RASLAN 2019
Recent Advances in Slavonic Natural Language Processing

A. Horák, P. Rychlý, A. Rambousek (eds.)

RASLAN 2019

Recent Advances in Slavonic Natural Language Processing

Thirteenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2019
Karlova Studánka, Czech Republic, December 6–8, 2019
Proceedings

Tribun EU
2019

Proceedings Editors

Aleš Horák
Faculty of Informatics, Masaryk University
Department of Information Technologies
Botanická 68a
CZ-602 00 Brno, Czech Republic
Email: hales@fi.muni.cz

Pavel Rychlý
Faculty of Informatics, Masaryk University
Department of Information Technologies
Botanická 68a
CZ-602 00 Brno, Czech Republic
Email: pary@fi.muni.cz

Adam Rambousek
Faculty of Informatics, Masaryk University
Department of Information Technologies
Botanická 68a
CZ-602 00 Brno, Czech Republic
Email: rambousek@fi.muni.cz

ISBN 978-80-263-1530-8
ISSN 2336-4289

Table of Contents

IV Text Corpora

The Concept of 'empire' in Russian and Czech

Victor Zakharov

Saint-Petersburg State University
Universitetskaya emb. 7-9
199034 Saint-Petersburg, Russia
v.zakharov@spbu.ru

Abstract. The paper is dealing with subject of forming semantic fields. The 'empire' semantic field in 2 languages (Russian, Czech) was chosen as an object of investigation. The paper describes a descriptive statistical method of forming semantic fields based on linguistic corpora. The result is a specific lexicographic product (distrubutive thesaurus) for each language with quantitative characteristics of the connectedness of lexical units. At the last step linguistic correlation between elements of these two thesauri is shown. The research is implemented on the basis of Sketch Engine and Czech National Corpus. In the aspect of theory, we get a fragment of the semantic description of Russian and Czech languages and a description of new methods for analyzing vocabulary and semantics of the language.

Keywords: concept, semantic field, concept of empire, distributional thesaurus, Russian, Czech, corpora

1 Introduction

The subject of the study is to consider the concept of 'empire' in Russian and Czech. We mean here the term, by which one determines the content plan of the word, i.e. the notion, fixed in the language and correlated with other notions associated with it. Our task is to reveal the lexical content of these interconnected notions defined by the named concept.

Concepts underlie what linguists and cognitologists call the linguistic image of the world. This is a set of ideas about the world historically formed in the everyday consciousness of a given community and reflected in the language, in other words, this is conceptualization of reality. The linguistic image of the world determines the various aspects of the language, its vocabulary, its capability to generate words, to influence the syntax of phrases and sentences, as well as paremiological layer of language. Only linguistic images of the world in specific national languages really exist and can be analyzed, this is national linguistic images of the world. The linguistic image of the world is time-varying. In this work, we are interested in the current state of the language.

The set of lexical units each of which has some common component of the meaning forms a semantic field. The field is characterized by the presence of an inventory of elements connected by systemic relationships. It has a central

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 89–97, 2019.

part, the core, the elements of which have a complete set of features that define this grouping, and the periphery, elements of which do not have all the features characteristic of the field. The field implies the continuity of the connections of set units. Fields are characterized by the possibility of quantitative expression of the strength of relations between field members.

The choice of Russian and Czech languages is due to the fact that in both languages the concept of 'empire' is strongly connected with the historical memory of the people and that it is "alive" in the linguistic consciousness of native speakers.

2 Statement of the problem and research methodology

Unlike of psycholinguistics the task of computational linguistics is automatic selection of lexical units for semantic fields. The method uses distributional statistical analysis based on linguistic corpora. The corpus approach, however, does not exclude the subsequent involvement of expert knowledge.

The objective of our study is to create two associative thesauri with quantitative characteristics of lexical units for two languages with examples from corpora. In this paper, we solve the problem of selecting the vocabulary of the empire semantic field in each language, getting statistical characteristics of lexical units on the basis of corpora, and the identification of Russian-Czech (Czech-Russian) translation equivalents of semantic field units.

There are two aspects of functioning of a linguistic unit, syntagmatics and paradigmatics. The methodology of the study is a corpus-oriented analysis of the paradigmatics and syntagmatics of lexical units, which form the semantic field for the word empire. Our materials are corpora with linguistic tagging and corpus linguistic processors. At the same time, the other lexicographic resources might be included in the analysis if necessary.

Corpus linguistics made it possible to "calculate" different types of compatibility which are combined under the term multiword expressions. But if syntagmatic relations are explicitly presented in text and can be extracted from it on the basis of a linear sequence, paradigms are hidden and it requires extralinguistic knowledge and/or sophisticated procedures should be developed to extract them from texts.

The building of semantic field is the task of modeling the conceptual subsystem of a language. Since our knowledge of the world is reflected in texts, we can set the task of extracting a system of concepts from texts. In this paper, we try to extract interrelated units grouping around the core notion of empire, starting from keywords that most closely express the meaning of the core concept.

Already at the dawn of computational linguistics, the idea was put forward that paradigmatic connections could be deduced from syntagmatic connections. The principle of the transition from the study of textual (syntagmatic) links to systemic (paradigmatic) underlies various distribution and statistical techniques

[6, 9, 10]. It was believed that two elements were connected paradigmatically if both of them are textually systematically connected with some third elements.

However, the capabilities of computational technology for a long time did not allow to put these ideas into practice. In order to talk about the regularity of any statistical distributions, very large data sets are needed. That became possible only with the development of the web and the creation of large text corpora. At the same time, appropriate software tools appeared [1, 4, 8]. Attention was drawn to the fact that it was also important to take into account the occurrence of a syntactic relationship between contextually close elements of the text [2, 5].

3 Research material and tools

In this work, the Sketch Engine system (https://app.sketchengine.eu) was mainly used for the research. We used ruTenTen 2011 corpus and csTenTen 2017 corpus and also we used the Czech National corpus (ČNK) (syn v7 and Treq).

The advantage of the Sketch Engine for our purposes is its special tools that make possible distributional statistical analysis, they are "Thesaurus" (building a distributional thesaurus) and "Clustering" (grouping of thesaurus units in clusters, i.e. lexical-semantic groups).

The thesaurus in Sketch Engine allows to see which words have a similar distribution with the given word, which, as a rule, is caused by their semantic proximity, i.e., in fact, this tool forms a uniterm semantic field. Word distribution similarity is calculated statistically, calculation is based on the association measure logDice [7] and lexical-syntactic patterns [3]. In the next step the inclusion in the semantic field the characteristic stable phrases is provided by the Collocations tool.

4 Technology of formation of the core of the empire semantic field

At the first stage, various lexicographic sources were used to describe the concept of empire in terms of keywords. Analysis of dictionary definitions from various Russian and Czech dictionaries made it possible to identify the main meanings and, respectively, semantic attributes of the concept of empire:

1) monarchy, headed by the emperor;

2) large state, consisting of several parts, possibly colonies;

3) metaphoric meanings derived from one of the first two (e.g. a large enterprise, parts of the natural world, etc.).

In our analysis, we deal only with vocabulary related to the first concept.

A technology of formation of semantic fields based on the diachronic approach was developed and tested, then data from text corpora printed in different historical periods were analyzed. For a detailed account, see [12]. In this paper, we are interested in a synchronous approach, how the concept of empire in modern Russian and Czech texts is implemented.

As a result of a definitional analysis of explanatory dictionaries and dictionaries of synonymy, elementary units of a meaningful plan were identified, 10 lexemes in each language. In doing so, we sought that these terms be monosemic.

Lexical identifiers of the concept of empire in Russian are as follows: государь (sovereign), держава (power), династия (dynasty), император (emperor), императрица (empress), империя (empire), монарх (monarch), монархия (monarchy), правитель (ruler), самодержавие (autocracy). Lexical identifiers of the concept of empire in Czech are as follows: *císař* (the emperor), *císařství* (empire). *dynastie* (dynasty), *impérium* (empire), *král* (king), *mocnářství* (monarchy), *monarchie* (monarchy), *panovník* (ruler), *říše* (empire), *vládce* (ruler).

Then for each of them 10 distributional thesauri were built in Sketch Engine on the basis of ruTenTen 2011 and csTenTen 2017 corpora (Fig. 1). In order to avoid getting into the resulting field of nonrelevant vocabulary the volume of the distributional thesaurus was limited to 15.

císař (noun)
Czech Web 2017

Lemma	Score	Freq
král	0.373	975,654
panovník	0.350	88,279
papež	0.348	208,554
kníže	0.321	151,336
vůdce	0.296	300,593
královna	0.294	249,985
vládce	0.292	119,743
biskup	0.279	231,650
prezident	0.276	1,510,494
generál	0.265	216,111
bratr	0.262	773,133
velitel	0.259	310,254
otec	0.254	1,384,487
ministr	0.250	1,316,050
premiér	0.248	493,872

Fig. 1: The distributional thesaurus (semantic field) for the word *říše*

The important characteristics here are the coefficient of the semantic proximity of the lexemes with a headword (score) and their frequency (freq).

We can suggest the language homogeneity in the selected corpora. Both of them are created on the base of texts from web, and contain mainly modern texts, both corpora are created using the same technology. We can say that they contain the vocabulary of the modern language and thereby reflect the modern state of linguistic consciousness.

On next stage, all 10 thesauri were put together into one dataset. Moreover, for each term, the average score was calculated. The assumption was made empirically that if a lexeme occurs in at least N thesauri (we call N the stability

coefficient), it is a candidate for inclusion in the core of the semantic field. The lexemes with value of the score less than N form its periphery. Both in the center and in the periphery area the lexemes can be sorted according to their score.

Further, for each element of the field core the most characteristic bigram collocations were identified using ČNK syn v7 corpus and Collocations tool. Bigrams were sorted by the MI.log_f association measure as one of the most effective ones.

5 The results obtained

5.1 Empire semantic field in modern Russian

The intersection of 10 thesauri (150 lexical units) yielded 79 unique lexemes, of which 17 met 3 or more times (in 3 or more thesauri), 19 - 2 times and 43 once. 17 units that have occurred 3 or more times form the core of the empire semantic field. They are as follows (in alphabetic order): владыка (lord), вождь (leader), государственность (statehood), государь (sovereign), держава (power), династия (dynasty), император (emperor), императрица (empress), империя (empire), князь (prince), король (king), монарх (monarch), монархия (monarchy), папа (pope), правитель (ruler), принц (prince), самодержавие (autocracy), царство (kingdom), царь (tsar), цивилизация (civilization).

It is interesting to note that the initial lexical identifiers of the concept of empire, which we took from dictionaries, appeared in the consolidated distributional thesaurus only 2 times (*power* and *dynasty*) and 1 time (*autocracy*). We have included them in the core for now. But since we consider our corpora as a model of a modern language, we can say with caution that these concepts are gradually leaving the concept of empire.

Perhaps the explanation of the appearance in this list the polysemous word папа (in conversational Russian 'dad') requires clarification. An analysis of corpus contexts showed that it was about the concept of the Pope of Rome which has a close connotation with the monarchs.

The periphery of the field includes 59 lexemes such as абсолютизм (absolutism), Англия (England), аристократия (aristocracy), властитель (lord, sovereign), Германия (Germany), государство (state) etc.

Also collocations will be added to the empire semantic fields both in Russian and in Czech.

5.2 Empire semantic field in modern Czech

The intersection of 10 thesauri (in total 150 lexical units) gave 88 unique lexemes, of which 13 met 3 or more times, 20 - 2 times and 55 once. If we take the stability coefficient equal to 3, then 13 lexemes form the core of the empire semantic field for the Czech language. Interestingly, for the Czech language, three of the original identifiers of the concept of empire which we took from Czech dictionaries were found in the combined distributional thesaurus for the Czech

language only 2 times (*císařství*, *dynastie*, *mocnářství*). However, we included them in the core of the semantic field for the Czech language.

The full list of the core of the empire semantic field for the Czech language is as follows (in alphabetic order): *císař* (emperor), *císařství* (empire), *dynastie* (dynasty), generál (general), impérium, impérium (empire), kníže (prince), král (king), *královna* (queen), *království* (kingdom), *mocnářství* (monarchy), *monarchie* (monarchy), *panovník* (ruler), *říše* (empire), *velitel* (commander), *vládce* (ruler), *vůdce* (leader). The periphery of the field includes 72 lexemes.

5.3 Comparison of the core of the empire semantic field in Russian and Czech

Let's try to compare the filling of the empire sematic field in Russian and Czech. If we temporarily exclude from consideration lexemes that mean roughly the same in Russian and Czech and lexemes that are present only in one of the language fields (государственность (statehood), папа (pope), князь (prince), цивилизация (civilization), generál (general), *velitel* (commander)), then lexemes related to the two microfields will remain.

The first microfield contains different names for the concept of empire: in Russian they are империя (empire), царство (kingdom), держава (power), partly монархия (monarchy); in Czech *impérium* (empire), *říše* (empire), *království* (kingdom), *císařství* (empire), *mocnářství* (monarchy), partly *monarchie* (monarchy). The second microfield contains different names for the concept of emperor: in Russian, they are монарх (monarch), правитель (ruler), царь (tsar), владыка (ruler), государь (sovereign), император (emperor), императрица (empress); in Czech *panovník* (ruler), *vládce* (ruler), *císař* (emperor), *král* (king), *královna* (queen). A few more words can be added to these microfields from peripheral vocabulary.

If we approach the analysis of these microfields from the point of view of historical science, we can show the national-cultural and historical conventionality and feature of each term in each language. However, we are interested in their relationship in two languages from the point of view of ordinary language consciousness. We can say that two ways to put together semantically similar terms in different languages are bilingual dictionaries and examples of translation.

6 Translation equivalents of lexemes of the empire sematic field in Russian and Czech

The last stage of the work is study of interlanguage equivalents. A preliminary assessment was carried out on the base of 2-volume dictionaries edited by L.V. Kopecky (Russian-Czech, Czech-Russian). Vocabulary equivalents can be seen in the left column in Table 1. When analyzing translation dictionaries, we cannot say with what probability one or another equivalent is used.

It is interesting to see which words (and why?) will prevail when translating the same concept. For example, the Czech "říše" in Russian can sound like империя (empire), королевство (kingdom), царство (kingdom), рейх (Reich),

Германия (Germany). The Russian империя (empire) can be translated into Czech as *impérium*, *říše*, *císařství*, *država* and others. The same applies to other terms, too.

Using the terms from our semantic field as an example, we made an attempt to evaluate this using the InterCorp parallel corpus that is a part of ČNK. ČNK programmers developed the Treq tool on the basis of the InterCorp [11], which allows to get all the translations of a given word and statistics on the frequency of translation equivalents that were found in the corpus.

The results obtained (for the lack of place only for translations from Czech to Russian) are shown in Table 1. The left column contains a word in the input language with a translation from the dictionary, the top row contains words of the output language (translations). In cells, quantitative characteristics of translated equivalents are given: the upper number is the number of translations for a given pair of words encountered in the InterCorp corpus, the lower number is the percentage of this translation from all translations of this word (the percentage value is rounded). Rare and erroneous cases are not included, so percent sum is not always 100%. The most frequent translations are highlighted in bold.

Table 1: Translation equivalents for words from the core of the empire semantic field for the Czech language according the InterCorp corpus

	империя	царство	держава	рейх	королевство	монархия	владение	метрополия	государство
říše империя царство	200 51%	56 14%	4 1%	50 13%	37 10%		4 1%		10 2.5%
impérium империя	230 97%								
království королевство		61 20%			216 70%	1 0.3%			
císařství империя	6 86%							1 14%	
mocnářství монархия			1 8%			12 92%			
država владение	1 6%		2 12%				7 44%		1 6%
carství царство		3 75%							1 25%
monarchie монархия						68 97%			

In dictionaries, usually only the main translation is given, and it is usually the most frequent in corpus, but the number of translation equivalents in real

texts is greater (see, for example, the translations for říše) and we see their ratio, too.

7 Conclusion

We see that the use of text corpora and "smart" corpus instruments allows one to identify syntagmatic and paradigmatic connections in an automated mode and create an adequate filling of the term system, in this case it is the semantic field that describes the concept of empire. Lists of words were obtained, greatly expanding available lexicographic manuals.

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

References

1. Blancafort, H. Daille, B., Gornostay, T., Heid, U., Méchoulam, C., Sharoff, S.: TTC: Terminology extraction, translation tools and comparable corpora. In: 14th EURALEX International Congress, pp. 263-268 (2010)
2. Gamallo, P., Gasperin, C., Augustini, A., Lopes, G. P.: Syntactic-Based Methods for Measuring Word Similarity. In: Text, Speech and Dialogue: Fourth International Conference TSD–2001. LNAI 2166, Springer-Verlag, pp. 116–125 (2001)
3. Kilgarriff, A., Rychly, P.: An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Czech Republic, June 2007, pp. 41–44 (2007)
4. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the XIth Euralex International Congress, Lorient: Universite de Bretagne-Sud, pp. 105–116 (2004)
5. Pazienza, M., Pennacchiotti, M., Zanzotto, F.: Terminology extraction: an analysis of linguistic and statistical approaches. In: Knowledge Mining Series: Studies in Fuzziness and Soft Computing, Springer Verlag, Berlin, pp. 255–279 (2005)
6. Pekar, V.: Linguistic Preprocessing for Distributional Classification of Words. In: Proceedings of the COLING–04 Workshop on Enhancing and Using Electronic Dictionaries, Geneva, pp. 15–21 (2004)
7. Rychlý, P.: A lexicographer-friendly association score. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, Brno, pp. 6–9 (2008)
8. Sharoff, S.: Open-source corpora: Using the net to fish for linguistic data. In: International journal of corpus linguistics, John Benjamins Publishing Company, Vol. 11, No. 4, pp. 435–462 (2006)

9. Shaykevich, A.Ya.: The distributive and statistical analysis in semantics [Distributivno-statisticheskij analiz v semantike]. In: Principles and methods of semantic researches [Principy i metody semanticheskih issledovanij], Moscow, pp. 353-378 (1976)
10. Smrž, P., Rychlý, P.: Finding Semantically Related Words in Large Corpora. In: Text, Speech and Dialogue: Fourth International Conference (TSD–2001), LNAI 2166, Springer-Verlag, pp. 108–115 (2001)
11. Škrabal, M., Vavřín, M. The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In: Lexical Computing CZ s. r. o. Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference (2017)
12. Zakharov, V. Ways of automatic identification of words belonging to semantic field. In: Jazykovedný časopis. 2019. Vol. 70. No. 2, pp. 234–243 (2019)