Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2017"

Moscow, May 31—June 3, 2017

RUSSIAN COLLOCATION EXTRACTION BASED ON WORD EMBEDDINGS[1]

Enikeeva E. V. (protoev@yandex.ru),
Mitrofanova O. A. (o.mitrofanova@spbu.ru)

Saint Petersburg State University, St. Petersburg, Russia

Collocation acquisition is a crucial task in language learning as well as in natural language processing. Semantics-oriented computational approaches to collocations are quite rare, especially on Russian language data, and require an underlying semantic formalism. In this paper we exploit a definition of collocation by I. A. Mel'čuk and colleagues (Iordanskaya, Mel'čuk 2007) and apply the theory of lexical functions to the task of collocation extraction. Distributed word vector models serve as a state-of-the-art computational basis for the tested method. For the first time experiments of such type are conducted on available Russian language data, including Russian National Corpus, SynTagRus and RusVectōrēs project resources. The resulting collocation lists are assessed manually and then evaluated by means of precision and MRR metrics. Final scores are quite promising (reaching 0.9 in precision) and described algorithm improvements yield a considerable performance growth.

Keywords: distributional semantics, compositional collocations, "Meaning ⇔ Text" theory, collocation extraction

[1] The reported study is supported by RFBR grant № 16-06-00529 "Development of a linguistic toolkit for semantic analysis of Russian text corpora by statistical techniques".

ИСПОЛЬЗОВАНИЕ ВЕКТОРНЫХ МОДЕЛЕЙ ДЛЯ ИЗВЛЕЧЕНИЯ КОЛЛОКАЦИЙ ИЗ КОРПУСОВ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

Еникеева Е. В. (protoev@yandex.ru),
Митрофанова О. А. (o.mitrofanova@spbu.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

1. Introduction

Collocability is an important factor in a vast majority of natural language processing and language modelling tasks, namely, syntactic parsing, machine translation, paraphrase generation, automatic and semi-automatic dictionary acquisition, semantic role labelling, word sense disambiguation, etc. In fact, contemporary research in most fields of computational linguistics rests upon the achievements of "contextualist" framework, cf. (Khokhlova 2010, etc.).

Presumably the first definition of collocation can be found in (Palmer 1933): "A collocation is a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts". These linguistic units are usually treated as restricted co-occurrences of two (or more) syntactically bound elements (Kilgariff 2006). At the same time they should be distinguished from idioms, because target word or collocation base can co-occur with a number of other lexical units (e.g. collocations *еловая*, *сосновая*, *кедровая*, etc. *шишка* '*fir, pine, cedar*, etc. *cone*' vs. idiom *важная шишка* '*boss*'; collocations *бить тревогу*, *рекорд*, *поклоны*, etc. '*sound the alarm*, *beat the record*, *beat bows*, etc.' vs. *idioms бить баклуши* '*twiddle*').

In our study much attention is given to the treatment of collocations in "Meaning ⇔ Text" theory (Iordanskaya, Mel'čuk 2007, Mel'čuk 1998) reflected in Explanatory Combinatorial Dictionary of Contemporary English (Mel'čuk, Zholkovsky 1984) and in SynTagRus Treebank (http://www.ruscorpora.ru/search-syntax.html). The theory allows to describe collocation structure in terms of lexical functions (LFs) that associate one lexical unit (argument, base) with another (value, collocate) which is selected by the rules of a language to express a meaning of given LF (cf. Section 3.1. below). Therefore, it is obvious that collocations are language-specific, for example, a meaning of '*do*, *perform*' for a base '*lecture*' is in English expressed by a lexeme '(*to*) *hold*' while in Russian the same meaning is conveyed by '*читать*' ('*chitat*", '(*to*) *read*'). Significance of "Meaning ⇔ Text" approach rests on the idea that collocations are expected to reveal both syntagmatic unity and lexical correlation of its parts.

In recent years we have witnessed rapid expansion of various collocation extraction techniques, which are based on co-occurrence statistics. Automatic tools for collocation extraction usually produce a list of ranked bigrams or n-grams. The ranking

(reflecting the so called 'collocation strength') is obtained in most cases by means of a statistical association measure such as t-score or PMI. Morphosyntactic annotation of processed corpora allows to bring into action such linguistic data as lexical-syntactic patterns and/or valency frames defining boundaries of syntactic groups for collocations and possibly their inner argument structure (e.g., Word Sketch Engine, https://www.sketchengine.co.uk/; RNC Sketches, http://ling.go.mail.ru/synt/). However, even the most sophisticated techniques of collocation extraction fail to take into account lexical and semantic peculiarities of collocations.

The purpose of our study is to prove the possibility of LF-oriented automatic collocation extraction for Russian. Our aim is to extract sets of collocations for target LFs from large corpora by means of machine learning. In this paper we try to combine a formal theory of collocations in "Meaning ⇔ Text" theory with distributed word representations (Mikolov et al. 2013a). Distributed word vectors can be used to extract linguistic regularities from a large corpus in an automated way. We are the first to perform experiments for Russian corpora in the given settings. The expected output is fine-grained classification of collocations according to their lexical meaning and syntactic structure which is important both in language learning and NLP applications (Apresjan et al. 2002; Kolesnikova, Gelbukh 2012, etc.).

The paper is structured as follows: first of all, an outline of the research in the field is presented. Then, we briefly describe a theoretical background of our study and present a computational model. In the following sections we present experimental settings and evaluation framework, and conclude with the results and its discussion.

2. Related work

Although publications on statistical collocation extraction seem to be overwhelming, there is much to be done in this field. In this section we mention several remarkable studies on Russian data. As observed in the introduction, most of approaches implemented and applied to Russian text corpora do not take into account the semantic structure of collocations, describing more or less free word combinations alongside with idioms, for example, in (Yagunova, Pivovarova 2010) collocations *сердечный приступ* '*heart attack*' and *круглый стол* '*round table*' take neighboring positions in the list.

The most popular statistical measures used to compute word association within collocations include Mutual Information (MI), Dice coefficient, Log-Likelihood and t-score (Khokhlova 2010). Multiword expressions scoring by means of learning-to-rank methods involved in information retrieval is discussed in (Tutubalina, Braslavski 2016). The approach is based on machine learning techniques, it makes use of bigrams from dictionaries as training data and, as authors say, treats collocations, idioms, set phrases in a uniform way. In (Kormacheva et al. 2014) six metrics (frequency, refined frequency ratio, weighted frequency ratio, MI, Dice score and t-score) were tested on Russian prepositions and the best performance was shown by refined frequency ratio score. Previously mentioned (Yagunova, Pivovarova 2010) compared MI and t-score and prove that the former is more suitable for extracting collocations reflecting domain-specific terms. The latter (t-score) gives preference to phrases that may be called auxiliary (two-word parentheses, discourse phrases).

The recent results of Collocations, Colligations and Corpora Project (CoCoCo, (Kopotev et al. 2015)) are presented in (Kormacheva et al. 2016). Collocations and colligations are classified according to the association between phrase constituents: some of them are marked as idioms and others are subject to semantic generalization: e.g., *sleight of* [*hand/mouth/mind*]. The procedure is fully automated and based on multiple grammatical and lexical features.

A study concerning association strength measurement in syntactic constructions and testing methodology is described in (Bukia et al. 2015). The authors study adjective-noun collocations, and their algorithm predicts association even for the combinations absent from corpus. Verb-noun collocation extraction from Russian texts is studied in (Akinina et al. 2013). The approach is PMI-based and takes into account syntactic information without further semantic classification.

Association strength measurement is closely related to identification of abnormal lexical compositions (Vecchi et al. 2011) and automatic lexical error detection (Kochmar, Briscoe 2013). The latter work presents a number of semantic anomaly measures in a vector space. A distributional approach applied to Russian error correction in collocations can be found in (Panicheva, Mitrofanova 2016).

Compositional distributional semantics provides successful solution of the collocation extraction task. As far as evidence for Russian is concerned, several vector models were evaluated during RUSSE workshop (Panchenko et al. 2015). Nowadays attention of researchers working with distributed word vector representations for Russian is focused on RusVectōrēs (Kutuzov, Kuzmenko 2017), AdaGram (Bartunov et al. 2015) and RDT (Panchenko et al. 2016) models. However, semantic relatedness evaluation only involves paradigmatic relations between lexical units. In (Bukia et al. 2016) two distributional approaches to selectional preference modelling are compared. The first one implies semantic similarity calculation based on cosine distance; the second one relies on Mikolov's (Mikolov 2013b) assumption about linguistic regularities captured by distributed word vector models. The metric similar to the latter is used as a baseline in (Rodríguez-Fernández et al. 2016): collocates are evaluated against a difference between example headword and collocate added to test headword. The main method proposed in the same paper is based on linear transformation between headword and collocate space. The approach is tested on manually classified samples drawn from Macmillan Collocations Dictionary (Rundell 2010).

Our study follows the experience of our colleagues and takes into account peculiarities of Russian data and resources.

3. Theoretical model

3.1. Collocations in "Meaning ⇔ Text" theory

"Meaning ⇔ Text" theory provides an exhaustive analysis of phraseological expressions, taking into account various types of interaction between lexical and semantic components constituting the meaning of a word group as well as well as syntactic relations established between co-occurring words. Collocations are considered

as a subclass of non-free utterances (or phrasemes). A formal definition of collocations (or semi-phrasemes) runs as follows: "A collocation **AB** of language **L** is a semantic phraseme of **L** such that its signified 'X' is constructed out of the signified of the one of its two constituent lexemes—say, of **A**—and a signified '*C*' ['*X*' = '*A* ⊕ *C*'] such that the lexeme **B** expresses 'C' contingent on **A**" [Mel'čuk 1998: 30]. A collocation includes a base constituting a freely chosen semantic nucleus of a word group and a collocate being a restricted component which determines the meaning of the whole as a function of the base. Opposite to idioms which reveal non-compositional nature, collocations are treated as compositional phrasemes conforming to lexical constraints imposed on collocates (*сильный акцент* '*heavy accent*', *високосный год* '*leap year*', *спать глубоким сном* '*be soundly asleep*' и т. д.)

In case of restricted lexical co-occurrence the relations between a base and a collocate reproduced in semantically and syntactically similar expressions are represented by lexical functions (LFs) which are formally defined as follows: $f(A) = B$, for example, MAGN(*болезнь* '*disease*') = *тяжелая* ('*serious*').

Some of the most frequent syntagmatic LFs are:

- MAGN means 'very', 'to a (very) high degree', 'intense(ly)': MAGN(смеяться 'laugh') = от души 'heartily';
- OPER1 introduces a support verb meaning 'do', 'perform': OPER1(поддержка 'support') = оказывать '(to) lend';
- FUNC0 means that an event described by a headword takes place: FUNC0(снег 'snow') = идёт 'falls', etc.

In fact, LFs describe not only syntagmatic relations but also paradigmatic (SYN(*врач* '*doctor, physician*') = *доктор* ('*doctor, physician*'), ANTI(быстрый 'fast') = медленный 'slow', CONV(покупать 'buy') = *продавать* '*sell*', etc.), and derivational ones (S0(*гордый* '*proud*') = *гордость* ('*pride*'), A1(*голод* '*hunger*') = *голодный* '*hungry*', CAUS(*понимать* '*understand*') = *объяснять* '*explain*', etc.).

LFs of different types can be combined in complex functions to express one meaning:

> INCEPOPER1 = INCEP (= '*to start*') × OPER1: INCEPOPER1(*привычка* '*habit*') = *приобретать* '*acquire*'.

At present the inventory of LFs comprises 116 varieties of standard and non-standard LFs (Apresjan et al. 2007). Russian collocations revealing LF relations are thoroughly described in the Explanatory Combinatorial Dictionary of Modern Russian (Zholkovsky, Melchuk 1984) and annotated in Russian National Corpus (RNC) subcorpus SynTagRus (Frolova, Podlesskaya 2011).

3.2. Predicting LF values by means of vector model

Mikolov and colleagues (Mikolov et al. 2013b) prove that regular linguistic relations between two word spaces may be described as a linear transformation on them. In case of collocations the relation to be modelled is perfectly formalized as a lexical function in "Meaning ⇔ Text" theory.

Our task is to predict values of a particular LF for an argument in question given training instances of this LF. Following (Rodríguez-Fernández et al. 2016),

we define an argument space *A* and collocate space *C* produced by word2vec toolkit. Let *T* be a set of collocations t_i comprising argument-value pairs (a_{t_i}, c_{t_i}), that represent a given lexical function *L*. Argument matrix $A_T = [a_{t_1}, \dots, a_{t_n}]$ and collocate matrix $C_T = [c_{t_1}, \dots, c_{t_n}]$ are made up of corresponding word vectors. Then, given examples of a particular LF (e.g., MAGN: *тяжёлая болезнь* '*hard illness*', *сильный акцент* '*heavy accent*', etc.), we should find a transformation which converts an argument vector to a value vector of this LF, for instance, predicts a collocate *бурный* '*wild*' (MAGN value) for an argument *аплодисменты* '*applause*'.

A linear transformation matrix $\Psi \in \mathbb{R}^{B \times C}$ learnt from training set *T* satisfies the following: $A_T \Psi_T = C_T$.

Therefore, Ψ can be approximated using singular value decomposition to minimize the sum:

$$\sum_{i=1}^{|T|} \left\| \Psi_T a_{t_i} - c_{t_i} \right\|^2.$$

Thus, we obtain a transformation matrix for a given LF. Applying it (multiplying it by argument vector representation) we obtain a ranked list of potential collocates for a given headword and lexical function. Following (Rodríguez-Fernández et al. 2016) we then use part-of-speech collocation patterns and NPMI filters. NPMI stands for normalized pointwise mutual information and is calculated as follows:

$$NPMI = \frac{PMI(a, c)}{-\log p(c)}.$$

4. Experiments

4.1. Test data

The resources containing LF markup for Russian language are quite limited. In our experiments we use SynTagRus Treebank (http://www.ruscorpora.ru/instruction-syntax.html)[2] and Verbal collocations of Russian abstract nouns dictionary (http://dict.ruslang.ru/abstr_noun.php). SynTagRus is a subset of Russian National Corpus (http://www.ruscorpora.ru) where each sentence is assigned a parse tree as well as a list of LFs in "Meaning ⇔ Text" notation. Verbal collocations dictionary uses its own markup scheme based on LF inventory. Collocations are classified in terms of 'regular abstract meanings', such as necessity, existence, action, with additional labels such as phase (start, finish) or semantic class (cognition, perception etc.)

The authors of (Rodríguez-Fernández et al. 2016) have proved their assumption that headword and collocate embeddings should be trained on different corpora. In their work headword vectors are obtained from a small corpus containing primarily literal usage (Wikipedia), while collocate vectors are trained on a large corpus full of various figurative meanings. Our experiments are aimed at testing this hypothesis once again on available Russian language data. Thus, we use precomputed word embeddings by RusVectōrēs project (http://ling.go.mail.ru/ru/, version 3) trained

[2] We are deeply grateful to SynTagRus team, especially to Leonid L. Iomdin and colleagues form IPPI RAS, for providing access to the data on LF.

on Russian National Corpus and Web corpus. NPMI scores are precomputed on Russian fiction corpus (collected from M. Moshkov's library, URL: http://lib.ru).

Table 1. Lexical functions and its frequency

LF	argument	value	Syntagrus frequency	gloss in (Rodríguez-Fernández et al. 2016)	rank in (Rodríguez-Fernández et al. 2016)
OPER1	цель 'aim'	иметь 'have'	818	'perform'	2
MAGN	каблук 'heel'	высокий 'high'	799	'intense'	1
CAUSFUNC0	соревнование 'competition'	проводить 'hold'	256	—	—
FUNC0	открытие 'opening'	состояться 'be held'	226	—	—
INCEPOPER1	работа 'work'	приступать 'start'	210	'begin to perform'	4
OPER2	правка 'correction'	подвергаться 'undergo'	140	—	—
REAL1-M	ракета 'rocket'	запускать 'launch'	109	—	—
REAL1	средства 'means'	расходовать 'spend'	97	—	—
INCEPFUNC0	речь 'conversation'	заходить 'turn to'	94	—	—

4.2. Test setup and evaluation

First of all, we conducted experiments on 9 most frequent LFs[3] from SynTagRus (table 1). In the table we present also semantic glosses from (Rodríguez-Fernández et al. 2016) and its frequency ranks for comparison. It is quite surprising that LFs' frequency distribution in Russian corpus differs from Macmillan Collocations Dictionary. An initial collocation set was extracted from the treebank and 10 headwords of these collocations were randomly chosen as a test set. The remaining part comprises a training set. For each LF top-10 ranked collocates were assessed manually. The performance was then evaluated using precision and mean reciprocal rank (MRR) on this list of 10 collocates:

$$precision = \frac{tp}{tp + fp}.$$

[3] LOC lexical function was excluded from top-10 as its value is expressed by preposition

where *tp* is a number of correct collocates among the retrieved list, *fp* is a number of false collocates in the list;

$$precision' = \frac{ep}{e},$$

where *ep* is a number of expected collocates (found in SynTagRus) among the top-10 retrieved ones and *e* is a number of collocates found in SynTagRus;

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{ran},$$

where *Q* is the top-10 list and $rank_i$ is a rank of the first correct collocate (according to experts' annotation).

Filtering was conducted using Universal Dependencies part-of-speech tags assigned to lexemes in RusVectōrēs models. As seen from table 1, all test lexical functions have verbal values, so POS tags filtering in our case means simply eliminating collocates with other POS tags. NPMI threshold was chosen experimentally on some headwords different from testset.

Following (Rodríguez-Fernández et al. 2016) we present several models:

- M1—a baseline vector model from (Bukia et al. 2016) which is virtually the same as a baseline in (Rodríguez-Fernández et al. 2016). For each candidate collocate we compute its cosine similarity to $vec(a_i) - vec(c_i) + vec(a_j)$, where (a_j, c_i) is an example collocation for a given LF and a_j is a test headword;
- M2—the same baseline filtered by POS tags and NPMI scores;
- M3—the model described above using the same vector spaces for headwords and collocates trained on RNC;
- M4—model M3 filtered by POS tags and NPMI scores;
- M5—model M3, but collocate vectors are obtained from Russian Wikipedia corpus;
- M6—model M5 filtered by POS tags and NPMI scores.

Table 2. Precision scores

LF	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*
OPER1	0.11	0.31	0.10	0.14	0.37	**0.63**
MAGN	0.23	0.28	0.24	0.28	0.63	**0.84**
CAUSFUNC0	0.10	0.33	0.22	0.23	0.54	**0.64**
FUNC0	0.21	0.40	0.29	0.33	**0.42**	**0.42**
INCEPOPER1	0.10	0.38	**0.64**	**0.64**	0.15	0.15
OPER2	0.17	0.28	0.12	0.11	0.29	**0.39**
REAL1-M	0.20	**0.66**	0.24	0.26	0.40	0.52
REAL1	0.15	0.37	0.32	0.33	**0.66**	**0.66**
INCEPFUNC0	0.13	0.28	0.24	0.23	0.35	**0.43**

Table 3. Expected precision scores

LF	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*
OPER1	0.50	0.50	0.48	**0.65**	0.57	**0.65**
MAGN	0.68	0.70	0.70	0.70	**0.83**	0.78
CAUSFUNC0	0.33	0.45	0.87	**0.9**	0.80	0.80
FUNC0	0.70	0.80	**0.90**	0.81	0.58	0.58
INCEPOPER1	0.40	**0.55**	0.50	0.50	0.50	0.50
OPER2	0.55	0.60	0.53	0.52	**0.75**	0.70
REAL1-M	0.55	0.70	**0.87**	**0.87**	0.70	0.75
REAL1	0.40	0.45	**0.73**	0.70	0.60	0.60
INCEPFUNC0	0.50	0.70	0.72	0.67	**0.82**	0.77

Table 4. MRR scores

LF	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*
OPER1	0.22	0.55	0.11	0.48	0.34	**0.73**
MAGN	0.30	0.68	0.30	0.68	0.50	**0.90**
CAUSFUNC0	0.11	0.43	0.37	0.76	0.41	**0.82**
FUNC0	0.30	0.82	0.47	**0.89**	0.36	0.64
INCEPOPER1	0.11	0.60	0.01	**0.64**	0.15	0.48
OPER2	0.19	0.64	0.23	0.48	0.37	**0.66**
REAL1-M	0.08	0.77	0.36	0.70	0.34	**0.86**
REAL1	0.23	0.50	0.37	**0.70**	0.30	0.59
INCEPFUNC0	0.10	0.64	0.37	**0.73**	0.42	0.72

4.3. Discussion

The results are presented in tables 2–4. As expected, the presented models outperform the baseline except for several cases. In general, a considerable improvement in precision and MRR scores is achieved by filtering. On the other hand, as far as precision on expected collocations is concerned, NPMI filters discard some relevant examples, so that the scores without filtering are higher. As regards ranking (assessed by MRR metric), we do not observe a steady improvement when using a different corpus to model collocate vector space on several test LFs: FUNC0, INCEPOPER1, REAL1, INCEPFUNC0. We suppose, that collocates corresponding to these LFs' values are quite rare in general domain corpus. On the contrary, as these meanings are quite specific (not abstract), they are better represented in standard register corpus.

It should be mentioned, that there is a number of headwords where an expected LF value is absent from a retrieved top-10 list. However, in the majority of such lists there is at least one correct collocate.

The examples of ranked collocates are presented in table 5. Correct collocates corresponding to target LF values are underlined. Correct collocates coinciding with the expected LF values are given in bold type. More frequent LF collocates (MAGN,

FUNC0) generally seem to be retrieved with higher precision because of wider collocability of its arguments and their high frequency. On the other hand, more specific LFs (REAL1, INCEPOPER1) are also processed correctly because such combinations are quite specific and usually both headword and collocate occur in quite specific contexts.

Table 5. Retrieved collocates examples. Correct LF values are underlined

headword	LF (Mel'čuk, Zholkovsky 1984)	retrieved collocates
довод	MAGN(*довод*)= *убедительный*	*решительный*, ***убедительный***, *основательный*, *веский*, *главный*, *бесспорный*, *достаточный*…
домино	OPER1(*домино*) = *играть*	***играть***, *поиграть*, *стучать*, *резаться*, *сыграть*, *игра*, *бильярд*, *футбол*…
арест	OPER2(*арест*) = *сидеть*	*подвергаться*, *находиться*, *подвергать*, *брать*, *миновать*, *попадать*…
азарт	INCEPOPER1(*азарт*) = *входить*	*приходить*, *игра*, *увлекаться*…
дорога	FUNC0(*дорога*) = *проходить*	*идти*, *пойти*, *тянуться*, *лежать*, *плестись*, *тащиться* …
день	INCEPFUNC0(*день*) = *наставать*	*наступать*, ***наставать***, *начинаться*, *приходить*, *прийти*, *намечаться*, *длиться*, *заканчиваться*…
встреча	CAUSFUNC0(*встреча*) = *назначать*, *проводить*, *устраивать*	***назначать***, *намечать*, ***проводить***, *уславливаться*, *потребовать*, *приглашать*…
газета	REAL1(*газета*) = *читать*	***читать***, *прочитывать*, *перечитывать*, *почитывать*, *читывать*, *листать*, *писать*, *пересказывать*…
долг	REAL1-M(*долг*) = *выполнять*, *исполнять*, *отдавать*	***исполнять***, *погашать*, ***исполнить***, *уплачивать*, *погасить*, ***отдавать***, *заплатить*, *повиноваться*, *обязывать*…

Alongside with correct LF values the output also includes several erroneous cases. First of all, some of the potential candidates do not represent a typical value of a given LF, for example, *главный довод* '*main reason*' may be treated as a realization of MAGN, though the sense of 'intense' is not the main one in the given phrase. Secondly, virtually all of the retrieved words may be interpreted as values of a lexical function (not necessarily the target LF). Consider the case of the headword '*ошибка*'. The lexemes *неточность*, *просчёт*, *промах*, *погрешность*, *дефект*, *описка* represent synonyms (SYN(*ошибка*)); *допускать* is a possible value of OPER1(*ошибка*). Other cases are *квартира* = COHYP(*дом*) given REAL1(*дом*) = *жить*; *долгий* = MAGN(*разбор*); *эскадра* = MULT(*корабль*), etc.

Thus, negative examples, although they go beyond the scope of our study, yield consistent explanation in terms of LF theory. Our data provide evidence on the possibility of retrieving "bundles" of lexical functions for a given word, e.g.

FUNC0(*речь*) = *идти*: OPER1(*речь*) = *произносить*, HYPO(*речь*) = *тирада*, *скороговорка*, S1(*речь*) = *оратор*, S-LOC(*речь*) = *митинг*, *банкет*, VER(*речь*) = *застольная*, etc. Thorough description of LF "bundles" obtained for a given headword allows to bridge a gap from the pure collocation analysis to the complex study of lexical-syntactic constructions.

5. Conclusion

Our study shows that enrichment of traditional statistical techniques of collocation extraction by means of vector space models and lexical-syntactic information (in our case, LF data) gives new insights into the problem of how word meanings interact in contexts. In most cases contemporary corpus-based data available for Russian ignore lexical structure of collocations and provide statistical information based on association measures and/or morphosyntactic patterns. On the one hand, by now profound semantic analysis has been performed for more complex linguistic units registered in Russian corpora, namely, constructions (cf. Lexicograph (URL: http://lexicograph.ruslang.ru/) and FrameBank (URL: http://framebank.ru/) projects). On the other hand, description of fine-grained lexical-semantic relations of LF type has been carried out within the lexicographic framework, given a limited list of headwords and narrow set of their collocations which maintain certain LFs (Mel'čuk, Zholkovsky 1984).

Our research is the first to provide reliable evidence on the possibilities of automated retrieval and classification of collocations exhibiting LFs in large corpora of Russian, thus bridging the gap between traditional dictionaries and corpus-based semantic representations. We have successfully applied a state-of-the-art approach (distributed word vector representations) to extracting potential collocates for given headwords and target lexical functions. The method requires only a dictionary of tagged collocations (the SynTagRus corpus with LF markup, in our case) and corpora for distributed representation learning. Since these corpora are quite large, the number of retrieved collocates exceeds the number of collocates listed in the dictionary.

The approach discussed and tested in our paper promises a vast field for future research.

We are going to improve experimental settings. In reported research we used word embeddings data from word2vec models in RusVectōrēs project, now we've got the possibility to use AdaGram (URL: https://github.com/sbos/AdaGram.jl) and/or RDT (URL: https://nlpub.ru/Russian_Distributional_Thesaurus) models. As experiments were carried out for 9 most frequent LFs and 10 randomly chosen headwords, it is also reasonable to expand the input data and to obtain a large list of LF-specific collocations.

We consider several applications of results achieved in course of experiments. A list of LF-specific collocations may be used in a set of computational semantics tasks requiring co-occurrence data: lexicon expansion for machine translation tasks (Protopopova et al. 2015), fact extraction and opinion mining (Protopopova et al. 2016), psycholinguistic profiling (Panicheva et al. 2016), automatic topic labelling, bigram-based topic modelling (Mirzagitova, Mitrofanova 2016), etc.

References

1. *Akinina, Y., Kuznetsov I., Toldova, S.* (2013) The impact of syntactic structure on verb-noun collocation extraction. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference "Dialogue". Pp. 2–17.
2. *Apresjan, Ju. D., Boguslavsky, I. M., Iomdin, L. L., Tsinman, L. L.* (2002) Lexical Functions in NLP: Possible Uses. In: Computational Linguistics for theNew Millenium: Divergence or Synergy? In: Proceedings of the International Symposium held at the Ruprecht-Karls-Universität Heidelberg, 21–22 July 2000. Manfred Klenner / Henriëtte Visser (eds.) Frankfurt am Main. Pp. 55–72.
3. *Apresjan, Ju. D., Djachenko, P. V., Lazursky A. V., Tsinman L. L.* (2007) On the Digital Textbook on Russian Lexica [O kompjuternom uchebnike russkogo jazyka]. In: The Russian Languiage in a Scientific Light [Russkij jazyk v nauchnom osveschenii], vol. 2(14). Pp. 48–112.
4. *Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D.* (2015) Breaking Sticks and Ambiguities with Adaptive Skip-gram. ArXiv preprint.
5. *Bukia, G. T., Protopopova, E. V., Panicheva, P. V., Mitrofanova, O. A.* (2016) Estimating Syntagmatic Association Strength Using Distributional Word Representations. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference "Dialogue", 2016. Vol. 15. pp. 112–122.
6. *Bukia, G., Protopopova, E., Mitrofanova, O.* (2015) A corpus-driven estimation of association strength in lexical constructions. In: Sergey Balandin, T. T., Trifonova, U.(eds.) Proceedings of the AINL-ISMW FRUCT. pp. 147–152. FRUCT Oy, Finland, http://fruct.org/publications/ainl-abstract/files/Buk.pdf
7. *Frolova, T. I., Podlesskaia, O. Ju.* (2011) Tagging Lexical Functions in Russian Texts of SynTagRus Tagging lexical functions in Russian texts of SynTagRus. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference "Dialogue". 2011. Vol. 10(17). Pp. 219–230.
8. *Khokhlova, M. V.* (2010) The Study of lexical-syntactic co-occurrence in Russian by statistical methods (based in text corpora) [Issledoanie leksiko-sintaksicheskoj sochetaemosti v russkom jazyke s pomoschju statisticheskih motodov]. PhD Thesis. Saint-Petersburg, 2010.
9. *Kilgarriff, A.* (2006) Collocationality (and how to measure it). In: Proceedings of the Euralex International Congress.
10. *Kochmar, E., Briscoe, T.* (2013) Capturing anomalies in the choice of content words in compositional distributional semantic space. In: RANLP. Pp. 365–372.
11. *Kolesnikova, O., Gelbukh A.* (2012) Semantic relations between collocations—A Spanish case study. Vol. 45, No. 78. Pp. 44–59.
12. *Kopotev, M. Escoter L., Kormacheva, D., Pierce, M., Pivovarova, L., Yangarber, R.* (2015) CoCoCo: Online Extraction of Russian Multiword Expressions. In: Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing BSNLP 2015. Hissar, Bulgaria. Pp. 43–45.
13. *Kormacheva, D., Pivovarova, L. & Kopotev, M.* (2014) Automatic Collocation Extraction and Classification of Automatically Obtained Bigrams. In: Proceedings: Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014). Pp. 27–33.

14. *Kormacheva, D., Pivovarova, L. & Kopotev, M.* (2016) Constructional generalization over Russian collocations. In: Mémoires de la Société néophilologique de Helsinki.
15. *Kutuzov, A., Kuzmenko, E.* (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Vol. 661. Springer.
16. *Mel'čuk, I., Zholkovsky A.*(1984) Explanatory Combinatorial Dictionary of Modern Russian [Tolkovo-kombinatornyj slovar russkogo jazyka]. Vienna, 1984.
17. *Mel'čuk, I. A.* (1998). Collocations and Lexical Functions. In: Anthony P. Cowie (ed.) Phraseology. Theory, analysis, and applications. Pp. 23–53. Oxford: Clarendon.
18. *Mikolov T., Yih, W.-t., and Zweig G.* (2013) Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of NAACL HLT.
19. *Mikolov, T., Chen, K., Corrado, G., and Dean, J.* (2013) Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR.
20. *Mirzagitova, A., Mitrofanova, O.* (2016) Automatic assignment of labels in Topic Modelling for Russian Corpora. In: Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016 / ed. A. Botinis.—Saint Petersburg: International Speech Communication Association.
21. *Palmer, H. E.* (1933) Second Interim Report on English Collocations, Tokyo: Institute for Research in English Teaching.
22. *Panchenko, A., Loukachevitch, N., Ustalov, D., Paperno, D., Meyer, C., Konstantinova, N.* (2015) RUSSE: The first workshop on Russian semantic similarity. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference "Dialogue". Pp. 89–105.
23. *Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N. and Biemann, C.* (2016) Human and Machine Judgements about Russian Semantic Relatedness. In: Proceedings of the 5th Conference on Analysis of Images, Social Networks, and Texts (AIST'2016). Communications in Computer and Information Science (CCIS). Springer-Verlag Berlin Heidelberg.
24. *Panicheva, P., Bogolyubova, O., Ledovaya, Y.* (2016) Revealing Interpetable Content Correlates of the Dark Triad Personality Traits. In: RUSSIR-2016. Russia, Saratov. Springer.
25. *Panicheva, P., Mitrofanova, O.* (2016) Developing a Toolkit for Distributional Analysis of Abnormal Collocations in Russia. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS), 2016. pp. 203–208.
26. *Protopopova, E, Bukia, G., Mitrofanova, O.* (2016) Sentiment analysis of reviews based on automatically developed lexicon. In: Proceedings of the 45th International Philological Conference. St. Petersburg State University, March 2016.
27. *Protopopova, E., Antonova, A., Misyurev, A.* (2015) Acquiring relevant context examples for A translation dictionary. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference "Dialogue".
28. *Rodríguez-Fernández, S., Anke, L., Carlini, R., Wanner, L.* (2016) Semantics-driven recognition of collocations using word embeddings'. In: Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany.

29. *Rundell, M.* (2010) Macmillan Collocations Dictionary, Macmillan.
30. *Tutubalina, E., Braslavski, P.* (2016) Multiple Features for Multiword Extraction: a Learning-to-Rank Approach, Computational Linguistics and Intellectual Technologies. In: Proceedings of the Annual International Conference “Dialogue”.
31. *Vecchi, E. M., Baroni, M., Zamparelli, R.* (2011) (linear) maps of the impossible: capturing semantic anomalies in distributional space. In: Proceedings of the Workshop on Distributional Semantics and Compositionality.
32. *Yagunova, E. V., Pivovarova, L. M.* (2010) The nature of collocations in the Russian language. The experience of automatic extraction and classification of the material of news texts. In: Automatic Documentation and Mathematical Linguistics, vol. 44, issue 3, Springer. Pp. 164–175.