva N. N., Semenova S. Yu.* (2003), Semanticheskiy slovar RUSLAN kak instrument kompyuternogo ponimaniya [Semantic dictionary of RUSLAN as a tool for automatic semantic analysis]. In: Ponimanie v Kommunikatsii. Materialy nauchno-prakticheskoy konferentsii. 5–6 marta 2003 g. [Understanding in Communication: Proceedings of Scientific and Practical Conference. March 5–6. 2003]. Moscow, pp. 41–46.
3. *Semenova S. Yu.* (2003), Primery v kompyuternom semanticheskom slovare: nekotorye nablyudeniya nad protsessom podbora [Text examples in the automated semantic dic-

tionary: some reflections on the procedure of choice]. In: Kompyuternaya lingvistika i intellektualnye tekhnologii. Trudy Mezhdunarodnoy konferentsii Dialog'2003 (Protvino, 11–16 iyunya 2003 g.) [Computational Linguistics and Intellectual Technologies: Proceedings of International Conference "Dialogue'2003". Protvino, June 11–16.2003]. Moscow, pp. 593–598.

Семенова Софья Юльевна
ИНИОН РАН, РГГУ (Россия)
Semenova Sophia
INION RAS, RSUH (Russia)
E-mail: sonya_sem@mail.ru

В. Д. Соловьев, В. В. Бочкарев, Л. А. Янда
V. D. Solovyev, V. V. Bochkarev, L. A. Janda

ДИНАМИКА ЧАСТОТ УПОТРЕБЛЕНИЯ СЕМАНТИЧЕСКИ БЛИЗКИХ СЛОВ[1]

FREQUENCY DYNAMICS OF SEMANTICALLY CLOSE WORDS

Абстракт. В работе рассматривается новая модель установления семантической тождественности слов через сопоставление частот их словоупотребления на большом временном интервале. Частоты словоупотребления берутся по корпусу Google Books Ngram. Модель применяется к проблематике аспектуальных пар. Показано, что корреляция частот словоупотребления аспектуальных пар с префиксальным образованием глагола совершенного вида выше, чем аспектуальных пар с суффиксальным образованием глагола несовершенного вида. Это неожиданный результат, так как исходя из современной теории русского вида, следовало ожидать обратного.
Ключевые слова. Русский язык, аспект, частота слов, корпус n-грамм.

Abstract. We propose a new model for establishing the semantic identity of words by comparing the frequencies of their usage over a large time interval. The word frequencies are taken from the Google Books Ngram corpus. The model is applied to the problem of aspectual pairs. It is shown that the correlation of the word frequencies for aspectual pairs formed via prefixation (with simplex imperfectives and prefixed perfectives) is higher than that of aspectual pairs formed via suffixation (with secondary imperfectives). This is an unexpected result, because from the modern theory of the Russian aspect, one would expect the opposite.
Keywords. Russian language, aspect, word frequency, corpus of n-gram.

1. Введение

Корпус Google Books Ngram позволяет анализировать и сопоставлять динамику частот словоупотреблений и снабжен удобной визуализацией в виде графиков частот (https://books. google.com/ngrams/). В основополагающей работе [Michel et al. 2011] и ряде последующих показаны возможности использования Google Books Ngram для изучения эволюции, как отдельных слов, так и всего лексикона языка в целом, а также культурных трендов в обществе.

Кроме собственно частот словоупотреблений, весьма информативным является форма графиков. Естественным является предположение, что частоты употребления семантически идентичных слов (например, словоизменение внутри одной леммы: *читать — читал*) под влиянием внешних факторов меняются схожим образом, т. е. графики частот имеют схожую форму.

[1] Работа выполнена при поддержке РФФИ, грант № 15-06-07402

Аспектуальная система русского языка находится в процессе становления, многие ее элементы получили неоднозначное освещение в литературе и вызвали большие споры. Неясным остается характер суффиксального образования вторичных имперфективов и префиксального образования перфективов с точки зрения словообразовательного или словоизменительного характера этих процессов. В данной работе изучается степень синхронности изменения частот слов в аспектуальных парах, как один из подходов к рассматриваемой проблеме. Графики частот трактуются как временные ряды. Мы используем терминологию по аспекту, принятую в [Janda et al. 2013].

2. Данные и методы

Мы сравниваем корреляции частот словоупотребления внутри следующих групп слов — глаголов: 1) случайно выбранные (в качестве baseline), 2) словоизменительные формы внутри одной леммы (например, *читать — читал*), 3) аспектуальные пар с префиксальным образованием перфективов (*делать — сделать*), 4) аспектуальных пар с суффиксальным образованием вторичных имперфективов (*убаюкать — убаюкивать*). Словоизменительная парадигма была взята по открытому ресурсу OpenCorpora (http://opencorpora.org/), основанному на словаре Зализняка. Случайная выборка включала 20 тыс. пар словоформ из парадигм 6947 глаголов, все формы полной парадигмы которых присутствуют в Google Books Ngram. Рассматриваются частоты словоупотребления за период с 1920 по 2005 гг. Исследование начато с 1920 г., чтобы избежать влияния старинных форм слов, использовавшихся до реформы орфографии 1917 г. Аспектуальные пары брались по базе данных "Exploring Emptiness" [Janda et al. 2011]. Считаются коэффициенты корреляции временных рядов по Пирсону. Для контроля вычислялась также корреляция по Спирмену, давшая аналогичные результаты. Следует отметить, что в настоящее время не существует универсального и полностью адекватного способа определения степени похожести временных рядов [Koplenig 2015]. Мы используем корреляцию по Пирсону, следуя большинству работ.

3. Результаты и обсуждение

Получены следующие результаты. Для случайно выбранных пар глаголов и пар словоизменительных форм внутри леммы коэффициенты корреляции по Пирсону оказались равны 0,0404 и 0,1864 соответственно. Для случайно выбранных пар корреляции быть не должно,

результат ожидаем. Для пар словоизменительных форм корреляция оказалась мала (что несколько странно и требует дальнейших исследований), но статистически значима, ввиду огромного числа проанализированных данных.

Для префисальных и суффиксальных аспектуальных пар глаголов коэффициенты корреляции равны соответственно 0,3020 и 0,2435. Это соотношение также неожиданно. Традиционная точка зрения [Маслов 1984], разделяемая многими авторами, гласит, что суффиксальное образование вторичных имперфективов — это словоизменение, а префиксальное образование перфективов — словообразование. Если это так, то это должно приводить к большей семантической схожести и более высокому коэффициенту корреляции перфективов с вторичными имперфективами. Полученный результат можно трактовать как аргумент в пользу того, что оба способа образования аспектуальных пар с точки зрения этого противопоставления имеют примерно одинаковый статус, что согласуется с [Filip 1999], а также с статьей [Janda et al. 2011], в которой показано, что префискальные и суффиксальные пары ведут себя одинаково, нет статистически значимой разницы между их грамматическими профилями.

Вероятно, одним из влияющих факторов является то, что при образовании вторичного имперфектива часто привносится дополнительное значение многократности действия (как в рассматривавшемся Ю. Д. Апресяном [Апресян 1995] примере: *пить — выпить — выпивать*). Другим фактором является выборка глаголов: в единственной существующей аспектуальной базе для русского языка [Janda et al. 2011] содержатся только тройки базовый имперфектив — естественный перфектив — вторичный имперфектив. Эта база данных не включает пары, в которых вторичный имперфектив образуется из специализированного перфектива. Учет таких пар может привести к увеличению соответствующего коэффициента корреляции.

В целом не очень высокие значения коэффициентов корреляции легко объяснить многозначностью слов. В настоящее время у нас нет хорошего способа автоматически выделять различные значения слов и получать для них различные графики. Для нивелирования трудностей с многозначностью мы выделили 22 глагола, не имеющих совершенно различных значений (*делать, работать, казаться, играть, просить, верить, ставить, звать, звонить, хранить, рисовать, прятать, влечь, плевать, жечь, растить, мерзнуть, жрать, копать, нюхать, красть, щупать*). Рассмотрен более узкий временной интервал — с 1950, с устоявшимся современным русским языком и боль-

шим количеством изданных книг по сравнению с предшествующим периодом.

Для выбранных 22 глаголов среднее значение корреляции внутри словоизменения оказалось равно 0,568, а корреляция между базовым имперфективом и естественным перфективом — 0,758. Далее для этих слов были выбраны некоторые специализированные перфективы (*переделать* и т. д.), подсчитаны коэффициенты корреляции между ними и соответствующими вторичными имперфективами. Среднее значение оказалось равно 0,633. Таким образом, и после устранения явной многозначности и учета специализированных перфективов все равно корреляция между базовым имперфективом и естественным перфективом оказалась выше. Этот результат указывает на то, что префиксальный способ образования естественного перфектива не в меньшей степени сохраняет значение исходного слова, чем суффиксальное образование вторичных имперфективов.

4. Заключение

Google Books Ngram дает в руки исследователей совершенно новую методологию и позволяет по новому взглянуть на аспектуальную систему русского языка. Полученные результаты являются аргументом в пользу известной общей гипотезы Дж. Байби [Bybee et al. 1994] о том, что словоизменение и словообразование не полярные противоположности, а, скорее, точки на континуальной школе лексико-грамматических процессов.

Литература

1. *Апресян Ю. Д.* (1995), Трактовка избыточных аспектуальных парадигм в толковом словаре // Апресян Ю. Д. Избранные труды. Т. II: Интегральное описание языка и системная лексикография. М., с. 102–113.
2. *Маслов Ю. С.* (1984), Очерки по аспектологии. Л.: ЛГУ.
3. *Bybee J. L., Perkins R., and Pagliuca W.* (1994), The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World, University of Chicago Press.
4. *Filip H.* (1999), Aspect, Eventuality Types, and Nominal Reference, London.
5. *Janda L. A., Lyashevskaya O.* (2011), Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian, Cognitive Linguistics, vol. 22, no. 4, pp. 719–763.
6. *Janda L. A.* et al. (2013), Why Russian Aspectual prefixes aren't empty. Prefixes As Verb Classifiers, Bloomington: Slavika Publishers Indiana University.
7. *Janda L. A.* et al. (2011), Database "Exploring Emptiness". URL: http://emptyprefixes.uit.no.

8. *Koplenig A.* (2015), Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions, Digital Scholarship in the Humanities, Oxford University Press. URL: doi:10.1093/llc/fqv030.
9. *Michel J.* et al. (2011), Quantitative Analysis of Culture Using Millions of Digitized Books, Science, vol. 331, no. 6014, pp. 176–182.

References

1. *Apresjan Yu.D.* (1995), Traktovka izbytochnyh aspektualnyh paradigm v tolkovom slovare [Interpretation of redundant of aspectual paradigms in the explanatory dictionary]. In: Apresjan Yu. D. Izbrannye trudy. T. II. Integralnoe opisanie yazyka i sistemnaya leksikografiya [Selected papers. Vol. II: Integral description of the language and systemic lexicography]. Moscow. pp. 102–113.
2. *Maslov Yu.S.* (1984), Ocherki po aspektologii [Essays on Aspectology]. Leningrad, LGU.
3. *Bybee J. L., Perkins R., and Pagliuca W.* (1994), The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World, University of Chicago Press.
4. *Filip H.* (1999), Aspect, Eventuality Types, and Nominal Reference. London.
5. *Janda L. A., Lyashevskaya O.* (2011), Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian. In: Cognitive Linguistics, vol. 22, no. 4, pp. 719–763.
6. *Janda L. A.* et al. (2013), Why Russian Aspectual prefixes aren't empty. Prefixes As Verb Classifiers, Bloomington, Slavika Publishers Indiana University.
7. *Janda L. A.* et al. (2011), Database "Exploring Emptiness". Available at: http://empty-prefixes.uit.no.
8. *Koplenig A.* (2015), Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions, Digital Scholarship in the Humanities, Oxford University Press. Available at: doi:10.1093/llc/fqv030.
9. *Michel J.* et al. (2011), Quantitative Analysis of Culture Using Millions of Digitized Books. In: Science, vol. 331, no. 6014, pp. 176–182.

Соловьев Валерий Дмитриевич
Solovyev Valery
E-mail: maki.solovyev@mail.ru

Бочкарев Владимир Владимирович
Bochkarev Vladimir
E-mail: vbochkarev@mail.ru

Казанский федеральный университет (Россия)
Kazan Federal University (Russia)

Янда Лора
Университет Тромсё (Норвегия)
Janda Laura
University of Tromsø (Norway)
E-mail: laura.janda@uit.no

Д. Ш. Сулейманов, А. М. Галиева,
А. Р. Гатиатуллин

D. Suleymanov, A. Galieva, A. Gatiatullin

РАЗРАБОТКА СИСТЕМЫ АННОТАЦИИ ДЛЯ АНАЛИТИЧЕСКИХ КОНСТРУКЦИЙ ДЛЯ ТАТАРСКОГО НАЦИОНАЛЬНОГО КОРПУСАХ[1]

DEVELOPMENT OF THE SYSTEM FOR ANNOTATION OF MULTIWORD CONSTRUCTIONS FOR THE TATAR NATIONAL CORPUS

Аннотация. В статье представлен подход по разработке системы грамматической аннотации многословных конструкций на примере аналитических глагольных конструкций татарского языка. Авторами составлен рабочий инвентарь типов аналитических форм и конструкций, и построены правила для их извлечения, а также предложена система обозначений для конструкций разного типа. Предложенная система обозначений используется в программе расширенного морфологического анализа татарских словоформ. Программа расширенного морфологического анализа способна размечать синтетические и аналитические формы татарских глаголов.
Ключевые слова. Аналитические конструкции, корпусная грамматическая аннотация, татарский язык, корпус, глагол.

Abstract. This paper presents an approach to develop the system of annotation of Tatar multiword constructions on example of analytical verb constructions. The authors compiled preliminary inventory of multiword verb forms and constructions, distinguished basic types of them, and built rules for extracting of them. They developed also the system of tagging these constructions.
Keywords. Multi word constructions, **c**orpus grammatical annotation, the Tatar language, the verb, lexicography, linguistic corpus, verb.

1. Введение

Вопросы совершенствования корпусной разметки сохраняют свою актуальность, так как от количественных и качественных параметров разметки зависит применимость корпуса для различных целей и его удобство для пользователя.

Татарский национальный корпус «Туган тел» (ТНК — http://corpus.antat.ru) является лингвистическим ресурсом современного татарского языка с текстовой коллекцией более 100 миллионов словоупотреблений. В настоящее время ТНК имеет систему тэгов для обозначения морфологических категорий в пределах словоформы: при анализе определяется часть речи основы и дается описание последовательно-

[1] Исследование выполнено при финансовой поддержке РФФИ (проект № 15-07-09214).

сти формообразующих и словоизменительных аффиксов. Тем не менее система разметки, нацеленная только на анализ в пределах словоформы, имеет существенные ограничения.

В связи с тем, что в тюркских языках многие лексические единицы и грамматические категории выражаются многословными единицами, а модальность, как правило, задается не лексически, а посредством специальных конструкций, в текущей версии корпусной разметки аналитические формы и конструкции могут быть получены только путем сложных запросов, содержащих различные комбинации конъюнкции, дизъюнкции, отрицания с учетом аффиксов и лексических ограничений, Поэтому для извлечения многословных единиц требуется описание параметров не одной, а двух и более единиц с заданным расстоянием между единицами; такие запросы являются весьма громоздкими и трудоемкими и требуют от пользователя специальных навыков по их формализации.

Поэтому мы в настоящее время осуществляем доработку системы грамматической разметки введением новых обозначений для многословных аналитических форм (АФ) и конструкций (АК) татарского языка. На корпусных данных нами выделены основные типы многословных конструкций (МК), изучена их структура, разработаны правила для извлечения многословных конструкций, основанные на наборе необходимых грамматических и лексических признаков с учетом порядка расположения компонентов и возможности/невозможности вставки внешних элементов между компонентами МК.

2. Многословные конструкции в татарском языке

В лингвистической литературе отражены различные подходы к определению и выделению многословных конструкций. Среди работ по технологиям автоматической обработки многословных конструкций, можно выделить работу М. Копотева [Копотев 2004], где описывается классификация многословных конструкций, сделанная для аннотированного корпуса русских текстов ХАНКО (www.slav.helsinki.fi/hanco).

В тюркских языках многословные конструкции разных типов образуют большой пласт на стыке лексики и синтаксиса, включая в свой состав образования разной природы и структуры. Тем не менее в специальной литературе по автоматической обработки текстов на тюркских языках данная тема освещена достаточно слабо, хотя имеются

специальные работы по их выделению [Adalı et al. 2016, Mersinli et al. 2016]. Отсутствие справочных словарей, документирующих многословные конструкции, а также надежной теоретической основы и методологии извлечения таких конструкций констатируется даже для турецкого языка [Mersinli et al. 2016] для которого среди тюркских языков имеется больше всего описательных лингвистических исследований и систем автоматического анализа текста. Поэтому задача по реализации системы корусной аннотации многословных конструкций татарского языка из прикладной области плавно перетекает в теоретическую. В существующих исследованиях по грамматике татарского языка [Закиев 1992], как привило, дано лишь поверхностное описание структуры отдельных типов МК; а специальных словарей татарских МК не существует; электронный корпус позволяет получить исчерпывающий список таких конструкций и построить их модели.

Один из типов аналитических форм в тюркских языках образуют аналитические формы глаголов, которые образуются с помощью служебных неполнозначных глаголов. В татарском языкознании [Закиев 1992] также выделяется категория недостаточных глаголов, к которой относятся такие глаголы, как *иде, икән, исә, имеш*, служащие для образования аналитических временных форм и наклонений.

Способы образования аналитических форм с помощью функционально-вспомогательных глаголов наиболее полно рассмотрены в работах М. З. Закиева [Закиев 1992], Д. Г. Тумашевой [Тумашева 1986], А. А. Юлдашева [Юлдашев 1965] и Ф. А. Ганиева [1982], И. Ф. Исламовой [2012]. М. З. Закиев относит глаголы, образованные таким способом к сложным и составным глаголам. По определению М. Закиева сложные глаголы это глаголы, образованные с помощью словообразующих глаголов, а составные глаголы — глаголы, образованные с помощью модифицирующих и модальных глаголов.

3. Выделение и аннотирование аналитических глагольных форм для корпусной разметки

При автоматическом выделении многословных конструкций в электронном корпусе важными являются структурно-функциональные особенности таких конструкций.

С точки зрения автоматической разметки в модели необходимо отразить те свойства многословных конструкций, которые можно определить, используя словари и программы морфологического анализа.

Нами разработаны правила для извлечения конструкций, основанные на наборе грамматических тэгов и лексических единиц с учетом порядка расположения компонентов и возможности/невозможности вставки внешних элементов между компонентами конструкции. В частности, двухкомпонентные глагольные аналитические формы имеют следующий стандартный вид: первый компонент имеет произвольную (для большей части форм) глагольную основу с обязательным аффиксом (набором аффиксов), описываемым специальной формулой, и грамматически является относительно инвариантной единицей), второй компонент, как правило, может иметь все словоизменительные аффиксы, допустимые для глагольных единиц.

Нами выделяются и размечаются аналитические глагольные формы двух основных типов:

1) аналитические формы, выражающие аналитические времена глагола;
2) аналитические формы, с фазовыми или модальными глаголами.

Таблица 1 дает общее представление о структуре и характере разметки аналитических форм, которые образуются при помощи недостаточного глагола *иде* в функции вспомогательного.

Таблица 1. **Примеры разметки аналитических форм**

Морфологическая категория	**Структура**	**Пример**	**Перевод**	**Тэг АФ**
Будущее-прошедшее время	V+PCP_FUT(АчАк) +V_aux(иде)	кайтачак иде	должен был вернуться	FUTURE-PAST
Преждепрошедшее	V+PCP_PS(ГАн) +V_aux(иде)	кайткан иде	вернулся	PAST-PAST
Оптатив	V+COND(сА) +V_aux(иде)	*барса иде*	*он пошел бы*	OPT

Пример разметки:

кайткан иде ‘возвращался’
V(кайт) + PAST-PAST (PCP_PS (ГАн); V_aux(иде))

Таблица 2 дает общее представление о конструкциях аналитических глаголов со значением фазовости и модальности.

Таблица 2. **Примеры разметки АК со значением фазовости и модальности**

Тип	Структура	Пример	Перевод	Тэг АФ
Фазовый	V+PRES_3SG+ Phase_V(башла)	Җырлый башлау	начать петь	INCHOAT_1
Фазовый	V+ADVV +Phase_V(бет)	Агып бетү	вытечь	COMPLET_2
Модальный	OBL +Vf(кил)	Җырлыйсы килү	хотеть петь	DESID_2
Модальный	PRES_3SG+Vf(бел)	Ясый белү	уметь делать	CAPAC_1

Поскольку многие аналитические формы и конструкции являются многозначными (их значение может зависеть от множества факторов: лексического наполнения предложения и синтаксической позиции конструкции, например, употребления в главной или зависимой клаузе, и т. п.), тэг фиксирует лишь основное, прототипическое значение конструкции.

Для осуществления предложенной системы разметки аналитических форм создана программа расширенного морфологического анализа. Эта программа расширенного морфологического анализа способна размечать, как синтетические, так и аналитические формы татарских глаголов.

4. Заключение

К настоящему времени описаны наборы правил для извлечения аналитических глагольных времен и наклонений, а также конструкций с фазовыми и модальными глаголами и предложена система специальных тэгов для их обозначения. Система обозначений строится на тэгах, заимствованных из Лейципских правил глоссирования, а также из системы обозначений глагольной базы Verbum (httlp://www.mccme.ru/ling/verbum.htm), разработанной В. А. Плунгяном. Проведенная работа является важным шагом в реализации системы разметки многословных конструкций татарского языка.

Литература

1. *Ганиев Ф. А.* (1982), Образование сложных слов в татарском языке. М.: Наука.
2. *Закиев М. З.* (1992), Татарская грамматика. Т3. Синтаксис. Казань: Таткнигоиздат.

3. *Исламова И. Ф.* (2012), Модальные глаголы в татарском языке // Вестник Челябинского государственного университета. Челябинск, № 2 (256). С. 46–49.
4. *Копотев М.* (2004), «Несмотря на» «потому что», или Многокомпонентные единицы в аннотированном корпусе русских текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции 'Диалог'2004'. С. 335–339.
5. Татарский национальный корпус «Туган тел» «Tugan Tel» Tatar National Corpus, available at: http://corpus.antat.ru
6. *Тумашева Д. Г.* (1986), Татарский глагол (Опыт функционально-семантического исследования грамматических категорий). Казань: Изд-во Казанского университета, 1986.
7. *Юлдашев А. А.* (1965), Аналитические формы глагола в тюркских языках. Монография. М.: Наука.
8. *Adalı K., Dinc T., Gokırmak M., Eryigit G.* (2016), Comprehensive Annotation of Multiword Expressions Turkish // Proceedings of The First International Conference on Turkic Computational Linguistics TurCLing 2016, pp. 60–66.
9. *Mersinli Ü., Aksan Y.* (2016), A Methodological Consideration for Multi-word Unit Extraction in Turkish // Proceedings of The First International Conference on Turkic Computational Linguistics TurCLing 2016, pp. 27–31.
10. Вербум. URL: httlp://www.mccme.ru/ling/verbum.htm (01.04.2017).

References

1. *Ganiev F. A.* (1982), Obrazovanie slozhnyh slov v tatarskom yazyke. M., Nauka.
2. *Zakiev M. Z.* (1992), Tatarskaja grammatika. T. 3. Sintaksis. Kazan', Tatknigoizdat.
3. *Islamova I. F.* (2012), Modalnyje glagoly v tatarskom jazyke. In: Vestnik Chelabinskogo gosudarstvennogo universiteta. Chelabinsk, no. 2 (256), pp. 46–49.
4. *Kopotev M.* (2004), "Nesmotrja na" "potomu chto", ili Mnogokomponentnye edinicy v annotirovannom korpuse russkikh tekstov. In: Kompjuternaja lingvistika i intellektualnye tekhnologii. Trudy Mezhdunarodnoj konferencii 'Dialog' 2004', pp. 335–339.
5. Tatarskij nacionalnyj korpus "Tugan tel". Tatar National Corpus. Available at: http://corpus.antat.ru
6. *Tumasheva D. G.* (1986), Tatarskiy glagol (Opyt funkcionalno-semanticheskogo issledovanija grammaticheskikh kategoriy). Kazan', publ. Kazanskogo univerditeta, 1986.
7. *Juldashev A. A.* (1965). Analiticheskije formy glagola v turkskikh yazykach. Monografija. M., Nauka.
8. *Adalı K., Dinc T., Gokırmak M., Eryigit G.* (2016), Comprehensive Annotation of Multiword Expressions Turkish // Proceedings of The First International Conference on Turkic Computational Linguistics TurCLing 2016, pp. 60–66.
9. *Mersinli Ü., Aksan Y.* (2016), A Methodological Consideration for Multi-word Unit Extraction in Turkish // Proceedings of The First International Conference on Turkic Computational Linguistics TurCLing 2016, pp. 27–31.
10. Verbum. Available at: httlp://www.mccme.ru/ling/verbum.htm (01.04.2017).

Сулейманов Джавдет Шевкетович
Suleymanov Djavdet
E-mail: dvdt.slt@gmail.com

Галиева Альфия Макаримовна
Galieva Alfiya
E-mail: amgalieva@gmail.com

Гатиатуллин Айрат Рафизович
Gatiatullin Ayrat
E-mail: agat1972@mail.com

НИИ "Прикладная семиотика" Академии наук Республики Татарстан (Россия)
Research Institute of Applied semiotics (Russia)

Ю. Тао, В. П. Захаров
Yu. Tao, V. P. Zakharov

КОРПУСНО-ОРИЕНТИРОВАННЫЙ АНАЛИЗ ПЕРЕВОДА БЕЗЛИЧНЫХ ПРЕДЛОЖЕНИЙ С РУССКОГО ЯЗЫКА НА КИТАЙСКИЙ[1]

CORPUS-BASED STUDY OF THE TRANSLATION OF RUSSIAN IMPERSONAL SENTENCES INTO CHINESE

Аннотация. Представлены результаты исследования перевода безличных предложений с русского языка на китайский. Выявлено, что чаще всего они переводятся бессубъектными предложениями китайского языка. Показано, что таких предложений в переводном китайском языке намного больше, чем в оригинальном китайском. Мы полагаем, что в данном случае китайский переводной язык уподобляется русскому, отклоняясь от норм нормативного китайского.
Ключевые слова. Параллельный корпус текстов, безличные предложения, перевод.

Abstract. The paper presents the results of a research of the translation of impersonal sentences from Russian to Chinese. It has been revealed that impersonal sentences of Russian are often translated by subject-less sentences of Chinese. It is shown that their quantity in translated texts is more, than in original Chinese. We mean Chinese translated language assimilates Russian, deviating of standard Chinese.
Keywords. Parallel corpus, impersonal sentences, translation.

1. Введение

В настоящем исследовании ставится задача на основе параллельного русско-китайского корпуса изучить вопрос об особенностях перевода безличных предложений с модальными словами выражения возможности. Модальность возможности — один вид из модальностей. При всем терминологическом разнообразии определений этой категории наиболее общим является то, что модальность выражает отношение к действительности с точки зрения говорящего.

Безличные предложения — это односоставные предложения, в которых говорится о действии или состоянии, возникающем и существующем независимо от производителя действия или носителя состояния. Главный член предложения (или члены) может быть выражен формой 3-его лица единственного числа безличного или личного глагола; формой среднего рода; сочетанием слов категорий состояния (с модальным значением) с инфинитивом (составное глагольное

[1] 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».

сказуемое). Субъект может быть выражен родительным или дательным падежом. Безличные предложения являются характерным явлением русского языка. По мнению Чжан Хуйсэна, в китайском языке безличных предложений нет [张会森 Чжан Хуйсэн 2001: 326]. Некоторым аналогом им являются бесподлежащные (бессубъектные) предложения, но они существуют преимущественно в устной речи.

2. Постановка задачи

В данной статье мы анализируем особенности перевода безличных предложений с модальными словами возможности. На основе поиска в параллельном корпусе были выбраны все безличные предложения со словами *можно, возможно/невозможно* и *нельзя* и их перевод, проанализированы все переводные конструкции и определена доля в них бессубъектных предложений китайского языка. Далее проводилось сравнение количества бессубъектных предложений в параллельном переводном корпусе и в сопоставимом корпусе. Данный анализ показывает отклонения переводного языка от правил целевого.

3. Корпус

Материалом для данной работы послужил параллельный корпус русского и китайского языков, составленный из русских научных текстов и их переводов, и сопоставимый корпус оригинальных китайских научных текстов[2] [Tao, Zakharov 2015]. Для данного исследования были выбраны 3 подкорпуса (см. табл. 1).

Таблица 1. **Объемные характеристики корпусов**

	Параллельный корпус (объем в токенах)		**Сопоставимый корпус (объем в токенах)**
	Русский язык	**Китайский язык**	**Китайский язык**
Политика и м/н отношения	418 100	710 856	657 718
Лингвистика	568 738	855 326	795 546
Литература	208 643	315 258	316 328
Итого	1 193 481	1 881 440	1 769 592

[2] Поддержан грантом Бюро национального фонда социальных и гуманитарных наук Китайской Народной Республики.

Как видно из табл. 1, подкорпусы русского и китайского языков несколько различаются по общему количеству токенов. Количество токенов китайского языка в параллельном корпусе и сопоставимом корпусе практически одинаково.

4. Экспериментальное исследование

Посмотрим, как переводятся безличные предложения с модальными словами *можно*, *возможно/невозможно* и *нельзя* с русского языка на китайский (табл. 2).

Таблица 2. **Перевод безличных предложений с модальными словами *можно, возможно/невозможно и нельзя***

Исходный текст	Перевод субъектными предложениями	Перевод бессубъектными предложениями	Всего	Доля бессубъектных предложений
Предложения со словом *можно*	189	413	602	68.6%
Предложения со словом *возможно/ невозможно*	45	171	216	79.2%
Предложения со словом *нельзя*	78	160	238	67.2%

Данные табл. 2 свидетельствуют, что более двух третей безличных предложений с модальными словами *можно*, *возможно/невозможно* и *нельзя* были переведены на китайский бессубъектными предложениями. Хотя субъектные и бессубъектные предложения китайского языка отличаются от личных и безличных предложений русского языка в когнитивном механизме, но их структуры схожи. В то же время анализ переводов показывает, что многие безличные предложения следовало бы переводить субъектными предложениями китайского языка, что более бы соответствовало правилам китайского языка. См. пример 1.

(1) *Против такого произвольного заявления нельзя, конечно, ничего возразить, но* ***можно поставить все-таки вопрос, возможно ли осуществление подобного желания****.*

当然不能对这样随意的声明反对什么，但还是可以提出一个问题，能否实现这样的愿望。(бессубъектное)

当然，我们没有能力对这样随意的声明提出反对，但是还是可以提出自己的疑问，即这样的愿望到底能否实现 (субъектное)

Бессубъектные предложения могут появляться в китайских текстах, но они носят оттенок устной речи. Чжу Дэси убедительно доказал, что в современным литературном китайском языке бессубъектные предложения — это явления устной речи, а норма для письменного языка – субъектные предложения [朱德熙 Чжу Дэси 1987].

Далее мы разделили слова китайского языка, которые могут составлять бессубъектные предложения, на 5 групп, так что в каждой группе слова схожи по значению, например, “不能” и “不应当”, и сравнили количество бессубъектных предложений со словами этих групп в переводном корпусе и в сопоставимом (табл. 3).

Таблица 3. **Бессубъектные предложения с модальными словами возможности в переводном и сопоставимом корпусе**

Модальные слова возможности китайского языка	Количество бессубъектных предложений	
	в параллельном корпусе	в сопоставимом корпусе
能，**能够** (можно)	237	46
可能 (возможно)	209	40
可以 (мочь)	156	29
不能，**不**应当 (нельзя, не следует)	144	20
不可能 (невозможно)	105	15

Поскольку объем корпусов практически одинаков (см. табл. 1), то можно говорить о явной диспропорции. Предложения в параллельном корпусе являются переводом с русского, а предложения в сопоставимом корпусе изначально писались по-китайски, в соответствии с нормами китайского языка. Данные табл. 3 свидетельствуют о том, что количество бессубъектных предложений во всех пяти группах в переводных текстах намного больше, чем в оригинальных текстах. Данные табл. 2 показывают, что в большинстве случаев безличные предложения русского языка переводятся именно бессубъектными предложениями.

В предыдущей работе мы выдвинули гипотезу о «проникновении языка оригинала» при переводе [Тао, Захаров 2016]. В Китае проник-

новение языка оригинала обозначает влияние европейских языков на китайский и заимствование элементов из европейских грамматик и лексиконов.

Мы полагаем, что и в данном случае обилие бессубъектных предложений в переводном китайском языке объясняется «проникновением языка оригинала», т. е. влиянием исходного русского языка на переводной китайский. Исследование показало, что безличные предложения русского языка влияют на китайский переводной язык, и поэтому число бессубъектных предложений в переводном китайском больше, чем в оригинальном.

5. Заключение

Данное исследование было посвящено переводу безличных предложений с модальными словами *можно*, *возможно/невозможно* и *нельзя* с русского языка на китайский. Большое количество таких предложений было переведено бессубъектными предложениями. Их количество в переводном языке гораздо больше, чем в целевом. И это происходит под влиянием русского языка. Важно отметить, что исследование проводилось на репрезентативном параллельном и сопоставимом корпусах и обладает всеми признаками достоверности.

Данная статья имеет и практическое значение: 1) так как некоторых грамматических значений русского языка в целевом (китайском) языке не существует, то переводчики должны глубоко понимать эти значения исходного языка, чтобы правильно переводить оригинальный текст; 2) чтобы уйти от «бессубъектности», с учетом норм целевого языка, при переводе целесообразно добавлять слова “我们” (мы), “本文” (данная работа), “本章” (данная глава) и др., чтобы переводной текст был более удобочитаемым и приемлемым для китайского читателя.

Литература

1. *Тао Ю., Захаров В. П.* (2016), «Иностранизация» сочетаемости в конструкциях с предлогом 对 (duì) при переводе научных текстов с русского на китайский // Вестник МГУ, Сер. 22, № 3, с. 58–72.
2. *Чжан Хуйсэн* (2001), 张会森. 俄汉语对比研究（上） [M]. 上海：上海外语教育出版社.
3. *Чжу Дэси* (1987), 朱德熙句子和主语－印欧语影响现 代书面汉语和汉语句法分析的一个实例[J]. 世界汉语教学, 1987(3), 31–34.

4. *Tao Y., Zakharov V.* (2015), The Development and Use of Russian-Chinese Parallel Corpus // Automatic Documentation and Mathematical Linguistics. Vol. 49, no. 2, pp. 65-75.

References

1. *Tao Y., Zakharov V.* (2015), The Development and Use of Russian-Chinese Parallel Corpus. In: Automatic Documentation And Mathematical Linguistics, 49(2), pp. 65–75.
2. *Tao Y., Zakharov V.* (2016), Foreignization of Dui Collocation Constructions in the Frame of Translation of Academic Texts from Russian to Chinese. In: The Moscow University Herald, Series 22, Translation Theory, no. 3, pp. 58–72.
3. *Zhang Huisen* (2001), A Contrastive Study of Russian and Chinese, vol. 1. Shanghai, Shanghai Foreign Language Education Press.
4. *Zhu Dexi* (1987), Sentence and Subject: A Case Study of the Indo-European Language Impact on Modern Written Chinese and Syntactical Analysis of Chinese. In: Chinese Teaching in the World, 1(3), pp. 31–34.

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

Захаров Виктор Павлович
Санкт-Петербургский государственный университет (Россия)
Zakharov Victor
St Petersburg State University (Russia)
E-mail: v.zakharov@spbu.ru

М. К. Тимофеева
M. K. Timofeeva

КОРПУСНАЯ ПРАГМАТИКА: О ВОЗМОЖНОСТИ СОЗДАНИЯ УЧЕБНОГО КОРПУСА

CORPUS PRAGMATICS: ABOUT POSSIBILITY OF CREATING EDUCATIONAL CORPUS

Аннотация. Прагматика как раздел лингвистики возникла относительно недавно (позже фонетики, синтаксиса, семантики), традиция включения её в число языковых уровней, в целом, не сложилась. Это проявляется в том, что 1) корпусная прагматика – один из поздних разделов корпусной лингвистики; 2) при обучении языку (родному или иностранному) прагматике целенаправленно не обучают; 3) навыки распознавания прагматических составляющих часто недостаточно развиты даже по отношению к родному языку. Ситуация усугубляется трудностью выявления прагматических составляющих по текстовым показателям. Исследуется возможность создания учебного прагматического корпуса, предназначенного для развития навыков распознаванию неоднозначностей, обусловленных неопределённостью выбора пресуппозиций; рассматривается ряд связанных с этой задачей общих вопросов корпусной прагматики.
Ключевые слова. Корпусная прагматика, учебный корпус, пресуппозиция, неоднозначность.

Abstract. Pragmatics is relatively young (comparatively with phonetics, syntax, or semantics); tradition of considering this language level, in whole, has not formed. Consequently: 1) corpus pragmatics emerged within corpus linguistics only recently; 2) standard linguistic courses (for native or foreign language) lack special chapters for pragmatics; 3) people often are insufficiently skilled in recognizing pragmatic phenomena even in the native language. Peculiarity of pragmatics – deficit of formal markers – complicates the situation. We study possibility of creating educational pragmatic corpus for developing the skill of recognizing ambiguities caused by uncertain presuppositions; and discuss several conjoined general questions of corpus pragmatics.
Keywords. Corpus pragmatics, educational corpus, presupposition, ambiguity.

1. Введение

Прагматика моложе других, более традиционных, разделов лингвистики; её статус и внутренняя структура до сих пор имеют немало конкурирующих версий. Корпусная прагматика, как одна из частей корпусной лингвистики, также появилась относительно недавно. Эти факты в значительной степени объясняются тем, что прагматика имеет дело преимущественно с имплицитными составляющими: прагматика — это «искусство анализа несказанного» (“the art of the analysis of the unsaid”) [Corpus pragmatics: a handbook 2014: 2].

При обучении (родному или иностранному) языку прагматике, как одному из аспектов языка, целенаправленно не обучают. Соответствующий раздел отсутствует в учебниках, традиционно включаю-

щих сведения о фонетике, лексике, синтаксисе, семантике. Прагматика может присутствовать лишь неявно, например, когда для анализа смысла конкретного текста необходимы общие сведения о мире или обсуждаемой ситуации, или важны особенности выражения речевых актов.

Недостаточное развитие навыков распознавания прагматических составляющих характерно даже для носителей языка. Отсюда следует, что прагматике необходимо учить как отдельному языковому уровню, разрабатывая и систематизируя соответствующий учебный материал, в частности, используя корпусные технологии. Далее рассматривается возможность создания учебного корпуса, предназначенного для освоения тех составляющих прагматики, которые пока не попали в фокус внимания современной корпусной прагматики.

2. Основные проблемы

При создании прагматического корпуса прежде всего возникает проблема разметки. Прагматические явления и представляющие их теории многообразны; нет единства в классификации. Для многих прагматических составляющих текстовые маркеры отсутствуют. Следовательно, выбор отображаемых в корпусе явлений представляет собой нетривиальную задачу.

В рамках англо-американской традиции в современную прагматику обычно включают четыре класса явлений — импликатуры, пресуппозиции, речевые акты, дейксис. В имеющихся прагматических корпусах фиксируются, как правило, только речевые акты. Для русского языка они отражены, например, в мультимедийном корпусе НКРЯ; на их анализ ориентирована также обсуждаемая в [Шерстинова 2015] прагматическая разметка речевого корпуса «Один речевой день». Изучение другого круга прагматических явлений возможно на основе корпуса рассказов о сновидениях [Рассказы о сновидениях: Корпусное исследование устного русского дискурса 2009], в разметке которого фиксируется коммуникативная организация.

Конечно, возможно исследовать прагматические явления на основе корпусов без специальной прагматической разметки, когда нужные элементы выявляются путём подбора подходящих косвенных контекстных признаков. Однако это не позволяет получить приемлемый результат для многих прагматических явлений из-за невысокой частоты их встречаемости и сложности подбора эффективных признаков.

3. Функции и содержание прагматического корпуса

Назначение. Основное назначение обсуждаемого прагматического корпуса: накопление структурированного текстового материала, создаваемого людьми, изучающими определённые разделы прагматики. Получаемый в результате учебный корпус, после внесения необходимых исправлений, предполагается использовать для изучения вошедших в него прагматических явлений.

Языковой материал откорректированного учебного прагматического корпуса может использоваться и в других целях, например, при разработке алгоритмов автоматического анализа текста, учитывающих прагматические составляющие. Это применение, однако, в данном обсуждении не рассматривается.

Функции корпуса: 1) пополнение корпуса новыми примерами использования прагматических составляющих в текстах русского языка; 2) анализ добавленных текстов; 3) внесение изменений в корпус; 4) поиск прагматических явлений, относящихся к заданному классу.

Единица корпуса. Корпус имеет прежде всего учебное назначение. Поэтому оптимальным представляется «точечный» подбор коротких фрагментов текстов (от одного до нескольких предложений), содержащих одно или более вхождений прагматических явлений, относящихся к рассматриваемым типам. Таким образом, корпус создаёт среду с «повышенной концентрацией» изучаемых прагматических явлений. Тексты могут быть взяты из печатных изданий или из реальной коммуникации. Возможный вариант: сочинение таких текстов самими пользователями.

Рассматриваемые прагматические составляющие. Существует ряд естественных ограничений. Некоторые прагматические составляющие довольно тривиальны, их обнаружение не представляет особого труда и не нуждается в тренировке. Таковы, в частности, импликатуры, обусловленные соблюдением некоторых постулатов Г. П. Грайса; например, соблюдение первого постулата качества фактически означает добавление к сказанному пропозиции ‘То, что сказано, автор полагает истинным’. Имеются прагматические явления, соотносимые скорее с семантикой, чем с прагматикой. Например, пресуппозиции, триггерами которых являются импликативные глаголы или глаголы изменения состояния, либо одна из частиц типа *даже*, *только*, *уже*, обычно включают в число семантических пресуппозиций. Ряд прагматических явлений (например, импликатуры) затруднительно рассматривать в общем случае из-за большого числа конкурирующих

теорий; сначала необходимо выработать единую теоретическую базу классификации. Некоторые прагматические явления уже отражены в существующих корпусах.

С учётом этих ограничений при создании прагматического корпуса было решено сфокусироваться на определённых классах прагматических пресуппозиций.

Прежде всего речь идёт о неоднозначных текстах, воспринимаемых автором / адресатом как однозначные на основе явного или неявного использования прагматических пресуппозиций. Именно неоднозначность предлагается считать триггером пресуппозиции в данном типе текстов. Рассматриваются следующие типы явлений, обусловливающих неоднозначность текстов.

1. Неопределённость области действия кванторного слова / отрицания / обозначения логической операции; неопределённость границ / ролей аргументов предиката; неопределённость антецедента анафоры. Простые иллюстративные примеры: *Все тетради и книги в портфеле*: ‘(тетради и книги) в портфеле’ или ‘тетради и (книги в портфеле)’; *Это было не зелёное яблоко*: ‘(не зелёное) яблоко’ или ‘не (зелёное яблоко)’; *Верчение диска вызывало лёгкое дуновение*. Реальные примеры из СМИ: *В Бердске штрафуют владельцев собак, гуляющих без намордников и поводков*; *Люстру лучше брать незамысловатую, в которой нет места для скопления пыли и острых оконечностей*. Однозначное понимание этих двух предложений происходит благодаря наличию пресуппозиций ‘намордники и поводки носят собаки, а не их хозяева’ и ‘в люстрах может скапливаться пыль, а не острые оконечности’.

2. Неопределённость экзистенциального типа: существование / несуществование объекта / процесса / события, о котором говорится в тексте. Сюда относятся, в частности, предложения, которые Е. В. Падучева [Падучева 2014]) называет конструкциями с радикальным отрицанием, например, *Он не обрадовал нас своим приходом*. Выбор значения зависит от того, что является пресуппозицией: пропозиция ‘Он приходил’ (но это «нас» не обрадовало) или пропозиция ‘Он не приходил’ (тем самым, лишив «нас» случая порадоваться его приходу).

3. Неопределённость приписывания тексту прагматических пресуппозиций, по-разному разрешаемая разными людьми. Например: — *Послушай, Клэр, не надо всё-таки бросаться луковицами в мистера Брэддока. — Так я сходила наверх и подобрала её, — утешила Клэр свою хозяйку, — Она уже в супе.* (П. Г. Вудхаус). Первой реплике собе-

седницы приписывают разные пресуппозиции: ‘Бросаться луковицами в людей нехорошо’ и ‘Разбрасываться луковицами неэкономно’.

Включаемые в корпус тексты могут содержать ошибки или аномалии, в этом случае они зачастую ярче демонстрируют нужное прагматическое явление, чем более правильные тексты. Например, высказывание участника телевизионной передачи по экономике *Одной ногой находимся и понимаем, что происходит в развитых экономиках* имеет ясный смысл благодаря наличию пресуппозиции ‘человек понимает экономическую ситуацию умом, а не ногой’.

Во всех случаях прагматические составляющие рассматриваются как пропозициональные по своей природе, то есть как дополняющие явно высказанные в тексте пропозиции определённым набором неявных пропозиций (пресуппозиций).

Разметка отражает альтернативные роли, границы, связи, ставя им в соответствие неявные пропозиции. Таким образом, разметка, помимо традиционной расстановки кодов языковых явлений, включает также дополнительные пропозиции (пресуппозиции), для представления которых должна быть выработана определённая стандартная форма.

4. Общие проблемы прагматических корпусов

Обсуждаемый учебный прагматический корпус вынуждает задуматься о границах и некоторых понятиях корпусной лингвистики. Какие текстовые коллекции разумно соотносить с понятием «языковой корпус»? Этот вопрос возникает прежде всего в связи с добавлением неявных пропозиций в разметку. Допустимость такой операции, по-видимому, требует ограничений. Должен ли корпус быть коллекцией размеченных текстов (базой данных) с определённой системой поиска или возможен более широкий набор функций (при той же разметке)? Например, добавление функций обучения за счёт полного или частичного скрытия от пользователя информации, содержащейся в разметке. Остаётся ли корпус учебным после его корректировки и использования в целях обучения языку или для проведения лингвистических исследований?

Литература

1. *Падучева Е. В.* (2014), Может ли отрицание отрицать презумпцию? // Научно-техническая информация. Сер. 2, Информационные процессы и системы. № 3, С. 24–32.

2. *Рассказы о сновидениях*: *Корпусное исследование устного русского дискурса* (2009), Под ред. А. А. Кибрика и В. И. Подлесской. М.: Языки славянских культур. 736 с.
3. *Шерстинова Т. Ю.* (2015), Прагматическое аннотирование коммуникативных единиц в корпусе ОРД: микроэпизоды и речевые акты // Труды Международной конференции «Корпусная лингвистика–2015». СПб., с. 451–459.
4. *Corpus pragmatics: a handbook* (2014), Karin Aijmer, Christoph Rühlemann (Eds.). Cambridge University Press. 480 p.

References

1. *Paducheva E. V*. (2014), Mozhet li otricanie otritsat' prezumpciju? [Can negation negate the presumption?]. In: Nauchno-tehnicheskaja informacija, ser. 2, no. 3, pp. 24–32.
2. Rasskazy o snovidenijah: Korpusnoe issledovanie ustnogo russkogo diskoursa (2009) [Night dream stories: A corpus study of spoken Russian discourse]. Ed. by A. A. Kibrik, B. I. Podlesskaya. Moscow, Jazyiki slavjanskih kul'tur [Languages of Slavonic Culture]. 736 p.
3. *Sherstinova T. Y.* (2015), Pragmaticheskoe annotirovanie kommunikativnyh edinic v korpuse ORD mikroehpizody i rechevye akty [Pragmatic annotation of communicative units in the ORD corpus: micro episodes and speech acts]. In: Trudy mezhdunarodnoy nauchnoy konferentsii "Korpusnaya lingvistica-2015" [Proc. of the International Conference "Corpus linguistics-2015"]. St Petersburg, pp. 451–459.
4. *Corpus pragmatics: a handbook* (2014), Karin Aijmer, Christoph Rühlemann (Eds.). Cambridge University Press. 480 p.

Тимофеева Мария Кирилловна
Институт математики СО РАН, Новосибирский государственный университет (Россия)
Timofeeva Mariya
Institute of mathematics SB RAS, Novosibirsk State University (Russia)
E-mail: mtimof@inbox.ru

М.В.Хохлова
M.Khokhlova

ОСОБЕННОСТИ СТАТИСТИЧЕСКИХ МЕР ПРИ ВЫДЕЛЕНИИ БИГРАММ[1]

DISTINCTIVE FEATURE SOFASSOCIATION MEASURES FOR BIGRAM EXTRACTION

Аннотация. Статья посвящена сравнению семи статистических мер на материале русскоязычного корпуса текстов ruTenTen. Были рассмотрены биграммы для высокочастотных существительных, которые были выделены при помощи данных мер, были проанализированы их отличия.
Ключевые слова. Статистические меры, сочетаемость, корпусы текстов, биграммы, оценка.

Abstract. The research focuses on a comparison between seven statistical measures based on the ruTenTen corpus. The paper gives a survey of bigrams with a number of Russian high-frequency nouns that were extracted by the given measures and analyzes the differences.
Keywords. Statistical measures, collocability, text corpora, bigrams, evaluation.

1. Введение

В последнее время в связи с возросшей потребностью в автоматизированных системах большое внимание уделяется вопросу, связанному автоматическим выделением словосочетаний сочетаний в текстах. Существуют различные статистические метрики для оценки сочетаемости слов. Ряд мер получил название мер ассоциации, или ассоциативных мер. Они позволяют вычислять силу связи между элементами словосочетаний и основываются на частотах данных словосочетаний и входящих в них отдельных слов. Таким образом, может быть вычислена некоторая устойчивость, присущая лексическим единицам, позволяющая их расположить на шкале: от свободных сочетаний до фразеологизированных структур. Всего существует более 80 мер, позволяющих оценить силу связанности словосочетаний [Pecina 2009].

В свою очередь в статье мы обратимся к нескольким статистическим мерам и сравним данные, которые они выделяют.

2. Статистический аппарат

В рамках исследования были отобраны 7 мер: MI, log-likelihood (LL), t-score, MI^3, minimum sensitivity (MS), logDice и MI.log-f. Наибо-

[1] Статья подготовлена в рамках работы по гранту Президента РФ для государственной поддержки молодых российских ученых № МК-5274.2016.6 «Исследование статистических закономерностей сочетаемости лексических единиц».

лее часто в литературе, посвященной вычислению силы связанности, упоминаются первые три меры. Их подробный обзор дан в ряде работ (см., например, [Evert 2004]).

Особенностью меры *MI* (или *коэффициента взаимной информации*) является то, что она позволяет найти в корпусе редкие словосочетания. Таким образом, вес каждой отдельной коллокации тем больше, чем реже она встречается. Поэтому в случаях, когда частота сочетания мала, использование данной формулы может привести к неправильным результатам. Чтобы решить эту проблему в ряде работ были предложены модификации данной меры (примерами являются следующие меры).

В работе [Oakes 1998: 171–172] эмпирически выводится формула, в которой величина $f(n,c)$ возводится в куб (MI^3):

$$MI^3 = \log_2 \frac{f^3(n,c) \times N}{f(n) \times f(c)},$$

где: n — ключевое слово; c — коллокат; $f(n,c)$ — частота встречаемости ключевого слова n в паре с коллокатом c; $f(n)$, $f(c)$ — абсолютные (независимые) частоты ключевого слова n и слова c в корпусе; N — общее число словоформ в корпусе.

Следующая мера MI.log-f также является вариантом взаимной информации и была введена в работе [Kilgariff, Rychly, Smrz, Tugwell 2004]:

$$MI.\log\text{–}f = MI \times \ln(f(n,c)+1).$$

В работе [Rychly 2008] была предложена мера log *Dice*.Она является модифицированным вариантом меры Дайса [Smadja 1993; Diasetal. 1999] и призвана устранить ее недостаток, связанный с выдачей повышенных значений для редко встречающихся биграмм:

$$\log Dice = 14 + \log_2 \frac{2f(n,c)}{f(n)+f(c)}.$$

Мера MS (minimum sensitivity) была предложена в работе [Pedersen, Bruce 1996]. Ее значения колеблются в диапазоне от 0 до 1.

$$MS = \min\left\{\frac{f(n,c)}{f(n,c)+f(n,\overline{c})}, \frac{f(n,c)}{f(n,c)+f(\overline{n},c)}\right\}.$$

Как отмечается в работе [Evert 2004], данная мера показала наилучшие результаты по сравнению с другими, и была реализована в ряде систем.

3. Методика исследования

Наше исследование было проведено на материале корпуса русского языка ruTenTen [Jakubíček et al. 2013], содержащего более 18,3 млрд. словоупотреблений.

Нами были рассмотрены биграммы для десяти высокочастотных существительных, отобранных по словарю [Ляшевская, Шаров 2009]: *год, человек, время, дело, жизнь, день, рука, работа, слово* и *место*. Были проанализированы 100 левосторонних нелемматизированных биграмм для каждого из приведенных существительных, которые были выделены при помощи 7 вышеописанных статистических мер и отсортированы по убыванию значения каждой из мер. Для каждой пары мер были найдены пересечения списков сочетаний. Необходимо отметить, что списки не были предварительно отфильтрованы, так как планировалось проверить, какие биграммы выделяются в «сыром» виде. В ходе экспериментов проверялось, насколько совпадают между собой биграммы, выделяемыми разными мерами, т. е. результативность статистических метрик.

4. Результаты

В таблице 1 показаны результаты попарного сравнения количества выделенных биграмм описанными мерами.

Таблица 1. **Количество совпадающих биграмм (среднее)**

	MI	t-score	MI^3	LL	MS	logDice	MI.log-f
MI		0,6	0,9	0,7	0,5	0,6	2,1
t-score	0,6		78,7	77,9	91,4	97,0	41,6
MI^3	0,9	78,7		83,2	73,1	77,5	57,5
LL	0,7	77,9	83,2		73,3	77,8	49,2
MS	0,5	91,4	73,1	73,3		93,3	37,0
logDice	0,6	97,0	77,5	77,8	93,3		40,6
MI.log-f	2,1	41,6	57,5	49,2	37,0	40,6	

4.1. Мера MI

Мера MI показала минимальное пересечение с остальными мерами по выдаваемым результатам, что подтверждает результаты, описанные в других работах: мера присваивает большие значения низкочастотным словосочетаниям, а также сочетаниям с опечатками и иноязычными вкраплениями. Отсутствие данных словосочетаний среди выделенных другими мерами может быть объяснено также объемом самого корпуса и, соответственно, большим количеством в нем hapaxlegomena, которые, в силу особенностей меры MI, оказываются в верхней части списков выделенных биграмм. Таким образом, можно утверждать, что выдаваемые ею биграммы являются уникальными, то есть практически отсутствуют в других списках.

4.2. Мера t-score

В верхней части списков оказалось большое количество сочетаний с предлогами и союзами (*по словам, из жизни, изо дня*). При этом данная мера показывает максимальное совпадение выделенных словосочетаний с мерой logDice.

4.3. Мера MI³

Наряду со служебной лексикой были также получены сочетания со знаками пунктуации, при этом они отсутствовали в списках, полученных при помощи других мер (за исключением меры *LL*). В отличие от меры взаимной информации ее модифицированный вариант продемонстрировал отсутствие выделенных биграмм с hapaxlegomena.

4.4. Мера LL

Мера *LL* продемонстрировала промежуточные результаты, с одной стороны, совпадающие с другими мерами в части выявления сочетаний с предлогами и союзами, с другой стороны, выделяющая сочетания, зафиксированные в словарях (*иностранных дел, повседневной жизни, покладая рук*).

4.5. Мера MS

Наибольшие значения меры имеют сочетания с цифровыми комплексами, числительными и предлогами. Также отметим, что ряд биграмм получил одинаковые значения меры (это может быть объяснено самой формулой), что затрудняет ранжирование результатов и их последующую обработку.

4.6. Мера logDice

Мера logDicepродемонстрировала максимальное совпадение в результатах с мерами *t*-score и *MS*. Как и в случае с мерой *t*-score, зафиксированы сочетания с предлогами, союзами и местоимениями.

4.7. Мера MI.log-f

Для всех 10 существительных данная мера выделила максимальное число биграмм с прилагательными и существительными (*ремонтные работы, календарных дней, образ жизни*). Для существительного «рука» в верхней части списка были отмечены сочетания с глаголами (*махнул рукой, развел руками, всплеснула руками*). Несмотря на то, что менее 50 % выделенных биграмм были зафиксированными другими мерами, однако в целом при анализе списков словосочетаний было установлено, что мера MI.log-f показала наилучшие результаты.

Заключение

Результаты демонстрируют относительную взаимозаменяемость статистических мер, однако требуется дальнейшая оценка выделенных сочетаний, в том числе экспертами. Полученные данные могут найти дальнейшее применение в работе по изучению возможностей количественных метрик. Статистические оценки для силы связанности лексических единиц также могут быть использованы при снятии лексической неоднозначности, при поиске переводных эквивалентов в параллельных текстах и корпусах, при идентификации синонимов и антонимов.

References

1. *Dias G., Guillore S., Lopes J.* (1999), Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In: Proceedings of of 6ème Conférence Annuellesur le Traitement Automatique des Langues Naturelles (TALN), 12–17 July, Cargèse, France, pp. 333–339.
2. *Evert S.* (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Available at: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf
3. *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
4. *Kilgariff A., Rychly P., Smrz P., Tugwell D.* (2004), The Sketch Engine. In: Proceedings of Euralex Lorient, France, July, pp. 105–116.

5. *Lyashevskaya O., Sharoff S.* (2009), Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Natsional'nogo Korpusa Russkogo Jazyka) [Frequency Dictionary of Contemporary Russian based on the Russian National Corpus data]. Moscow, Azbukovnik.
6. *Oakes M.* (1998), Statistics for corpus linguistics. Edinburgh, Edinburgh University Press.
7. *Pecina P.* (2009), Lexical Association Measures. Collocation Extraction. Prague, Institute of Formal and Applied Linguistics.
8. *Pedersen T., Bruce R.* (1996), What to infer from a description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.
9. *Rychly P.* (2008), A lexicographer-friendly association score. In: Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008, pp. 6–9.
10. *Smadja F.* (1993), Retrieving collocations from text: X tract. In: Computational Linguistics, 19 (1), pp. 143–177.

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

А. Я. Шайкевич
A. Shaikevich

СРЕДНИЙ ИНТЕРВАЛ В ДИСТРИБУТИВНО-СТАТИСТИЧЕСКОМ АНАЛИЗЕ ТЕКСТОВ

MID-INTERVAL IN DISTRIBUTIONAL-STATISTICAL ANALYSIS OF TEXTS

Аннотация. Корпус русской прозы (14 миллионов словоупотреблений) членится на фрагменты по 40 слов. Если совместная встречаемость двух слов во фрагментах существенно превышает величину, подсчитанную на основе нулевой гипотезы, делается вывод о наличии связи между этими словами. В ходе анализа возникает огромная сеть текстуальных связей слов.
Ключевые слова. Дистрибутивно-статистический анализ, текстуальные связи слов.

Abstract. A corpus of Russian prose (14 million running words) is divided into fragments of equal size (40 words each). If actual co-occurrence of two words exceeds significantly the expectation, calculated on the basis of the null hypothesis, the two words are said to be linked together. A huge network of textual links of words is constituted as a result of the analysis.
Keywords. Distributional statistical analysis, textual links of words.

Дистрибутивно-статистическим анализом (ДСА) называем формальный (без обращения к смыслу) метод изучения больших собраний текстов, опирающийся исключительно на статистику распределения графических слов. Вводимое понятие **интервала** текста позволяет надеяться на получение самых разнообразных содержательных результатов. В монографии [Шайкевич и др. 2013, 2016] были исследованы интервалы, задаваемые исходным членением текста. Объем изучаемого корпуса — около 14 млн словоупотреблений. Анализ комбинаторики букв внутри графических слов (**микроинтервал**) и бинарных словосочетаний (**минимальный интервал**) приводил к открытию грамматики и лексических единиц, больших чем слово.

Остальные интервалы задаются исследователем. В рамках заданного интервала корпус текстов автоматически членится на фрагменты равной длины, каждому фрагменту присваивается свой адрес. Для пары слов определяется число общих адресов, если оно существенно превосходит математическое ожидание, подсчитанное в предположении независимости слов, делается вывод о том, что между этими словами обнаружена **текстуальная связь**. Мерой неслучайности связи служит выражение

$$S = (x - m - 1)/\sqrt{m}, \qquad (1)$$

где x — наблюдаемое число общих фрагментов, а m — математическое ожидание. Все предварительные исследования показывают, что, начиная с $S = 3$, результаты оказываются осмысленными.

Перед окончательным выбором длины фрагмента для **среднего интервала** было проведено пробное исследование на материале текстов Тургенева (720 тыс. словоупотреблений). Сравнивались результаты при длине фрагмента в 40 слов, в 200 слов и в 1000 слов. В первом случае слово *дуэль*, например, обнаружило связь с *драться* ($S = 15$), при длине в 200 слов с *дуэлью* связаны ($S > 4$), *вызов, драться, пистолет, поединок, противник, пуля, секундант*. Выбранная мера неслучайности напрямую зависит от объема изучаемого корпуса. Все семь слов, ассоциированных с *дуэлью* у Тургенева, будут обнаружены на всем корпусе прозы уже при длине 40 слов, к ним присоединятся еще 47 слов — *бретер, вызвать, выстрелить, извиниться, Лермонтов, обида, оскорбить, отказаться, пощечина, разжаловать, ранить, скандал, соперник, стреляться, трус, убить, удовлетворение, целить* и т. п. Переход к длине фрагмента в 1000 слов покажет уменьшение значений S у слов, обозначающих стандартизованные ситуации вроде дуэли; напротив, появятся сюжетные текстуальные связи, приуроченные к конкретному тексту (*болгарин — Кунцево, резать — лягушка — нигилист*). Это свидетельствует о качественно ином — **большом интервале**. Что касается среднего интервала, то в качестве окончательной была принята длина фрагмента в 40 слов.

Для конкретных пар слов результаты выдаются в следующем виде:

волосы густой	195	3156	1687	48	13.90
лежать постель	281	5303	2142	45	29.66
лес куст	99	3692	1022	28	9.85

Левый столбец показывает число общих адресов двух слов, во втором столбце дается частота (число адресов) первого слова, в третьем — частота второго слова, в четвертом столбце указывается величина S, в пятом — величина m.

Как видим, средний интервал во многом наследует результаты минимального интервала, ср. пару *волосы — густой*, пары *ахиллесов пята* ($S = 170$), *выеденный яйцо* ($S = 129$), *несолоно хлебать* ($S = 123$), *отложной воротничок* ($S = 132$). Пару *лежать — постель* можно было бы обнаружить на **малом интервале** (длина фрагмента 3–5 слов), ср. также пары *пазуха — вынуть, скрестить — грудь, ухаживать — больной, шесток — сверчок, юг — север*. Однако большинство найденных

пар обнаруживаются именно на среднем интервале: не только *лес — куст*, но и *лес — дерево*.

Нередко при этом за парами слов видится некое более широкое объединение. У слова *волосы* обнаружено 913 текстуальных связей, из них 230 имеют $S = 3$, т. е. на самом пороге статистической значимости, но и в этой группе находим слова, явно связанные с волосами (*брюнет, дергать, копна, кружок, лезть, лен, немытый, непокорный, отвиснуть, погладить, причесываться, пушок, разглаживать, редеть, убранный*). Унаследованные от меньших интервалов связи нашего слова доминируют в группе максимальных значений *S*: начиная со связи со словом *прядь* ($S = 87$) и далее по мере уменьшения *S*: *седой, дыбом, ерошить, русый, черный, вьющийся, густой, проседь, зачесанный, каштановый, белокурый, длинный, расчесывать, растрепанный, всклокоченный, распущенный, причесанный, смоль* ($S = 31$). Уже при $S = 47$ появляется первое пополнение среднего интервала — слово *лицо*, за которым следуют *борода, лоб, висок, глаза, рост, голова, губы, затылок, карий, голубой, нос* ($S = 20$). Постепенно проясняется весьма характерная для прозы совокупность слов, описывающих внешность человека.

Такая совокупность (кластер) скорее исключение в прозе. В целом же можно заключить, что работа ДСА в среднем интервале привела к порождению грандиозной сети текстуальных связей многих тысяч слов. Связи отдельного слова могут вести к разным участкам этой сети.

Сравним текстуальные связи двух слов *Невский проспект* и *Кузнецкий мост* (словом в ДСА может быть и словосочетание, единство которого доказано на минимальном интервале). Частота первого слова — 357, частота второго — всего 53; контраст частот заставляет ожидать и контраста числа текстуальных связей: у *Невского проспекта* находим 119 связей, у *Кузнецкого моста* их всего 12. Есть два слова (*магазин* и *экипаж*), с которыми связаны оба топонима. Связь с *магазин* подкрепляется у *Невского проспекта* связями с *модный, пассаж*, у *Кузнецкого моста* аналогичны связи с *купить, готовый*. Связь *Кузнецкого моста* с *экипажем* остается единственной, у *Невского проспекта* она подкрепляется связями с *английский, бал, великолепный, извозчик, изящный, карета, кататься, коляска, копыто, кучер, мелькать, мостовая, мчаться, набережная, остановиться, отправиться, пара, Петербург, пешеход, пешком, площадь, понестись, развалиться, разъезжать, рысак, сани, толпа, тротуар, улица, час, шляпа, щегольской*. Тридцать два слова, одновременно связанных с *экипажем* и с *Невским*

проспектом, несомненно свидетельствуют в пользу существования особого кластера «Невский проспект». Связь *Невского проспекта* с *гулять* подкрепляется связями с *встретить, зайти, обед, праздничный, прогуливаться, публика и* еще семью связями из числа перечисленных выше: *кататься, отправиться, пешком, толпа, тротуар, улица, час*. Таким образом, обнаруженный кластер стал еще шире, В Москве слово *гулять* связано только с *Нескучный сад*.

У *Невского проспекта* есть еще один «топографический» аспект — его связи с *Английской набережной, Большой Морской, Летним садом, Лиговкой*. У *Кузнецкого моста* аналогична связь с *Тверским бульваром*.

Сравнение слов *магазин* и *лавка* демонстрирует одну из проблем, с которой сталкивается ДСА. У первого слова зафиксировано 198 текстуальных связей, у второго слова находим 260 связей. Оба слова непосредственно связаны друг с другом ($S = 11$), кроме того есть 38 слов, одновременно связанными с двумя нашими словами (связи второго порядка), *базар, бакалейный, бульвар, вывеска, город, гостиный, европейский, зайти, китаец, китайский, книжный, колониальный, купец, купить, лавочка, мастерская, материя, мостовая, окно, переулок, покупатель, покупать, приказчик, прилавок, продаваться, продавец, ресторан, табачный, товар, торговать, торговец, торговля, торговый, тротуар, узенький, улица, фруктовый, харчевня*.

Казалось бы, семантическая близость двух слов налицо. Однако, обратившись к текстуальным связям слова *лавка* и на минуту отказавшись от «асемантического» формализма, мы тут же обнаружим аномалию. В списке 16 связей, упорядоченных по уменьшению S, находим связи со словами, никак не вписывающимися в сферу торговли (ниже они отмечены звездочкой), *мелочной* ($S = 51$), *товар, книжный, изба*, полати*, овощной, печь*, покупатель, бакалейный, гостиный, мясной, купец, приказчик, сидеть*, торговать, пол**. Идя дальше по списку, найдем и другие аномальные слова: *присесть, стряпной, подостлать, зыбка, печка, садиться, сесть*. Знание русского языка подскажет — *лавка* соотносится с двумя семантическими объектами: 1) торговая точка, 2) скамья. Рассмотренная аномалия есть лишь отражение этого факта в сети текстуальных связей.

В процессе настоящего исследования для нескольких десятков слов мы отошли от строгого формализма и с учетом нашего знания языка создали «искусственные» слова, соответствующим образом маркируя их в тексте. В данном случае текстуальные связи определены для двух слов *лавка* (=лавка-1 с частотой 1020 и с 293 связями) и *лавкаъ* (=лав-

ка-2 с частотой 470 и с 165 связями). Первое слово сохраняет «торговые» связи, второе — значительно увеличивает *S*, расширяет круг «бытовых» связей (любопытны связи со словами *Кутузов* и *Пугачев*). У слов *лавка и лавкаъ* есть общие связи со словами *баба, войти, вход, горшок, дверь, двор, деревянный, лучина, сени, стена, стоять, угол, хозяин, шапка*. Точно такой же операции расщепления подверглось слово *лавочка*.

По текстуальным связям нетрудно понять мнемонику следующих трех слов:

коса-ж *(213 связей), лента, заплетать, вплести, заплести, расплетать, заплетенный, русый, прядь, растрепаться, кокошник, расчесывать, густой, обрезать, длинный, волосы, монисто;*

коса-сх *(60 связей), точило, косить, серп, косец, лезвие, точить, брусок, грабли, косарь, цеп, трава, топор, кузнец;*

коса-в *(6 связей), песчаный, таможенный, неводчик, камыш, мелькать, море.*

Создание «искусственных» слов может быть мотивировано и учетом дискурсивных функций, ср.:

прощать *(54 связи), простить, прощение, извинять, вина, виноват, враг, грешник, Бог, великодушный, забывать;*

прощай *(242 связи), уходить, увидеться, свидание, лихо, целовать, завтра, провожать, целоваться, спасибо, заболтаться, кланяться, пора, пойти, поцеловать, проститься, уйти.*

Работа ДСА в среднем интервале привела к созданию сети текстуальных связей — новому объекту корпусной лингвистики, открывающему широкие возможности изучения текстов, далеко выходящие за пределы традиционных интересов лингвистов. От изучения корпуса как целого можно переходить к анализу авторских, жанровых и тематических подкорпусов (желательно большого объема).

Слово *лес* у русских писателей встречается с разной частотой. На 100 тыс. словоупотреблений частота этого слова составляет у Тургенева 24, у Достоевского 6, у Толстого 31. Тем не менее у этих авторов появляется одна и та же тройка слов, взаимно связанных текстуальными связями: *лес, дерево, куст*. У трех авторов слово *лес* связано со словами *кругом* и *поле*. Но за пределами этой пятерки начинаются различия. У Достоевского с двумя другими авторами находим по одной общей

связи, Тургенев и Толстой для слова *лес* дают 24 общих связей: *береза, ветер, виднеться, густой, дорога, дуб, зарасти, зеленый, земля, лист, небо, овраг, опушка, покрытый, поляна, река, ружье, солнце, сосновый, темнеть, трава, туман, широкий, яркий*.

В общем корпусе прозы слово *виднеться* имеет 580 связей. Лишь одна связь (со словом *черный*) обнаружена в подкорпусах всех трех авторов, у Тургенева и Толстого находим 20 общих связей данного глагола: *белый, ветер, высокий, из-за, из-под, красный, крыша, куст, лес, направо, небо, острый, поле, полоса, сквозь, сплошной, стена, темнота, туман, тянуться*. Фигура «наблюдателя» очень ярко отражена в этой группе слов.

Поскольку ДСА не привязан к конкретному языку, видны перспективы межъязыкового сравнения корпусов текстов.

Литература

1. *Шайкевич А. Я., Андрющенко В. М., Ребецкая Н. А.* (2013, 2016), Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг., тт. 1–2. М.

References

1. *Shaikevich A. Ya., Andrjuscenko V. M., Rebetskaya N. A.* (2013, 2016), Distributivno-statisticheskij analiz yaxyka russkoj prozy 1850–1870-h gg., tt. 1–2 [Distributional statistical analysis of the language of Russian prose of the 1850–1870s., vol. 1–2]. Moscow.

Шайкевич Анатолий Янович
Институт русского языка РАН (Россия)
Shaikevich Anatole
Institute for Russian Language RAS (Russia)
E-mail: ishaikev@mail.ru

В. В. Шеремет
V. V. Sheremet

ПОСЕССИВНЫЕ КОНСТРУКЦИИ С «У»-ЛОКАЛИЗАТОРОМ И ИХ ЭКВИВАЛЕНТЫ В АРАБСКОМ ЯЗЫКЕ (НА МАТЕРИАЛЕ ПАРАЛЛЕЛЬНОГО КОРПУСА)

THE STRUCTURES WITH RUSSIAN PREPOSITION «У» AND SPECIFICS THEIR TRANSLATION IN THE TURKISH AND THE ARABIC LANGUAGES (ON THE BASE OF THE PARALLEL GRAMMATICAL CORPUS)

Аннотация. В статье языка рассматривается выборка из созданного вручную параллельного русско-арабского корпуса по посессивной конструкции «у + X (Р.п.) + бытийный предикат + Y (И.п.)» на семантическом и синтаксическом уровнях. Сквозь призму арабского языка анализируется семантико-синтаксическая дифференциация описания моделей русского языка, на основе которой проецируется результат исследования – модель русского языка, находящая регулярное соответствие в арабском языке.
Ключевые слова. Русский язык, арабский язык, родительный падеж, предлог «у», параллельный корпус.

Abstract. In the article the translation of the Russian language construction with preposition «у» into the Arabic language as a one of the less learned aspects of comparative grammar is studied. The research on the base of self-created parallel grammatical corpus is provided. Basing on the analysis of samples the specifics of the transfer of this construction in the Arabic language is discussed.
Keywords. The Russian language, the Arabic language, genitive case, «у+»-construction, parallel corpus**.**

1. Введение

Посессивность — универсальная по своей денотативной природе категория, которая интерпретируется в различных языках различными формально-семантическими средствами [Бондарко 1996: 99].

В русском языке наблюдается два способа её реализации: предикативный и адъективный [Бондарко 1996: 3]. Основным формальным средством фиксации предикативной посессивности в русском языке являются «бытийно-посессивные» конструкции с «у»-локализатором: «у X есть Y» [Арутюнова, Ширяев 1983].

Возникающая в данных конструкциях проблема неоднозначности особенно релевантна при составлении эквивалентных моделей, требующих наличия в конструкции исходного языка уникального состава параметров. В свою очередь, ориентация на прикладное использова-

ние моделей, мотивирующая формализацию всех релевантных компонентов, не позволяет обращаться к контексту ввиду его многоаспектности и требует отличного от этого подхода.

Исследуемые конструкции не получили должного освещения в сопоставительных изучениях грамматики русского и арабского языков. В работах арабистов, как правило, изучается специфика именной функции родительного падежа [Гранде 2001: 325–338, Габучан 2000: 21, и др.]. Предложно-падежные сочетания рассматриваются факультативно, в рамках сопоставления семантических полей предлогов русского и арабского языков [Гранде 2001: 400–411].

В связи с этим нашей **целью** стало рассмотрение вариантов модели «у + X (Р. п.) + есть + Y (И. п.)» русского языка, находящих регулярные соответствия в арабском языке.

Цель достигалась на основе решения **задач**: 1) получить объективную и верифицируемую фактологическую базу исследования; 2) описать стратегию передачи посессивности русского языка в рассматриваемой конструкции и её арабском эквиваленте на семантическом и синтаксическом уровнях; 3) выявить модель их лингвистического соответствия.

Как показывает опыт исследователей, корпус текстов в наибольшей степени соответствует современным требованиям к уровню представления материала [Кибрик 2006: 16–46, Брыкина 2009].

В связи с этим фактологической базой исследования был избран параллельный русско-арабский корпус.

Стремление сформировать возможность получения исчерпывающей информации о конструкциях с падежами русского языка и их эквивалентов в арабском языке стало отличительной чертой создаваемого корпуса по сравнению с существующими русско-арабскими ресурсами (http://opus.lingfil.uu.se/). Исходя из этого, его разметка осуществлялась на базе системного, ономасиологического принципа: ключевым элементом корпуса является не слово или предложение, а монопредикативное, монопропозитивное выражение, содержащее все релевантные для передачи смыслосодержания конструкций с падежом элементы: девять парадигматических и семь синтагматических параметров, а также контексты выражений.

Основой выделенной номенклатуры параметров стали научные работы, описывающие падеж в ономасиологическом аспекте, а также списки семантических ролей, тематических классов и семантических типов Е. В. Падучевой, Ю. Д. Апресяна, Н. Ю. Шведовой.

Выбор источника материала неслучаен: 1) стремление наблюдать живую речь обусловило обращение к художественной литературе; 2) нацеленность на создание мультиязыкового корпуса, а также специфика переводной литературы на восточные языки сузили круг поиска источника материала до классики русской литературы XIX и первой половины XX века; 3) трудоёмкость ручной обработки текстов большого и малого объёма обусловила обращение к жанрам среднего размера — повестям.

Грамматическая ориентированность исследования в свете вышесказанного, позволила привлечь в качестве материала повесть А. С. Пушкина «Капитанская дочка» на русском языке и в переводе на арабский язык, выполненным сирийским писателем и переводчиком Сами ад-Дуруби (سامي الدروبي).

Общее число элементов, доступных для выборки, — 12 тысяч выражений.

Статья фиксирует результаты на экспериментальном этапе исследования и, иллюстрируя исследовательские возможности корпуса, не претендует на полноту описания языковых явлений в связи с малой частотностью выборки.

В связи с: 1) малочисленностью ресурсов, предоставляющих переводы русской художественной литературы на арабский язык в текстовом формате; 2) отсутствием средств распознавания арабского языка, позволяющих получить в среднем менее 10 % ошибочно распознанных знаков, арабский текст набирался в текстовом формате вручную. После нормализации русский и арабский тексты выравнивались при помощи программы, написанной на языке Python. Выровненные тексты размечались вручную и сохранялись в excel-файле. В связи со спецификой отображения арабской кодировки на сервере данные вручную переносились в базу данных MySQL. Поиск в базе данных по одному либо ряду параметров реализовывался на языке PHP и работает в тестовом режиме на сайте www.параллельныйкорпус.рф.

2. Результаты исследования

Выборка по конструкциям с предлогом «у» составила 94 выражения. В них конструкция «у + X + есть + Y», согласно параметрической разметке, классифицируется наличием как **посессивных** конструкций, так и конструкций **местоположения** («он был у Пугачёва»), **бенефицианта** («у нас разграбили»), **агенса** («у меня шары летали»), **свойства** («у меня были стянуты руки»).

Из данных выражений 12 обладают посессивной семантикой. Можно наблюдать три стратегии их перевода в арабском языке, формирующие соответствующие эквивалентные модели, что требует привлечения дополнительного дифференцирующего семантико-синтаксического описания.

1) Модель «имя (агенс) + предикат обладания + актант». «*У[1] Швабрина[2] было[3] несколько[4]* книг[5]» — «كان[2] شفابرين[3.2] يملك[4] عدداً من الكتب[5]» [ka:na ʃfabri:n jamlaku ʕadadan min al-kutubi] «*был[3.1] Швабрин[2] обладал[3.2] рядом[4] книг[6]*».
2) Модель «имя (агенс) + предикат раскрывающий содержание русской связки предикат + актант». «*Было[1] у[2] Ивана Кузмича[3] совещание[4]*» — «ثم ما لبثت أن علمت أن اجتماعا[4] قد عُقد اثناء غيابها» [θumma ma: labaθat ʔan ʕalamat ʔanna ʔidʒ͡ʒtima:ʕan qad ʕuqida ʔaθna:ʔa ɣajabiha] «*потом вскоре она узнала, что собрание[4] было проведено в её отсутствие*».
3) Использование предлогов «ل» [li], «لدى» [lada:], указывающих на обладателя: «*Солдат[1] у[2] нас[3] довольно[4]*» — «لدي[2]نا[3] عدد كاف[4] من الجنود[1]» [ladajna: ʕadadun ka:fun min al-dʒ͡ʒunudi] «*у[2] нас[3] количество[4.1] достаточное[4.2.] из солдат[1]*».

Рассмотрим первую стратегию более подробно.

Анализ параметрического описания структуры исходного выражения показывает её эмерджентность. На уровне синтаксем она состоит из сирконстанта («у Швабрина»), указывающего на местоположение, и предиката, связанного с актантом отношением бытийности («было несколько книг»), и только на уровне предложения реализуется бицентрическая структура посессивности, в которой существующее явление соотносится со сферой субъекта, который может им обладать. При этом реализуется коммуникативная стратегия понижения статуса субъекта.

В арабском примере наблюдается использование глагола собственно посессивной семантики, стилистически не свойственного в подобных ситуациях русскому языку. Это, при семантической изоморфности первого актанта арабского выражения сирконстанту русского, трансформирует семантический тип второго актанта арабской фразы, формируя в ней семантический аспект посессивности.

Исходя из анализа выборки, данная стратегия реализуется, если элементом выражения является объект, которым можно обладать физически (например, книга). Данное значение не было включено

в параметрическую номенклатуру корпуса, в связи с этим значения параметров корректировались исходя из эмпирической необходимости.

Обращение к синтаксису показывает, каким образом выбранная стратегия передачи посессивности закрепляется в структурах целевого языка. Выраженный дополнением в родительном падеже сирконстант «у Швабрина» в арабском языке становится первым актантом: «شفابرين» — «Швабрин», приобретает формант именительного падежа и роль подлежащего. Ключевая лексема первого актанта русского выражения в именительном падеже — «несколько», в свою очередь, преобразуется во второй актант в арабском предложении, участвует в формировании сложного дополнения и принимает арабский винительный падеж формально выраженной фатхой с танвином «ً»: «عدداً» — «ряд, несколько».

3. Заключение

В результате исследования было установлено, что фиксация эквивалентных моделей требует привлечения дополнительного параметра: возможности / невозможности физического обладания в качестве предмета или субъекта.

Достаточной глубиной дифференциации в русском языке обладает модель: «у+X (Р. п.)+есть+Y(И. п.)», где Y — конкретное неодушевлённое имя, относящееся к группе имён, которыми можно обладать физически, а X — одушевлённое или подразумеваемое таким имя, которое может участвовать в ситуациях обладания в качестве субъекта.

Эквивалентной ей в арабском языке является модель «X (И. п.) + глагол обладания + Y(В. п.)».

Дальнейшее корпусное изучение грамматики может способствовать расширению существующих знаний об лингвистических русско-арабских соответствиях. Эквивалентное моделирование требует достаточно большой глубины дифференциации, что обусловливает необходимость дальнейшего расширения корпуса.

Литература

1. *Бондарко А. В.* (1996), Теория функциональной грамматики. Локативность. Бытийность. Посессивность. Обусловленность. СПб.
2. *Арутюнова Н. Д., Ширяев Е. Н.* (1983), Русское предложение. Бытийный тип. М.
3. *Кибрик А. Г., Брыкина М. М., Леонтьев А. П., Хитров А. Н.* (2006), Русские посессивные конструкции в свете корпусно-статистического исследования. М.

4. *Брыкина М. М.* (2009), Языковые способы кодирования посессивности (на материале корпусного исследования русского языка. М.
5. *Гранде Б. М.* (2001), Курс арабской грамматики в сравнительно-историческом освещении. М.
6. *Габучан Г. М.* (2000), Арабское словоизменение. М.

References

1. *Bondarko A. V.* (1996), Teorija funkcional'noj grammatiki. Lokativnost'. Bytijnost'. Posessivnost'. Obuslovlennost' [The theory of functional grammar. Locality. Existentiality. Possessiveness. Conditionality]. SPb.
2. *Arutyunova N. D., Shiryaev E. N.* (1983), Russkoe predlozhenie. Bytijnyj tip [The Russian sentence. The existential type]. Moscow.
3. *Kibrik A. G., Brykina M. M., Leontiev A. P., Hitrov A. N.* (2006), Russkie posessivnye konstrukcii v svete korpusno-statisticheskogo issledovanija [The Russian possessive constructions in the light of corpus-statistical study]. Moscow.
4. *Brykina M. M.* (2009), Russkie posessivnye konstrukcii v svete korpusno-statisticheskogo issledovanija [The language methods of possessivness coding (based on the corpora language studying of the Russian language]. Moscow.
5. *Grande B. M.* (2001), Kurs arabskoj grammatiki v sravnitel'no-istoricheskom osveshhenii [Course of Arabic grammar in comparative-historical illumination]. Moscow.
6. *Gabuchan G. M.* (2000), Arabskoe slovoizmenenie [The Arabic inflection]. Moscow.

Шеремет Виталий Владимирович
ФГАОУ ВО «Крымский федеральный университет им. В. И. Вернадского»
Sheremet Vitaliy Vladimirovich
V. I. Vernadsky Crimean Federal University
E-mail: sheremetvitaliy@gmail.com

Т. Ю. Шерстинова
T. Y. Sherstinova

ПОДХОДЫ К ТЕМАТИЧЕСКОМУ АННОТИРОВАНИЮ ЗВУКОЗАПИСЕЙ ПОВСЕДНЕВНОГО БЫТОВОГО ОБЩЕНИЯ В КОРПУСЕ «ОДИН РЕЧЕВОЙ ДЕНЬ»

APPROACHES TO THEMATIC ANNOTATION OF EVERYDAY SPOKEN DISCOURSE IN "ONE DAY OF SPEECH" CORPUS

Аннотация. Рассматриваются подходы к тематическому аннотированию эпизодов речевой коммуникации корпуса «Один речевой день» (ОРД), который содержит большую коллекцию звукозаписей повседневного бытового общения на русском языке, выполненных в естественных условиях. Тематическое аннотирование материалов корпуса ОРД позволит осуществлять поиск речевого контента по новым параметрам, а также исследовать тематическое разнообразие повседневного речевого общения. Важным требованием к тематическому аннотированию, выдвигаемому при работе с мультимедийным контентом, является сегментация аудиофайлов на фрагменты, относительно однородные по теме разговора. Ставится задача разработки расширенного шаблона многоуровневого аннотирования речевого материала, включающего в себя дополнительные уровни, характеризующие тематику разговоров и эмоциональную окраску речи.
Ключевые слова. Современный русский язык, повседневное речевое общение, речевой корпус, мультимедиа, аннотирование, темы речевого общения, устный дискурс, эмоциональная речь, социолингвистика.

Abstract. The paper describes approaches to thematic annotation of everyday private conversations made in natural settings and gathered in the corpus "One Day of Speech" (ORD), which is the biggest collection of audio recordings of Russian everyday speech communication. Thematic annotation of corpus data will provide new search capabilities, and will allow to explore thematic variety of everyday spoken discourse. An important requirement for thematic annotation of multimedia content is segmentation of audio files into fragments that are relatively homogeneous from a thematic point of view. The necessity of an extended template for multilevel speech annotation, which will take into account the topic of conversation and the emotional state of speakers, is discussed.
Keywords. Modern Russian language, everyday speech communication, speech corpus, multimedia, annotation, topics of speech communication, oral discourse, emotional speech, sociolinguistics.

Корпус ОРД содержит наиболее крупную коллекцию звукозаписей повседневного бытового общения на русском языке, выполненных в естественных условиях [Богданова-Бегларян и др. 2015]. Запись осуществлялась в Санкт-Петербурге в 2007–2016 гг., общий ее объем составляет 1250 часов звучания (2800 коммуникативных макроэпизодов). Текстовые расшифровки получены для 17 % звукозаписей корпуса (480 макроэпизодов) и насчитывают 1 млн словоупотреблений.

Лингвистическое аннотирование выполнено на подкорпусе объемом в 125 000 словоупотреблений [Богданова-Бегларян и др. 2016; 2017].

Все звукозаписи корпуса ОРД проиндексированы на уровне макроэпизодов (крупных фрагментов речевой коммуникации) по месту коммуникации, типу коммуникативного сценария и социальной роли говорящих, что позволяет осуществлять поиск речевого материала по этим параметрам. Разметка данных осуществляется в профессиональной программе мультимедийного аннотирования ELAN, разработанной в Институте психолингвистики Макса Планка (Неймеген, Нидерланды), которая поддерживает визуализацию осциллограммы звукового сигнала и позволяет осуществлять привязку элементов аннотирования непосредственно к звуковой волне [ELAN 2017].

Сплошное тематическое аннотирование речевого материала ранее не проводилось. Некоторая информация о тематике разговоров может быть получена из поля описания *SceneName* базы данных речевого корпуса. Однако это поле является факультативным, оно заполнено лишь для части звуковых файлов, относится ко всему макроэпизоду и не является нормализованным. Поэтому осуществлять поиск по данному полю в настоящее время не представляется возможным.

В рамках реализации проекта РНФ «Русский язык повседневного общения: особенности функционирования в разных социальных группах» (№ 14–18–02070) на материале аннотиованной подвыборки в 125000 словоупотреблений на основе анализа лексики (частотных словарей), были получены предварительные выводы о тематических ценностях и приоритетах повседневного речевого общения. В частности, популярными оказались следующие темы: «питание/еда», «компьютеры, автомобили и технические гаджеты», а также «здоровье» и «условия труда». Очевидна перспективность продолжения исследования материалов корпуса ОРД в этом направлении. Поэтому кажется целесообразным осуществление сплошного тематического индексирования всех звукозаписей корпуса ОРД по теме речевого общения и проведение на его основе масштабного системно-статистического анализа, а также изучение языковых средств, присущих обсуждению той или иной ключевой темы.

В настоящий момент ставится задача разработки расширенного шаблона многоуровневого аннотирования речевого материала корпуса ОРД, включающего в себя дополнительные уровни, характеризующие тематику разговора. Перспективным было бы и включение в расширенный шаблон аннотирования еще одного уровня, отражающего

эмоциональную окраску речи, что позволило бы в дальнейшем проводить исследование тональности — личностного отношения говорящих к предмету обсуждения.

При этом важным требованием к тематическому аннотированию, выдвигаемому при работе с мультимедийным контентом, является сегментация аудиофайлов на фрагменты, относительно однородные по теме разговора. В корпусе ОРД такой единицей являются *микроэпизоды*, которые выделяются для всех расшифрованных звукозаписей корпуса [Шерстинова 2015]. При этом введение расширенного шаблона аннотирования может привести к необходимости пересмотра (коррекции) уровня микроэпизодов для уже расшифрованного подкорпуса звукозаписей, чтобы привести речевой материал к единому формату представления данных.

Для звукозаписей корпуса с текстовыми расшифровками, уже отсегментированных на тематически однородные микрэпизоды, возможно проведение тематического аннотирования с использованием статистических методов автоматического извлечения ключевых слов. Однако, учитывая высокую эллиптичность повседневной бытовой речи, когда «тематические слова» разговора довольно часто опускаются, эффективность применения таких методов требует специальной проверки. Кроме того, абсолютное большинство существующих в настоящее время программ ориентировано на письменные тексты, состоящие из предложений, поэтому «квантами» автоматического анализа для извлечения ключевых слов являются предложения. В спонтанной устной речи членение на единицы, соответствующие предложениям, далеко не всегда можно выделить однозначно, поэтому при обработке расшифровок устной речи имеет смысл ориентироваться не на предложения, а на другие показатели — длительные (разграничительные) паузы в потоке речи или определенное количество словоупотреблений.

Для тех звукозаписей корпуса, которые еще не имеют текстовой расшифровки (это примерно 83% коммуникативных макроэпизодов корпуса), тематическое аннотирование подразумевает экспертное прослушивание речевого материала и его параллельную сегментацию на микроэпизоды, относительно однородные фрагменты с точки зрения тематики разговора или же отражающие изменения в коммуникативной ситуации (например, телефонный разговор, прервавший естественное течение беседы). Для упрощения этой задачи может быть использована процедура автосегментации речевого сигнала на полезный сигнал (речь) и паузы, которую предоставляет программа ELAN.

Разговор по каждой теме имеет физические границы (начала и конца по файлу звукозаписи), что дает потенциальную возможность статистически оценить время (с поправкой на паузы), затраченное собеседниками на каждую из тем. Поэтому в будущем станет возможным получение статистических данных как о количестве коммуникативных эпизодов, в которых анализируемая тема была затронута, так и о конкретном времени ее обсуждения.

Часть разговоров могут характеризоваться одновременно отнесением к нескольким основным темам, как основным, так и сопутствующим. Определенную сложность при тематическом аннотировании речевого материала может вызвать одновременное обсуждение группой собеседников нескольких тематически далеких тем, а также аннотирование тех ситуаций, когда анализируемый коммуникативный эпизод состоит из нескольких параллельных разговоров (например, застольные разговоры или разговоры в офисе).

Следует также иметь в виду, что довольно большую долю повседневной речевой коммуникации составляют изолированные реплики (например, *Я ушёл*) или изолированные микродиалоги (*Дай спички! — Держи!*). Тематически-изолированные реплики или микродиалоги нередко нарушают формальную целостность разговора, но сами по себе также интересны для лингвистического анализа.

Тематическое аннотирование корпуса ОРД даст возможность проводить поиск данных по теме бытового общения (например, разговор «о здоровье», «о личных взаимоотношениях», «о работе», «о новых гаджетах» и др.). Помимо оптимизации поисковых возможностей тематическое аннотирование аудиозаписей имеет и другую важную цель, которая связана с исследованием тематического разнообразия повседневных бытовых разговоров, а именно — какие темы чаще других становятся предметом устного общения у жителей большого российского города в начале XXI века. Каждая историческая эпоха характеризуется своим набором приоритетных тем, однако статистический анализ реального тематического разнообразия стал возможен только в наши дни благодаря корпусной организаций аутентичного речевого материала.

Литература

1. *Богданова-Бегларян Н. В., Шерстинова Т. Ю., Асиновский А. С., Блинова О. В., Маркасова Е. В., Рыко А. И.* (2015), Звуковой корпус русского языка: новая мето-

дология анализа устной речи // Язык и метод. Русский язык в лингвистических исследованиях XXI века, вып. 2, ред. Д. Шумска, К. Озга, Краков, Изд-во Ягеллонского ун-та, с. 357–372.
2. *Богданова-Бегларян Н. В., Шерстинова Т. Ю., Блинова О. В., Мартыненко Г. Я.* (2017), Корпус «Один речевой день» в исследованиях социолингвистической вариативности русской разговорной речи // Анализ разговорной русской речи (AP^{3-}2017). Труды седьмого междисциплинарного семинара. СПб.: Политехника-принт, с. 14–20.
3. *Богданова-Бегларян Н. В., Шерстинова Т. Ю., Баева Е. М., Блинова О. В., Мартыненко Г. Я., Ермолова О. Б., Рыко А. И.* и др. (2016), Русский язык повседневного общения: особенности функционирования в разных социальных группах. Коллективная монография. СПб.: Издательство «ЛАЙКА», 2016. 244 с.
4. ELAN — Linguistic Annotator (2017). URL: http://www.mpi.nl/corpus/html/elan/index.html.
5. *Шерстинова Т. Ю.* (2015), Прагматическое аннотирование коммуникативных единиц в корпусе ОРД: микроэпизоды и речевые акты / Захаров В. П., Хохлова М. В. (ред.) // Труды международной конференции «КОРПУСНАЯ ЛИНГВИСТИКА — 2015». СПб.: СПбГУ, с. 436–445.

References

1. *Bogdanova-Beglarjan N. V., Sherstinova T. Ju., Asinovskij A. S., Blinova O. V., Markasova E. V., Ryko A. I.* (2015), Zvukovoj korpus russkogo jazyka: novaja metodologija analiza ustnoj rechi [Sound Corpus of Russian: New Methodology of Oral Speech Analysis]. In: Jazyk i metod. Russkij jazyk v lingvisticheskih issledovanijah XXI veka [Language and Methodogoly. The Russian Language in Linguistics Studies of the XXIth Century], iss. 2, D. Shumska, K. Ozga (eds.), Krakov, Wydawnictwo Uniwersytetu Jagiellońskiego, pp. 357–372.
2. *Bogdanova-Beglarjan N. V., Sherstinova T. Ju., Blinova O. V., Martynenko G. Ja.* (2017), Korpus "Odin rechevoj den'" v issledovanijah sociolingvisticheskoj variativnosti russkoj razgovornoj rechi ["One Day of Speech" corpus in studies of sociolinguistic variability of Russian colloquial speech]. In: Analiz razgovornoj russkoj rechi [Analysis of colloquial Russian speech] (AR3–2017), Trudy sed'mogo mezhdisciplinarnogo seminara [The works of the 7rd Interdisciplinary Seminar]. St Petersburg, Politehnika-print, pp. 14–20.
3. *Bogdanova-Beglarjan N. V., Sherstinova T. Ju., Baeva E. M., Blinova O. V., Martynenko G. Ja., Ermolova O. B., Ryko A. I.* et al. (2016), Russkij jazyk povsednevnogo obshhenija: osobennosti funkcionirovanija v raznykh social'nykh gruppakh [Russian everyday language in different social groups]. St Petersburg, Lajka, 244 p.
4. ELAN — Linguistic Annotator (2017). Available at: http://www.mpi.nl/corpus/html/elan/index.html
5. *Sherstinova T. Ju.* (2015), Pragmaticheskoe annotirovanie kommunikativnykh jedinic v korpuse ORD: mikrojepizody i rechevye akty [Approaches to pragmatic annotation in the ord corpus: micro episodes and speech acts]. In: Zaharov V. P., Khokhlova M. V.

(eds.) Trudy mezhdunarodnoj konferencii “KORPUSNAJA LINGVISTIKA — 2015” [Proc. of the Int. Conference “CORPUS LINGUISTICS — 2015”]. St Petersburg, St Petersburg State University, pp. 436–445.

Шерстинова Татьяна Юрьевна
Санкт-Петербургский государственный университет (Россия)
Tatiana Sherstinova
St Petersburg State University (Russia)
E-mail: sherstinova@gmail.com

П. М. Эйсмонт
P. Eismont

«КОНДУИТ»: КОРПУС УСТНЫХ ДЕТСКИХ ТЕКСТОВ[1]

"KONDUIT": CORPUS OF CHILD ORAL NARRATIVES

Аннотация. В статье представлен проект КОрпуса Неподготовленных Детских Устных (Извлеченных) Текстов «Кондуит». Корпус состоит из 213 однотипных устных неподготовленных текстов детей в возрасте 2;7–7;6 лет, каждый из которых имеет метатекстовую, структурную, синтаксическую и семантическую разметку. Основной целью создания корпуса является получение базы устных неподготовленных текстов для изучения процессов формирования навыков построения связного текста в онтогенезе.
Ключевые слова. Детская речь, связный текст, устный корпус, связность, цельность, глагольная структура.

Abstract. The project of a new corpus of child oral unprepared elicited narratives "Konduit" is described. The corpus consists of 213 oral narratives based on the same story or a similar sequence of actions, elicited by Russian native children at the age of 2;7–7;6. Each text has semantic, syntactic and discourse annotations. The main purpose of this corpus is to become an effective and useful tool for the studying of narrative acquisition in Russian.
Keywords. Child language, oral corpus, narrative, coherence, cohesion, verb structure.

В последнее время в отечественной лингвистике созданию устных корпусов уделяется значительное внимание — расширяется устный подкорпус Национального корпуса русского языка [Гришина, Савчук 2009], развивается проект «Один речевой день» [Богданова-Бегларян 2016], продолжается работа над полной фонетической расшифровкой Корпуса русских спонтанных текстов [Венцов и др. 2013]. Несмотря на бесспорную значимость для изучения современного русского языка и процессов порождения и восприятия речи в обществе и высокий уровень работы над этими корпусами, каждый из них уделяет большее внимание решению какой-либо конкретной задачи, а все они включают в себя в первую очередь образцы речи взрослых носителей русского языка.

Специфика КОрпуса Неподготовленных Детских Устных (Извлеченных) Текстов «Кондуит» заключается в том, что в нем представлены образцы устной связной речи детей разных возрастных групп от 2;7 до 7;6 лет. Основной целью создания корпуса является получение базы устных неподготовленных текстов для изучения процессов фор-

[1] Публикация подготовлена в рамках поддержанного РГНФ научного проекта №16-04-50114.

мирования навыков построения связного текста в онтогенезе. Всего в корпусе представлено 213 текстов, которые были получены в результате проведения серии экспериментов с детьми, посещающими дошкольные образовательные учреждения г. Санкт-Петербурга. Собранные тексты поделены на 5 возрастных групп: в возрасте от 2,7 до 3,6 (далее — 1 группа) — 37 детей, в возрасте от 3,7 до 4,6 (далее — 2 группа) — 50 детей, в возрасте от 4,7 до 5,6 (далее — 3 группа) — 49 детей, в возрасте от 5,7 до 6,6 (далее — 4 группа) — 42 ребенка и в возрасте от 6,7 до 7,6 (далее — 5 группа) — 35 детей.

При проведении экспериментов с детьми разного возраста был использован разный экспериментальный дизайн: с детьми 1 группы эксперимент проходил в форме игры, в которой дети должны были называть действия, производимые экспериментатором при помощи игрушек-бибабо; дети 2 группы рассматривали с экспериментатором книжку с картинками (сказка В. Г. Сутеева «Три котенка», рис. автора), рассказывая, что на них нарисовано; дети 3, 4 и 5 групп смотрели мультфильм без звука («Как стать большим?», «Союзмультфильм», 1967) и рассказывали происходящее одновременно с просмотром без предварительной подготовки. Таким образом, тексты, полученные при проведении эксперимента с детьми 1 группы, имеют форму диалога, тексты, полученные при проведении эксперимента с детьми 2 группы, также имеют форму диалога, но включают в себя монологические связные фрагменты, тексты, полученные при проведении эксперимента с детьми 3, 4 и 5 групп, имеют форму монологического связного текста. Общий объем корпуса — 25 689 словоформ (сюда входит только речевая продукция детей без высказываний экспериментатора). Каждый текст имеет следующие характеристики: пол и возраст автора (на момент записи), общее число словоформ и высказываний в данном тексте, средняя длина высказывания (MLU) (табл. 1).

Особой отличительной чертой корпуса является то, что все содержащиеся в нем тексты основаны на одинаковой (группы 3, 4 и 5) или схожей (группы 1 и 2) последовательности действий, что сюжетно унифицирует тексты и снимает многие дополнительные факторы, влияющие на семантическую и синтаксическую организацию связного текста. В этом корпус «Кондуит», с одной стороны, продолжает традиции изучения особенностей связного устного текста, заложенные в таких крупных исследованиях, как кросслингвистическая экспериментальная серия изучения детского нарратива под руководством Рут Берман и Дэна Слобина [Berman, Slobin 1994] или исследование устно-

Таблица 1. **Статистика корпуса по возрасту и полу, характеристика групп по числу словоформ и высказываний, по средней длине высказывания (MLU)**

Группа	Средний возраст	М	Ж	Число словоформ (Me)	Число высказываний (Me)	MLU (Me)
1	3 года 1 месяц	21	16	20	13	1,5
2	4 года 2 месяца	30	20	84.5	27	3,12
3	5 лет	25	24	108	23	4,61
4	6 лет	20	22	174,5	34	5,10
5	6 лет 9 месяцев	16	19	190	34	5,39

го спонтанного дискурса «Рассказы о сновидениях» [Кибрик, Подлесская 2009], а с другой, отличается от них, поскольку в последнем случае тексты были объединены одной темой, но были различны по сюжету и эмоциональной вовлеченности рассказчика, а при кросслингвистическом исследовании основной проблематикой является изучение именно сходства и различий в процессе усвоения нарративных принципов говорящими на разных языках, а не построение последовательности этапов усвоения нарративной структуры в онтогенезе.

Применяемая разметка обусловлена целью создания корпуса и включает три основных типа:

— структурная разметка: отражает эпизодическую структуру текстов и полноту описания события, персонажа и обстоятельств каждого из эпизодов. Поскольку тексты либо полностью, либо частично сюжетно идентичны, то такая разметка облегчает задачи исследования особенностей организации цельности текста в рассказах детей разного возраста, а также показывает важность отдельных действий разных типов, деталей и обстоятельств ситуации для детей на разном уровне когнитивного развития;

— синтаксическая разметка: отражает используемые детьми средства обеспечения связности текста на разных этапах усвоения навыков построения связного текста. Также включает в себя коммуникативную разметку, что позволяет выявить взаимосвязь между использованием в связном русскоязычном тексте различных синтаксических структур и организацией коммуникативной (тема-рематической) структуры высказывания;

— семантико-синтаксическая разметка глагольных структур: отражает взаимосвязь семантического класса и синтаксической структуры глагола, а также позволяет проводить сравнительное исследование усвоения семантико-синтаксической структуры глагола.

Взаимодействие всех трех типов разметки отражает роль, которую играют глаголы и их аргументы в организации структуры текста, в отражении событий и действий, в восприятии и понимании окружающей действительности детьми на различных этапах когнитивного и языкового развития, позволяет приблизиться к пониманию взаимодействия семантики и синтаксиса в процессах порождения и восприятия речи.

Данный корпус может использоваться для проведения сравнительных исследований процесса формирования навыков организации связности и цельности текста, формирования разнообразия синтаксических структур, развития лексического запаса, формирования таких когнитивных способностей, как внимание к деталям, эмпатия и эмоциональная оценка поведения другого, развитие фантазийного мышления и многого другого у русскоязычных детей в возрасте от 2,7 до 7,6 лет.

Литература

1. *Богданова-Бегларян Н. В.* (ред.) (2016), Русский язык повседневного общения: особенности функционирования в разных социальных группах / Н. В. Богданова-Бегларян, Т. Ю. Шерстинова, Е. М. Баева, О. В. Блинова, Г. Я. Мартыненко, О. Б. Ермолова, А. И. Рыко и др. СПб.
2. *Венцов А. В., Нигматулина Ю. О., Раева О. В., Риехакайнен Е. И., Слепокурова Н. А.* (2013), Корпус русских спонтанных текстов: структура и единицы. Труды международной конференции «Корпусная лингвистика–2013». СПб., с. 223–231.
3. *Гришина Е. А., Савчук С. О.* (2009), Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., с. 129–149.
4. *Кибрик А. А., Подлесская В. И.* (ред.) (2009), Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.
5. *Berman R., Slobin D.* (eds.) (1994), Relating events in narrative: A crosslinguistic developmental Study. New York.

References

1. *Bogdanova-Beglaryan N. V.* (ed.) (2016), Russkij yazyk povsednevnogo obshheniya: osobennosti funkcionirovaniya v raznykh social'nykh gruppakh [Russian everyday communication: its functioning within different social groups]. Saint Petersburg.
2. *Vencov A. V., Nigmatulina Yu. O., Raeva O. V., Riekhakajnen E. I., Slepokurova N. A.* (2013), Korpus russkikh spontannykh tekstov: struktura i edinicy [Corpus of Russian spontaneous texts: its structure and elements]. In: Trudy mezhdunarodnoj konferencii "Korpusnaya lingvistika — 2013" [Proceedings of international conference "Corpus linguistics-2013"]. Saint Petersburg, pp. 223–231.
3. *Grishina E. A., Savchuk S. O.* (2009), Korpus ustnykh tekstov v NKRYA: sostav i struktura [Corpus of oral texts in National Russian Corpus: elements and structure]. In: Nacional'nyj korpus russkogo yazyka: 2006–2008. Novye resultaty i perspectivy [National Russian Corpus: 2006–2008. New results and perspectives]. Saint Petersburg, pp. 129–149.
4. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009), Rasskazy o snovideniyakh: Korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow.
5. *Berman R., Slobin D.* (eds.) (1994), Relating events in narrative: A crosslinguistic developmental Study. New York.

Эйсмонт Полина Михайловна
Санкт-Петербургский государственный университет
аэрокосмического приборостроения
Eismont Polina
Saint Petersburg State University of Aerospace Instrumentation
E-mail: polina272@hotmail.com

И Ян
Yi Yang

НОВЫЙ КИТАЙСКО-РУССКИЙ ПАРАЛЛЕЛЬНЫЙ КОРПУС С ДИСКУРСИВНО-СТРУКТУРНОЙ РАЗМЕТКОЙ[1]

A NEW CHINESE-RUSSIAN DISCOURSE STRUCTURE PARALLEL CORPUS

Аннотация. Дискурсивно-структурная разметка заключается в описании структуры текста в виде сети дискурсивных единиц, соединенных дискурсивными отношениями. Выравнивание в корпусе проводится по элементарной дискурсивной единице и структуре каждого абзаца. Создаваемый корпус может быть в дальнейшем использован для решения задач машинного перевода, обучения иностранным языкам, сопоставительной лингвистики, теории перевода и т.д.
Ключевые слова. Китайско-русский параллельный корпус, выравнивание по клаузам, дискурсивно-структурная разметка, машинный перевод.

Abstract. This article is devoted to building a Chinese-Russian Parallel Corpus, in which comparable texts are annotated and aligned at the level of discourse structure. We show how to align and annotate clauses, discourse connectives and relations for Chinese-Russian comparable texts. In addition, the subject of this article is of great interest to comparative linguistics, practice of discourse analysis, teaching translation and machine translation.
Keywords. Chinese-Russian parallel corpus, clause-alignment, discourse structure annotation, machine translation.

1. Введение

Дискурсивно-структурная разметка определяется как лингвистическая информация о структуре дискурса, в том числе об идентификации элементарных дискурсивных единиц, дискурсивных отношений, организации текста и т.д. Создание корпуса с дискурсивно-структурной разметкой основывается на опыте создания англоязычных дискурсивных трибанков RST Discourse TreeBank [Carlson et al. 2003] и Penn Discourse TreeBank (PDTB) [Miltsakaki et al. 2004]. В нашем корпусе выравнивание параллельных текстов китайского языка и его перевода проводится не только по элементарной дискурсивной единице (ЭДЕ), но и по дискурсивной структуре каждого абзаца. Соответственно, создание корпуса производится в два этапа: 1) членение текстов и выравнивание на уровне ЭДЕ (см. п. 2); 2) выравнивание и разметка текстов на уровне дискурсивной структуры (см. п. 3).

[1] Исследование выполнено при поддержке Программы повышения конкурентоспособности Уральского федерального университета (номер соглашения 02.A03.21.0006) и «China Scholarship Council».

2. Выравнивание на уровне элементарной дискурсивной единице

Выравнивание в параллельном корпусе устанавливает соответствие между парами текстов. Суть выравнивания, по мнению В. П. Захарова, заключается в «параллельной сегментации оригинального текста и его перевода по предложениям, клаузам, словосочетаниям и словам» [Захаров 2011: 26]. Итак, параллельная сегментация является одной из ключевых процедур создания параллельного корпуса.

Сегментация текста связана с выделением ЭДЕ. Во многих современных теориях, ЭДЕ обычно равна клаузе. Идеальное соотнесение клауз двух текстов в принципе невозможно: клауза исходного языка (ИЯ) может быть передана несколькими синтаксическими структурами переводного языка (ПЯ). Если наша задача заключается в обработке параллельных текстов, то их членение должно быть взаимообусловлено. Поэтому в нашем корпусе предполагается выравнивание по клаузам ИЯ и их синтаксическим аналогам ПЯ.

В нашем корпусе при делении текста ИЯ клауза выделяется исходя из трех параметров: 1) клауза обязательно состоит из одной глагольной группы и одной или нескольких именных групп; 2) в клаузе содержится как минимум одно суждение; 3) между клаузами обычно ставится знак препинания (запятая, точка с запятой и т. д.), но не любое предложение со знаком препинания членится на ЭДЕ [Мухин, Ян 2016: 25]. Текст ПЯ членится на соответствующие синтаксические аналоги клауз ИЯ. Итак, мы переходим к следующему этапу создания параллельного корпуса.

3. Дискурсивно-структурная разметка в китайско-русском параллельном корпусе

Выравнивание по клаузам дает возможность дальнейшего выравнивания по дискурсивной структуре. Для того, чтобы разметка в параллельном корпусе в большей степени отражала процесс перевода и понимание переводчика, общая дискурсивная структура определяется по переводному тексту. Выравнивание проводится одновременно в процессе разметки.

3.1. Дискурсивные связки и отношения

Структурная разметка связана с определением дискурсивных связок и дискурсивных отношений. Дискурсивные связки определяются как средства для организации дискурса, соединяющие разные типы

конструкций (типы ЭДЕ) и указывают на дискурсивное отношение между ними. Функцию дискурсивных связок и китайского языка, и русского, могут выполнять не только традиционные союзы и союзные слова, но и предлоги, наречия, вводные конструкции. Классификация дискурсивных связок в дальнейшем будет уточнена в ходе работы над разметкой корпуса.

За основу классификации дискурсивных отношений мы решили принять концепцию Li, согласно которой существует 4 группы отношений, в том числе 17 разновидностей [см. Li 2014: 2111]. На практике создания китайско-русского параллельного корпуса различные виды отношений приходится адаптировать к материалу и понимать расширительно.

3.2. Структурный анализ дискурса

Приведем для пояснения выровненные тексты по клаузам, пример (1).

Пример (1).

Исходный текст (а)	***Переводной текст (б)***	
[a1]展望今后五年，	[61]Перспективы развития на предстоящее пятилетие	‖
[a2]我们充满必胜信心。	[62]придают нам уверенность в победе.	|
[a3]如期实现全面建成小康社会目标，	[63]Намеченные задачи полного построения среднезажиточного общества будут выполнены в установленные сроки,	‖
[a4]人民生活将会更加美好，	[64]жизнь народа непременно станет еще лучше,	‖
[a5]中国特色社会主义事业前景一定会更加光明！	[65]а дело социализма с китайской спецификой ждет еще более светлое будущее!	

«Доклад о работе правительства КНР 2016г.»

Как показано в этом примере, в каждой строке выделены ЭДЕ. Количество вертикальных черт (знак «|») между ЭДЕ указывает на уровень иерархии в структурном дереве, к которому ЭДЕ относится. Подчеркнутые слова являются дискурсивными связками. В сущности, структурный анализ дискурса показывает степень близости соседних дискурсивных единиц в семантическом и грамматическом

отношении: а1/б1 и а2/б2 имеют более близкое отношение, чем а2/б2 и а3/б3; между а3/б3, а4/б4 и а5/б5 — тождественное отношение в тексте.

4. Критерии отбора текстов для корпуса

Тексты в китайско-русском параллельном корпусе с дискурсивно-структурной разметкой на данном этапе относятся к официально-деловому стилю и содержат много повторяющихся элементов (от слов до текстовых структур), что имеет большое значение для возможной автоматической обработки. Чтобы не сомневаться в качестве перевода, мы используем основной источник текстового материала — официальный сайт правительства КНР (http://cn.theorychina.org/). На данный момент в процессе разработки находятся восемь законов и десять докладов правительства. Объем корпуса на сегодняшний день составляет 1 822 абзац текста, 196 567 текстоформ, в том числе 32 111 текстоформ в русской части и 130 285 — в китайской.

5. Заключение

На данном этапе создания корпуса определены общие принципы выравнивания и сегментации параллельных текстов, а также принципы разметки дискурсивной структуры. В будущем необходимо решить проблемы соотношения классификаций дискурсивных связок и типов дискурсивных отношений. Мы предполагаем, что корпус может быть использован для машинного перевода, обучения китайскому и русскому языку, сопоставительного анализа синтаксических и семантических структур, а также для решения других актуальных задач.

Литература

1. *Захаров В. П., Богданова С. Ю.* (2011), Корпусная лингвистика: учебник для студентов гуманитарных вузов. 161 с.
2. *Мухин М. Ю., Ян И* (2016), Проект создания китайско-русского параллельного корпуса официально-деловых текстов с дискурсивно-структурной разметкой // Вестник ЮУрГУ, Лингвистика, 13(4), с. 23–31.
3. *Carlson L., Marcu D., Okurowski M. E.* (2003), Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory, Current Directions in Discourse and Dialogue. Available at: http://www.aclweb.org/anthology/W01-1605
4. *Li Y., Feng W., Sun J., Kong F., Zhou G.* (2014), Building a Chinese Discourse Corpus with Connective-driven Dependency Tree Structure, Emporical Methods in Natural Language Processing, pp. 2105–2114.

5. *Miltsakaki E., Prasad R., Joshi A., Webber B.* (2004), The Penn Discourse Treebank, the 4th International Conference on Language Resources and Evaluation. Available at: http://www.cis.upenn.edu/~elenimi/lrec04-lisbon-miltsakaki.pdf.

References

1. *Zaharov V. P., Bogdanova S. Ju.* (2011), Korpusnaja lingvistika: uchebnik dlja studentov gumanitarnyh vuzov. 161 p.
2. *Muhin M. Ju., Jan I* (2016), Proekt sozdanija kitajsko-russkogo parallel'nogo korpusa oficial'no-delovyh tekstov s diskursivno-strukturnoj razmetkoj. In: Vestnik JuUrGU, Lingvistika, 13(4), pp. 23–31.
3. *Carlson L., Marcu D., Okurowski M. E.* (2003), Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory, Current Directions in Discourse and Dialogue. Available at: http://www.aclweb.org/anthology/W01-1605
4. *Li Y., Feng W., Sun J., Kong F., Zhou G.* (2014), Building a Chinese Discourse Corpus with Connective-driven Dependency Tree Structure, Emporical Methods in Natural Language Processing, pp. 2105–2114.
5. *Miltsakaki E., Prasad R., Joshi A., Webber B.* (2004), The Penn Discourse Treebank, the 4th International Conference on Language Resources and Evaluation. Available at: http://www.cis.upenn.edu/~elenimi/lrec04-lisbon-miltsakaki.pdf.

И Ян
Уральский федеральный университет (Россия)
Yi Yang
Ural Federal University (Russia)
E-mail: xwyang@mail.ru

Научное издание

ТРУДЫ МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА–2017»
27–30 июня 2017 г., Санкт-Петербург

Компьютерная верстка *Ю. Ю. Тауриной*

Подписано в печать 21.06.2017. Формат 60×84 $^{1}/_{16}$.
Усл. печ. л. 22,32. Тираж 120 экз. Заказ №

Типография Издательства СПбГУ.
199034, Санкт-Петербург, Менделеевская линия, д. 5.