JAZYKOVEDNÝ ČASOPIS

VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS

SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Hlavná redaktorka/Editor-in-Chief: doc. Mgr. Gabriela Múcsková, PhD.
Výkonní redaktori/Managing Editors: PhDr. Ingrid Hrubaničová, PhD., Mgr. Miroslav Zumrík, PhD.

Redakčná rada/Editorial Board: PhDr. Klára Buzássyová, CSc. (Bratislava), prof. PhDr. Juraj Dolník, DrSc. (Bratislava), PhDr. Ingrid Hrubaničová, PhD. (Bratislava), doc. Mgr. Martina Ivanová, PhD. (Prešov), Mgr. Nicol Janočková, PhD. (Bratislava), Mgr. Alexandra Jarošová, CSc. (Bratislava), prof. PaedDr. Jana Kesselová, CSc. (Prešov), PhDr. Ľubor Králik, CSc. (Bratislava), PhDr. Viktor Krupa, DrSc. (Bratislava), doc. Mgr. Gabriela Múcsková, PhD. (Bratislava), Univ. Prof. Mag. Dr. Stefan Michael Newerkla (Viedeň – Rakúsko), Associate Prof. Mark Richard Lauersdorf, Ph.D. (Kentucky – USA), prof. Mgr. Martin Ološtiak, PhD. (Prešov), prof. PhDr. Slavomír Ondrejovič, DrSc. (Bratislava), prof. PaedDr. Vladimír Patráš, CSc. (Banská Bystrica), prof. PhDr. Ján Sabol, DrSc. (Košice), prof. PhDr. Juraj Vaňko, CSc. (Nitra), Mgr. Miroslav Zumrík, PhD. (Bratislava), prof. PhDr. Pavol Žigo, CSc. (Bratislava).

Technický redaktor/Technical editor: Mgr. Vladimír Radik

Vydáva/Published by: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied
- v tlačenej podobe vo vydavateľstve SAP – Slovak Academic Press, s. r. o.
- elektronicky vo vydavateľstve Sciendo – De Gruyter
https://content.sciendo.com/view/journals/jazcas/jazcas-overview.xml

Adresa redakcie/Editorial address: Jazykovedný ústav Ľ. Štúra SAV, Panská 26, 811 01 Bratislava
Kontakt: gabriela.mucskova@juls.savba.sk
Elektronická verzia časopisu je dostupná na internetovej adrese/The electronic version of the journal is available at: http://www.juls.savba.sk/ediela/jc/

Vychádza trikrát ročne/Published triannually

Dátum vydania aktuálneho čísla (2019/70/2) – október 2019

CiteScore 2018: 0,24
SCImago Journal Rank (SJR) 2018: 0,122
Source Normalized Impact per Paper (SNIP) 2018: 0,476

JAZYKOVEDNÝ ČASOPIS je evidovaný v databázach/JOURNAL OF LINGUISTICS is covered by the following services: Baidu Scholar; Cabell's Directory; CEJSH (The Central European Journal of Social Sciences and Humanities); CNKI Scholar (China National Knowledge Infrastructure); CNPIEC – cnpLINKer; Dimensions; DOAJ (Directory of Open Access Journals); EBSCO (relevant databases); EBSCO Discovery Service; ERIH PLUS (European Reference Index for the Humanities and Social Sciences); Genamics JournalSeek; Google Scholar; IBR (International Bibliography of Reviews of Scholarly Literature in the Humanities and Social Sciences); IBZ (International Bibliography of Periodical Literature in the Humanities and Social Sciences); International Medieval Bibliography; J-Gate; JournalGuide; JournalTOCs; KESLI-NDSL (Korean National Discovery for Science Leaders); Linguistic Bibliography; Linguistics Abstracts Online; Microsoft Academic; MLA International Bibliography; MyScienceWork; Naviga (Softweco); Primo Central (ExLibris); ProQuest (relevant databases); Publons; QOAM (Quality Open Access Market); ReadCube; SCImago (SJR); SCOPUS; Sherpa/RoMEO; Summon (ProQuest); TDNet; Ulrich's Periodicals Directory/ulrichsweb; WanFang Data; WorldCat (OCLC).

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA
SLOVENSKEJ AKADÉMIE VIED

2

ROČNÍK 70, 2019

NLP, Corpus Linguistics, Language Dynamics and Change

SLOVKO 2019

Tematické číslo Jazykovedného časopisu venované počítačovému spracovaniu prirodzeného jazyka, korpusovej lingvistike, jazykovej dynamike a zmenám.

Prizvaní editori: Mgr. Jana Levická, PhD.
Mgr. Miroslav Zumrík, PhD.

sciendo

SAP
SLOVAK ACADEMIC PRESS

CONTENTS

CORPUS-BASED/DRIVEN LINGUISTIC AND DH RESEARCH

CORPUS BUILDING

CREATION AND USE OF LANGUAGE RESOURCES

NATURAL LANGUAGE PROCESSING

sciendo
DOI 10.2478/jazcas-2019-0054

WAYS OF AUTOMATIC IDENTIFICATION OF WORDS BELONGING TO SEMANTIC FIELD

VICTOR ZAKHAROV
Saint-Petersburg State University, Russia

ZAKHAROV, Victor: Ways of automatic identification of words belonging to semantic field. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 234 – 243.

Abstract: The paper presents results of the ongoing research on creation of the semantic field of the "empire" concept. A semantic field is a collection of content units covering a certain area of human experience and forming a relatively autonomous microsystem with one or several centers. Relations in such microsystems are also called associations. The idea is to extract from data on syntagmatic collocability a set of lexical units connected by systemic paradigmatic relations of various types and strength using distributional analysis techniques. The first goal of the study is to develop methodology to fill a semantic field with lexical units on the basis of morphologically tagged corpora. We were using the Sketch Engine corpus system that implements the method of distributional statistical analysis. Text material is represented by our own corpora in the domain of "empire". In the course of the work we have acquired lists of items filling the semantic space around the concept of "empire".

Keywords: semantic field, concept of empire, distributive and statistical analysis, corpus, thesaurus

1 INTRODUCTION

Many automated text processing systems are based on dictionaries, including semantic ones. Semantic fields can also be regarded as semantically simple dictionaries. The notion of "semantic field" is used in linguistics to denote a set of language units with a common semantic feature with some common meaning component. Words (both common and proper nouns) and word groups constitute such lexical units. To cite O. S. Akhmanova, "A field is a set of content units covering a certain field of human experience and forming a relatively autonomous microsystem" [2].

First attempts to identify semantic fields were undertaken when conceptual dictionaries, or thesauri (for example, Roget's Thesaurus) were created. As construed by V. G. Admoni, a field is characterized by an inventory of elements related through systemic relations. V. G. Admoni discovers a central part in a field, the nucleus, whose elements have a full set of features defining this group, and the periphery whose elements has some of the features characteristic of the field, but can have features peculiar to neighboring fields [1]. A field consists of a continuity of links between objects of a set, and the links in some areas are particularly dense. In such cases, we speak of lexico-semantic groups or elementary microfield grouping words that are, as a rule, of the same part of speech. In a general case, blurred boundaries between minifields and parts of speech are characteristic of a field. Many papers deal with the semantic field theory ([4], [18], [21], etc.).

The task of modelling a conceptual or terminological system has been a subject of computational linguistics for a long time. It can be divided into two parts, namely: identification of lexical identifiers of concepts and identification of relations between them. In this study, we are interested in the first task. It can be solved "manually" through explication and formalization of professional knowledge accumulated in the course of human activities. However, since our knowledge of the world is to a lesser or greater extent reflected in texts, then a task of extraction of a system of concepts from texts can be set.

In this paper, we deal with the semantic field "empire". The paper is a part of a greater research dedicated to comparative analysis of the content of this field in Russian, English, Czech and German. The choice of languages can be explained by the fact that the concept of "empire" in these languages is closely related to the historic memory of the people and that it is "alive" in the linguistic consciousness of the native speakers.

2 RESEARCH METHOD

The research method is a corpus-oriented analysis of the paradigmatics and syntagmatics of lexical units using distributional statistical methods. In our study, we used existing tool and corpus linguistic processors and the data from the ad hoc annotated corpora.

One of the earliest and best-known linguistic research methods is distributional statistical analysis where information on the distribution of text elements and their numerical parameters are used. There is an idea that paradigmatic links which belong to language system can be derived from the syntagmatic ones, i.e. neighborhood of words in a linear text chain (see [3], [14], [16], [17]). The principle of transfer from studying textual (syntagmatic) links to systemic (paradigmatic) ones underlies various distributional statistical methods. The assumption is that two elements are linked by paradigmatic relation if both of them are textually systematically linked with some other third element. However, for it to be possible to speak of regularities in any statistical distribution, large bodies of data and, consequently, massive computational power are needed.

3 MECHANISM OF CREATION OF LEXICO-SEMANTIC GROUPS AND FIELDS

As was already mentioned, paradigmatic links can be derived from syntagmatic ones. This concept was suggested by A. Ya. Shaykevich [17] and K. S. Jones [6] as early as in the 1960s. The mathematical apparatus for computing this similarity was developed later by D. Lin [13]. But this approach has not been implemented until now though it has become possible to create a large database of co-occurrences of lexical units on the basis of text corpora and to "compute" a set of "the closest neighbors" for every word on the basis of such database (see [11], [19]).

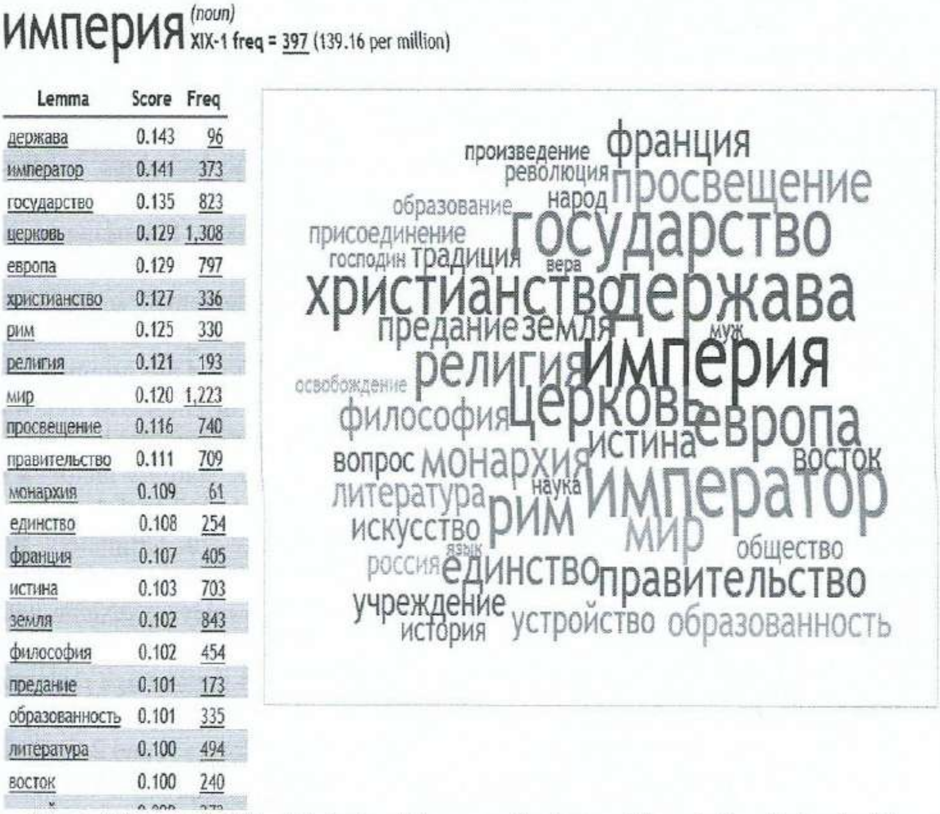
However, it is also important to take into account a syntactic relation between contextually close elements [5]. In the Sketch Engine system ([8], [9], [10], [12]), which was used to build corpora and identify syntagmatic and paradigmatic links, the concept of lexico-syntactic patterns is implemented as so-called *word sketches,* an automatically generated summary of the co-occurrences restricted to the set of syntactic formulas. Word sketches are based on sets of rules describing grammatical relations between words in a text (word sketch grammar).

When building a corpus on the basis of morphologically tagged data, a special database consisting of triplets of lexico-grammatical relations is built. Statistical processing of this database computes the data for a distributional thesaurus which, for us, is similar to a lexico-semantic group for a specific term. The algorithm to compute semantic distance between the elements in the group (candidates) is described in [20, sect. 3, 4]. The similarity in distribution of words is calculated statistically on the basis of an association measure logDice [15] taking into account the grammar of lexico-syntactic patterns [11]. In our research, the distributional statistical analysis is based on the grammar of lexico-syntactic patterns for the Russian language developed by M.V. Khokhlova [7].

4 STUDY MATERIAL AND TOOLS

The main study material is an ad hoc corpus (10.25 mln tokens). It has been built from texts relating to the topic of empire in the Russian literature and culture of the end of the 18th and the beginning of the 20th century collected within the scientific project of the Institute of the Russian literature, St. Petersburg. Words' meanings are subject to continuous change, thus we have to take into account the temporal dimension and the diachronic nature of words. That is why we divided the corpus into 4 subcorpora on a chronological principle: the 18th century (the corpus identifier –XVIII), the 1st half of the 19th century (XIX-1), the 2nd half of the 19th century (XIX-2) and the 20th century (XX). The boundary dates of the subcorpora were chosen as some sort of milestones in the perception of the "empire" concept in the development of the Russian social thinking.

As has been mentioned already, we have used the Sketch Engine system for purposes of our research. Its main feature is availability of special tools implementing the distributional statistical analysis method – *Thesaurus* (building a lexico-semantic group for a specific term; see Fig. 1), *Clustering* (grouping thesaurus units into narrower clusters), and *Collocations* (extraction of steady word combinations, collocations).



империя (noun) XIX-1 freq = 397 (139.16 per million)

Lemma	Score	Freq
держава	0.143	96
император	0.141	373
государство	0.135	823
церковь	0.129	1,308
европа	0.129	797
христианство	0.127	336
рим	0.125	330
религия	0.121	193
мир	0.120	1,223
просвещение	0.116	740
правительство	0.111	709
монархия	0.109	61
единство	0.108	254
франция	0.107	405
истина	0.103	703
земля	0.102	843
философия	0.102	454
предание	0.101	173
образованность	0.101	335
литература	0.100	494
восток	0.100	240

Fig. 1. A fragment of the distributional thesaurus for the word "empire" on the basis of the subcorpus of the 1st half of the 19th century with semantic link strength value (score)

Translation of terms in Fig. 1: держава 'state, power', император 'emperor', государство 'state', церковь 'church', Европа 'Europe', христианство 'Christianity, Рим 'Rome', религия 'religion', мир 'world', просвещение 'enlightenment', правительство 'government', монархия 'monarchy', единство 'unity', Франция 'France', истина 'truth', земля 'land', философия 'philosophy', предание 'legend', образованность 'education', литература 'literature', восток 'East'.

A considerable portion of terms in any subject area are, as a rule, represented by word combinations. The tool calculating the strength of syntagmatic links between lexical units is Collocations. It computes the association of the units in a linear sequence based on 7 association measures. It should be added that this tool identifies not only syntagmatic links, but also paradigmatic ones if the "window" for collocates is sufficiently large.

5 CORPUS-BASED ANALYSIS OF PARADIGMATIC AND SYNTAGMATIC RELATIONS

5.1 Research technique

The Russian and Czech corpora were used for the research. A technique for creating a semantic field was developed. The essence of the technique is given below.

Two approaches to identification of lexical units presumably belonging to the semantic field "empire" were used, namely: creation of a distributional thesaurus, creation of a list of collocations. These methods are implemented in the four subcorpora mentioned above. The thesauri terms derived using each subcorpus (minithesauri) are ranked by the association score (see Fig. 1) and are grouped into a summary array. The number of words in each minithesaurus in Thesaurus tool was limited to 40. As a result, it counts 160 term occurrences. But it important how stable an appropriate word is represented throughout the corpus. The "stability factor" k is assigned to each lexical unit (term or word group) We set k = 1, 2, 3 or 4 depending on the number of minithesauri that include an appropriate word. Ultimately, the average rank is calculated for all the units of the summary array as well as the normalized rank that represents the semantic association of a relevant lexeme with the head word "empire", i.e. it is the "appropriateness" factor for this semantic field (Table 1). The normalized rank is derived by multiplication of the average rank and the "rank normalization factor". The following factors were chosen empirically: 1 for the terms occurring in all the four minithesauri (k=4), 1.5 for k=3, and 2 for k=2. The terms extracted from one subcorpus only are not included in the field. Thus, these factors reduce the ranks of the terms related to the word "empire" in larger number of subcorpora (i.e. in more time periods).

Subcorpus	*Rank in the subcorpus thesaurus*	*Lemma*	*Score*	*Freq*	*Stability factor*	*Average rank*	*Norm. rank*
XIX-2	**36.**	**англия** 'England'	**0.131**	**1055**	**2**	**29**	**58**
XVIII	22.	англия 'England'	0.095	148	2		
XIX-2	19.	армия 'army'	0.149	478	1		
....	....	....	.......,,	......	.....	....	....
XIX-2	**24.**	**государственность** 'statehood'	**0.143**	**201**	**2**	**19**	**38**
XX	14.	государственность 'statehood'	0.141	143	2		
XX	**1.**	**государство** 'state'	**0.245**	**1016**	4	**2.25**	**2.25**
XIX-2	2.	государство 'state'	0.200	4240	4		
XVIII	3.	государство 'state'	0.184	766	4		
XIX-1	3.	государство 'state'	0.135	823	4		
XX	2.	гуманизм 'humanism'	0.188	195	1		

Subcorpus	*Rank in the subcorpus thesaurus*	*Lemma*	*Score*	*Freq*	*Stability factor*	*Average rank*	*Norm. rank*
XVIII	**2.**	**держава** 'state, power'	**0.189**	**424**	**3**	**4.3**	**6.45**
XIX-2	10.	держава 'state, power'	0.165	606	3		
XIX-1	1.	держава 'state, power'	0.143	96	3		
....	....	....	......,,	......	....	....	....
XIX-1	13.	единство 'unity'	0.108	254	1		
XIX-2	**5.**	**император** 'emperor'	**0.184**	**1381**	**3**	**4**	**8**
XX	5.	император 'emperor'	0.177	295	3		
XIX-1	2.	император 'emperor'	0.141	373	3		
XX	8.	империализм 'imperialism'	0.166	297	1		
....	....	....	....	....	....	....	....

Tab. 1. Summary distributional thesaurus for the word "empire" (a fragment)

5.2 Preliminary results

As stated above, summary distributional thesaurus included 160 entries. 79 words occur once, i.e. only in one of the minithesauri, and the distribution of these unique words in the subcorpora is as follows: subcorpus XVIII: 32 words, subcorpus XIX-1: 16, subcorpus XIX-2: 14, subcorpus XX: 17.

The remaining 81 entries which consisted of 33 different words occur in 2, 3 or 4 minithesauri. We call these 33 words the nucleus of the semantic field. Their distribution in subcorpora is as follows: 8 in the 18th century, 24 in the 1st half of the 19th century, 26 in the 2nd half of the 19th century, and 23 in the 20th century.

Here is the list of the nucleus of the semantic field "empire" derived through the implementation of the proposed technique, ranked by different bases (the beginning of the list is given):

a) by the normalized rank:

государство 'state', *император* 'emperor', *держава* 'state', *Европа* 'Europe', *царство* 'kingdom', *церковь* 'church', *Рим* 'Rome', *Франция* 'France', *христианство* 'Christianity', *монархия* 'monarchy', *правительство* 'government', *страна* 'country', *общество* 'society', *нация* 'nation', *Россия* 'Russia', *государственность* 'statehood'...

b) by the average semantic association factor (score):

держава 'state, power), *государство* 'state', *общество* 'society', *союз* 'union', *государственность* 'statehood', *нация* 'nation', *император* 'emperor', *политика* 'politics', *культура* 'culture', *страна* 'country', *община* 'community', *церковь* 'church', *царство* 'kingdom', *христианство* 'Christianity', *религия* 'religion', *мир* 'world', *просвещение* 'enlightment', *правительство* 'government', *монархия* 'monarchy'...

c) alphabetically:

Англия 'England', *государственность* 'statehood', *государство* 'state', *держава* 'state, power', *Европа* 'Europe', *император* 'emperor', *искусство* 'art', *история* 'history', *культура* 'culture', *литература* 'literature', *мир* 'world', *монархия* 'monarchy', *наука* 'science', *нация* 'nation', *общество* 'society', *община* 'community', *политика* 'politics', *правительство* 'government'...

d) by the relative frequency (ipm, instances per million):

Россия 'Russia', *общество* 'society', *церковь* 'church', *мир* 'world', *история* 'history', *государство* 'state', *наука* 'science', *просвещение* 'englightment', *правительство* 'government', *держава* 'state', power', *политика* 'politics', *царство* 'kingdom', *литература* 'literature', *революция* 'revolution', *союз* 'union', *страна* 'country', *Европа* 'Europe', *община* 'community', *культура* 'culture', *император* 'emperor'...

Let's give a list of bigram collocations that are candidates to the semantic field "empire" (ranked according to the log-Dice measure).

Total 115 bigrams were extracted, with the majority of them being bigrams including a form of the word *империя* 'empire' plus some other word: Adj+*империя*, *империя*+N (genitive), N+*империи* (genitive) where Adj states for adjective, N for noun.

24 bigrams are the nucleus of the syntagmatic collocations according to the normalized rank: *Российская империя* 'Russian Empire', *Византийская империя* 'Byzantine Empire', *Восточная империя* 'Eastern Empire', *Священная империя* 'Holy Empire', *падение империи* 'fall of the empire', *Австрийская империя* 'Austrian Empire', *Великая империя* 'Great Empire', *пределы империи* 'borders of the empire', *Турецкая империя* 'Turkish Empire', *столица империи* 'capital of the empire', *etc*.

The experiments with the Czech language were conducted using the synchronous National Corpus of the Czech Language. The collocation search was conducted using the corpus SYN2015 (`https://kontext.korpus.cz/first_form-?corpname=syn2015`). However, the Nosketch Engine system supporting Czech corpora lacks the Thesaurus tool. This is why a corpus on our topic has been created from the Czech Internet (342 mln tokens). To avoid peripheral vocabulary in the resulting field, the output of the distributional thesaurus was limited to 30.

Here are several examples for the Czech language for the word říše (empire) (the beginning of the lists is given):

a) by the normalized rank:

království 'kingdom', *civilizace* 'civilization', *Británie* 'Britain', *Rusko* 'Russia', *společenství* 'community', *vesmír* 'space', *Řím* 'Rome', *impérium* 'empire'...

b) by the semantic association factor (score measure in the distributional thesaurus):

civilizace 'civilization', *království* 'kingdom', *země* 'country', *impérium* 'empire', *Británie* 'Britain', *Amerika* 'America', *armáda* 'army', *lidstvo* 'mankind', *monarchie* 'monarchy'...

The beginning of the collocation list is as follows:

Třetí říše 'Third empire', *Římská říše* 'Roman empire', *Osmanská říše* 'Ottoman empire', *Německá říše* 'German empire', *Svatá říše* 'Holy empire', *Velkomoravská říše* 'Great Moravian empire', *vládce říše* 'ruler of the empire', *zánik říše* 'fall of the empire'...

Anyone who is familiar with Czech culture will agree that these terms do have a strong semantic link with the word *říše* 'empire'.

5.3 Conclusion and further work

As we can see, the use of a text corpus and "smart" corpus tools allow to automatically extract syntagmatic and paradigmatic relations and create rather reasonable content of a term system. Moreover, obtained lists significantly expand the existing lexicographic guides. However, the question is where are the limits of the field "empire". We see in the periphery of the summary distributional thesaurus such words as посол 'embassador', отечество 'Motherland', воевода 'battlemaster', воин 'soldier', etc. that hardly belong to the field "empire". This encourages us to repeat the experiments with "stricter" parameters of corpus tools. At the same time the work to identify elements semantically related to terms included in the nucleus of the field "empire" will be conducted, i.e. a task to create second-stage thesauri (minifields) and form a single list made, if possible, as a semantic network.

One can note that the concept "empire" in Russian had different connotations during different periods in the Russian culture defined by different parameters. For example, significant specific feature of the 18th century texts catches the attention. This is evident in the contents of the vocabulary – see Section 4.2: 32 words of 79 words "unique" for only one period belong to the 18th century and only 8 words from the field nucleus are listed in 18th century minithesaurus. In general, it can be carefully concluded that despite the fact that the empire existed de facto in the 18th century, the very concept of the empire did not form in the Russian culture at the time.

It is interesting to get interpretation of the results linked to historical or cultural aspects concerning different languages. That is why we have started selecting texts for a parallel English-Russian, Czech-Russian and English-Czech corpora. It appears that the volume of the corpora will be small due to the difficulties in selecting parallel texts, but we think that it is important to do that because the elements of the field will be in the same temporal and historical paradigm in these texts. It is also interesting to see which words (and why) will prevail in translation of the same concept: for example, the Czech *říše* can be translated into Russian as империя 'empire', королевство 'kingdom', царская власть 'reign', рейх 'Reich'. The Russian империя is translated into Czech as *impérium* 'empire', *říše* 'empire', *císařství* 'empire', *država* 'domain', etc. The same goes for other terms and other language pairs.

Finally, it can be stated that the task of building one small semantic field reflects the peculiarities of the lexico-semantic system of a language as well as opportunities and barriers in automation of semantic processing.

ACKNOWLEDGMENTS

This work was implemented with the financial support of the Russian Foundation for Basic Research, Project No. 18-012-00474 «Semantic field "empire" in Russian, English and Czech» and partly with the Project No. 17-04-00552 "Parametric modeling of the lexical system of the modern Russian literary language".

References

[1] Admoni, V. G. (1973). Syntax of modern German: The system of the relations and the system of construction, [Sintaksis sovremennogo nemeckogo jazyka: Sistema otnoshenij i sistema postroenija], Leningrad.

[2] Akhmanova, O. S. (1966). Dictionary of Linguistic Terminology [Slovar' lingvisticheskikh terminov]. Moscow.

[3] Arapov, M. V. (1964). Some principles of creation of the "thesaurus" dictionary NTI Serie 2(4), pages 40–46.

[4] Askoldov, S. A. (1980). Concept and word, [Koncept i slovo]. Moscow.

[5] Gamallo, P., Gasperin, C., Augustini, A., and Lopes, G. P. (2001). Syntactic-Based Methods for Measuring Word Similarity, In Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166, pages 116–125. Springer-Verlag.

[6] Jones, K.S. (1965). Experiments in semantic classification, Mechanical Translation and Computational Linguistics, 8(3–4), pages 97–112.

[7] Khokhlova, M.V. (2010). Development of the grammatical module of Russian for the specialized system of processing of corpus data [Razrabotka grammaticheskogo modulja russkogo jazyka dlja specializirovannoj sistemy obrabotki korpusnyh dannyh], Bulletin of St. Petersburg State University [Vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta], Series 9, Philology, oriental studies, journalism. 2(9), pages 162–169.

[8] Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus, In Proceedings of the 13th EURALEX International Congress. Spain, July 2008, pages 425–432. EURALEX.

[9] Kilgarriff, A., Rychlý, P., Jakubíček, M., Rundell, M. et al.: Sketch Engine [Computer Software and Informatiom Resource]. Accessible at: `http://www.sketchengine.co.uk`.

[10] Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwel, D. (2004), The Sketch Engine, In Proceedings of the XIth Euralex International Congress, pages 105–116. Lorient, Universite de Bretagne-Sud.

[11] Kilgarriff, A., and Rychlý, P. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments), In Proceedings of the 45th Annual Meeting of the ACL. Interactive Poster and Demonstration Sessions. Czech Republic, June 2007, pages 41–44. ACL.

[12] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1(1), pages 7–36.

[13] Lin, D. (1998). Automatic retrieval and clustering of similar words. In Proc. COLING-ACL, pages 768–774. Montreal.

[14] Pekar, V. (2004), Linguistic Preprocessing for Distributional Classification of Words. In Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries, pages 15–21, Geneva.

[15] Rychlý, P. (2008). A lexicographer-friendly association score, In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, pages 6–9. Brno.

[16] Shaykevich, A. Ya. (1982). Distributive and statistical analysis of texts [Distributivno-statisticheskij analiz tekstov], PhD thesis. Leningrad.

[17] Shaykevich, A.Ya. (1963). Distribution of words in the text and allocation of semantic fields [Raspredelenie slov v tekste i vydelenie semanticheskih polej], In Foreign languages in higher education, 2, pages 14–26, Moscow.

[18] Shchur, G.S. (1974). Field theory in linguistics, [Teorija polja v lingvistike], Moscow-Leningrad.

[19] Smrž, P., and Rychlý, P. (2001). Finding Semantically Related Words in Large Corpora, In Text, Speech and Dialogue: Fourth International Conference (TSD–2001), LNAI 2166, pages 108–115. Springer-Verlag.

[20] Statistics Used in Sketch Engine. Accessible at: `https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine`.

[21] Wierzbicka, A. (2001). Understanding of cultures through keywords, [Ponimanie kul'tur cherez posredstvo kljuchevyh slov]. Moscow.