Corpus Methods and Semantic Fields: the Concept of Empire in English, Russian and Czech*

Victor Zakharov
v.zakharov@spbu.ru

Svetlana Pivovarova
clairepivovar@gmail.com

Ekaterina Gvozdyova
ek.gvozd@yandex.ru

Natalya Semenova
nathalja.v.semenova@gmail.com

Saint-Petersburg State University
Saint-Petersburg, Russian Federation

Abstract

The paper presents the results of the ongoing research on creating the semantic field of empire. A semantic field is a collection of words and word combinations covering a certain area of human experience and forming a relatively autonomous microsystem with one or several centres. Relations in such microsystems are also called associations. We draw upon data on syntagmatic collocability and distributional analysis techniques to form a set of lexical units connected by systemic paradigmatic relations of various types and strength. We have developed a methodology to fill a semantic field with lexical units based on morphologically tagged corpora and Sketch Engine built-in tools of statistical distributional analysis. Text material is represented by our corpora in the domain of empire. As part of the study, we have retrieved lists of items filling the semantic field of empire. Our research is focused on the concept of empire in different languages; therefore, we also deal with translation equivalents in language pairs.

Keywords: *semantic field, concept of empire, corpora, distributional statistical analysis, distributional thesaurus*

1 Introduction

Languages tend to reflect everything native speakers think or know about the world. Such beliefs exist not only in the mindset of an individual, but are also characteristic of larger groups (families, nations, societies, etc.) and form what is called a picture of the world. Language and picture of the world are inextricable: a picture of the world can be described by language which in its turn creates a linguistic picture of the world depending on a nation's culture and experience. Researchers seem to be interested in studying linguocultural correlations between similar items within different linguistic pictures of the world.

Linguistic picture of the world determines the most important features of each language, its vocabulary, word formation and syntax [Weisgerber, 1993]. Studies of various linguistic pictures of the world contribute to the research of a native speaker mindset, but a proper study of linguistic picture of the world and linguistic consciousness requires that they should be presented in the form of something tangible and researchable. Associative fields and thesauri might play this role [Ufimtseva, 2014].

Associative fields relate to semantic fields, which can be defined as sets of language units (words and set phrases) with a common meaning component. Groups of lexemes within a semantic field have both linguistic and non-linguistic relations. Such links between elements within a field are of great importance because they reflect dependencies between words in a language. Words are not used separately; all lexemes tend to form a model of the lexical-semantic system.

2 Semantic Fields Description

The term "field" was introduced by G. Ipsen to denote a group of words which forms a unit of meaning [Ipsen, 1924]. The field theory of semantics (defined as the study of words with similar meaning) was developed by W. Humboldt and H. Osthoff [Shur, 1974]. Semantic fields are considered to be semantic groups which consist of lexemes and relations between them, such as synonymy, antonymy, hyponymy, meronymy, etc.

A semantic field can be described by its features. Each field has several semantic properties; elements of semantic fields are joined together by some common semantic property. V. G. Admoni argued that all semantic fields have the following structure: a nucleus, or a core, and a periphery [Admoni, 1973]. All elements of a field (lexemes) are connected to its nucleus. The nucleus must convey a general meaning of its field, therefore only lexemes with the simplest meaning components become a nucleus while meaning components of the nucleus must be present within other field elements. The further an element is from the field core, the more complex are its meaning components. Elements of the field periphery usually consist of several simple meaning components and thus may be connected to the nearby fields. Coherence between elements might differ within a field and could be measured using methods of computational linguistics.

Elements of a field have systemic syntagmatic (textual), paradigmatic (associative) and epigmatic (lexically derived) relations [Novikov, 2011]. In order to place an element into a lexical system of a language and form a semantic field, all three types of relations should be determined. However, many studies prove that two main types (textual and associative relations) are enough for this purpose. In order to form a semantic field, we need to implement the following steps: choose a list of lexemes that will be constituents of the field and determine relations between them.

There are various ways of making a list of candidates. The easiest one seems to be a monotonous task of selecting items from all types of dictionaries (glossaries, semantic dictionaries, dictionaries of collocations, etc.) representing syntagmatics as well as paradigmatics. However, this method is somewhat time-consuming and subjective. Besides, it does not guarantee that the list of elements will be comprehensive. Alternatively, we can use an association experiment, which might prove effective.

We believe that methods and approaches of computational linguistics provide the best way to form a semantic field. All lexemes are represented in texts, which can be used for extracting necessary items. First, we should find or create a large set of texts (corpus). The next step is to calculate coherence using statistical measures, which would provide us with enough data for the extraction of relevant texts and lexemes. This methodology allows us to make semantic field construction automatic.

3 Corpus-Based Semantic Research

A corpus is a big collection of annotated and structured language data presented in electronic form and designed for linguistic research [Zakharov, Bogdanova, 2019]. The number of corpus-based studies is rapidly increasing due to linguistic annotation within corpora and a large amount of language data available for research purposes. Moreover, corpora can be easily built for a specific purpose, which is appealing to many scholars.

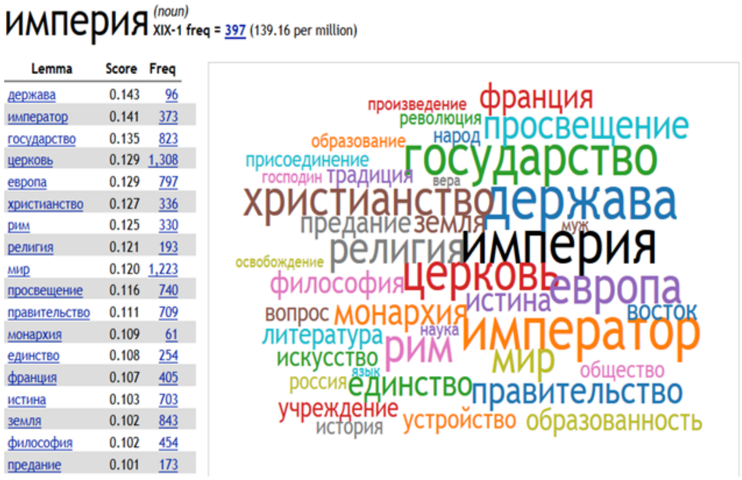
As corpus linguistics developed and new larger semantically annotated corpora emerged semantic corpus-based studies slowly evolved as well. A. Kilgariff, arguing in one of his works that semantics should be studied using corpora, stated that a word sense corresponds to a cluster of texts for this word [Kilgarriff, 1997]. Corpora are widely used in quantitative semantics: word sense disambiguation is often based on frequencies of words in context extracted from corpora; semantic space (semantic field) construction relies on calculating semantic distances and determining relations between words within a corpus.

Statistical scores can describe elements that tend to co-occur in corpora. Such co-occurring elements which form syntagmatic relations in texts can be called collocations. Coherence within collocations is calculated using various statistical measures implemented in modern corpus systems: the most widely known are MI (Mutual Information), log-likelihood, t-score, minimum sensitivity, logDice.

Paradigmatic relations describe semantically similar elements: semantic similarity is based on determining the agreement of words' lexical neighbourhood [Ruge, 1992]. Therefore, in order to form paradigmatic relations, we need to determine co-occurrence vectors for each word and compute their similarity. By comparing vectors, we compare distances between them; the smaller the distance, the closer the words are in meaning to each other.

Paradigmatic relations can also be formed based on syntagmatic ones. Two words are considered paradigmatically related if each of them is systematically connected to a third text element.

One more way to compute paradigmatic relations is using statistical distributional analysis. The analysis includes a set of algorithms for language description based solely on the distribution of elements within texts [Yartseva, 1976] and provides a clear indication of functional and semantic relations between field elements. Algorithms of statistical distributional analysis are implemented in many corpora systems (e. g. Sketch Engine (Fig.1)).



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

Figure 1: A distributive thesaurus for the word *империя* (subcorpus for the first half of the XIXth century)

4 Formation of Semantic Fields Using Corpus Managers

4.1 Sketch Engine

A corpus manager is a corpus analysis tool for data search and extraction. The most widely known are Sketch Engine, NoSketch Engine, AntConc, MonoConc, etc. Corpus managers allow users to perform the statistical analysis necessary for term extraction (these terms form a semantic field) and identification of semantic relations between the terms. For detailed description of corpus managers' tools, refer to documentation [Sketch Engine documentation], [AntConc Software].

In this research, we used Sketch Engine – a corpus analysis tool which allows users to create corpora by uploading files or downloading content from the web using WebBootCaT technology. Sketch Engine built-in tools deal with lexical-semantic patterns (word sketch), statistical analysis, distributional thesauri, clustering, keyword extraction, etc. The goal of our study – to form a semantic field – implies identifying syntagmatic and paradigmatic relations between corpus items, for which purpose we use the following Sketch Engine tools: "Collocations" for syntagmatic relations (using association measures) and "Thesaurus" for paradigmatic ones.

The "Thesaurus" tool in Sketch Engine retrieves words having a distribution similar to that of the given word, which, as a rule, is due to their semantic proximity, i.e., in fact, this tool forms a uniterm semantic field. Word distribution similarity is calculated statistically by using the association measure logDice [Rychlý, 2008] and lexical-syntactic patterns [Kilgarriff, Rychlý, 2007]. At the next step, the most typical word co-occurrences identified using the "Collocations" tool are added to the semantic field.

4.2 Material and Research Methodology

As this study deals with three languages (Russian, English and Czech) and their linguistic pictures of the world, we needed a concept equally significant for each of these linguocultures. The concept of empire seems to be the right one for our purposes. We have undertaken the following tasks: an automatic formation and comparative analysis of the semantic field of empire for three languages, as well as compiling a thesaurus for these semantic fields specifying quantitative features and examples from the corpora.

Our research was conducted on English, Russian and Czech text corpora in the domain of empire (the Russian corpus was created earlier in association with M. Khohlova). Russian and English text corpora consist of four subcorpora, each covering different periods: the XVIIIth century, the first half of the XIXth century, the second half of the XIXth century, and the XXth century.

Our methodology relies on the "Thesaurus" tool within the SketchEngine system for extracting a ranked list for each subcorpora and retrieving a list of lexemes that are present in all thesauri. The detailed description of our methodology for semantic field formation could be found in [Zakharov, 2018]. After following all the steps, we have retrieved summaries of distributional thesauri for Russian and English (Table 1, 2).

Table 1: Summary of the distributional thesaurus for the word *империя* (a fragment)

Subcorpus	**Rank**	**Lemma**	**Score**	**Freq**	**Stability coefficient**	**Average rank**	**Norm. rank**
XIX-2	1.	австрия (Austria)	0,216	1014	1		
XIX-2	36.	англия (England)	0,131	1055	2	29	87
XVIII	22.	англия (England)	0,095	148	2		
XIX-2	19.	армия (army)	0,149	478	1		
XIX-1	37.	господин (master)	0,085	363	1		
XIX-2	24.	государственность (statehood)	0,143	201	2	19	57
XX	14.	государственность (statehood)	0,141	143	2		

Table 2: Summary of the distributional thesaurus for the word *empire* (a fragment)

Subcorpus	**Rank**	**Lemma**	**Score**	**Freq**	**Stability coefficient**	**Average rank**	**Norm. rank**
XVIII	5.	church	0.227	2932	3	28,3	56,7
XIX-1	41.	church	0.114	744	3		
XIX-2	39.	church	0.185	3170	3		
XVIII	29.	dominion	0.158	487	1		
XVIII	30.	England	0.158	5499	4	31,25	31,25
XIX-1	38.	England	0.117	2253	4		
XIX-2	33.	England	0.196	5211	4		
XX	24.	England	0.165	3659	4		
XVIII	18.	Europe	0.172	1615	4	27,5	27,5
XIX-1	26.	Europe	0.126	1516	4		
XIX-2	17.	Europe	0.219	4361	4		

Due to the lack of Czech historical corpora, we failed to conduct a diachronic study for the Czech language. The research was carried out using corpora of modern texts: syn v7 (all synchronic written corpora of the Czech National Corpus) and csTenTen 2017 (Sketch Engine Czech web corpus).

At the first stage, various lexicographic sources were used to describe the concept of empire in terms of keywords. Having analysed definitions of empire in various Czech dictionaries, we identified the main meanings and corresponding semantic attributes of the concept of empire:

1. monarchy, headed by the emperor;

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

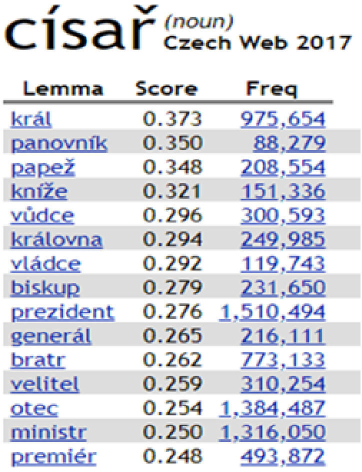
3. metaphoric meanings derived from one of the first two (e.g. a large enterprise, parts of the natural world, etc.).

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

We have carried out a definitional analysis of explanatory dictionaries and dictionaries of synonymy and identified elementary units of a meaningful plan. In doing so, we sought to make these terms monosemic.

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 these identifiers, 10 distributional thesauri were built in Sketch Engine based on csTenTen 2017 corpora (Fig. 2). In order to avoid retrieving nonrelevant vocabulary, the volume of the distributional thesaurus was limited to 15 items.



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

Figure 2: The distributional thesaurus (semantic field) for the word *říše*

At the next stage, all 10 thesauri were put together into one dataset. Furthermore, for each term, the average score was calculated. We have made the following empirical assumption: 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 a value of the score less than N form its periphery. Both in the centre and the periphery area the lexemes can be sorted according to their score.

Further, for each element of the field core, the most typical bigram collocations were identified using ČNK syn v7 corpus and the "Collocations" tool. Bigrams were sorted by MI.log_f – one of the most effective association measures.

4.3 Results

4.3.1 Semantic field of empire in Russian

Summary distributional thesaurus for the semantic field of empire in Russian included 160 entries with 79 unique words occurring once. 33 different words occur in 2 or more minithesauri. We call these 33 words the nucleus (or the core) of the semantic field. Their distribution in the subcorpora is as follows: XVIII: 8, XIX-1: 24, XIX-2: 26, XX: 23. The full alphabetical list of the core lexemes includes:

Англия (England), государственность (statehood), государство (state), держава (derzhava), 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 (policy), правительство (government), просвещение (education), революция (revolution), религия (religion), Рим (Rome), Россия (Russia), союз (union), страна (country), традиция (tradition), учреждение (institution), философия (philosophy), Франция (France), христианство (Christianity), царство (kingdom), церковь (church)

4.3.2 Semantic field of empire in English

In this section, we present the results of semantic field formation for English. The full list of lexemes includes 113 items, 46 being unique.

The alphabetical list of terms for the English semantic field includes the following lexemes: *affair, Africa, ally, America, arm, army, assembly, Austria, authority, body, Britain, camp, Canada, capital, church, city, colony, commerce, community, conquest, Constantinople, constitution, corps, country, court, crown, dominion, dominions, dynasty, East, emperor, Empire, enemy, England, Europe, family, fleet, force, fortune, France, freedom, frontier, garrison, Gaul, Germany, government, group, happiness, history, house, India, industry, interest, Ireland, island, Italy, king, kingdom, land, language, law, liberty, life, line, man, monarch, monarchy,* *movement, name, nation, order, part, party, people, person, policy, population, position, possession, power, prince, property, province, Prussia, question, race, religion, republic, Republic, revolution, right, Rome, rule, Russia, service, settlement, ship, society, sovereign, Spain, state, States, subject, system, territory, throne, time, town, trade, troop, war, work, world.*

Lexemes that occur in 2 or more thesauri form the core of the semantic field of empire. In English, the core includes 67 lexemes, which are distributed the following way: XVIII: 42, XIX-1: 12, XIX-2: 7, XX: 6. Here is the full list of the core lexemes in alphabetical order:

ally, army, Austria, authority, Britain, capital, church, city, colony, conquest, constitution, country, court, crown, dominions, emperor, Empire, enemy, England, Europe, family, fleet, force, France, Germany, government, history, interest, island, Italy, king, kingdom, land, law, liberty, life, line, man, monarch, monarchy, name, nation, part, party, people, position, power, prince, province, race, religion, republic, Rome, Russia, society, sovereign, Spain, state, system, territory, throne, town, trade, troop, war, work, world

4.3.3 Semantic field of empire in 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 semantic field of empire for the Czech language. Interestingly, 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 semantic field of empire for the Czech language is as follows (in alphabetical 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.

A list of bigram collocations – candidates for the semantic field of empire – can also be formed, but this question is beyond the scope of this paper.

5 Translation Equivalents of Lexemes of the Semantic Field of Empire in Language Pairs

It is also interesting for us to find out how one and the same concept is translated in different languages, namely, English, Russian and Czech. This task can be accomplished by means of parallel corpora containing texts in different languages but in the same cultural and historical paradigm. Thus, such an instrument can be used to interpret our results from historical and cultural perspectives.

5.1 Russian and English equivalents

In this study we have set a goal to identify translation equivalents for the elements of the semantic field of empire, English being a source language, while Russian is a target language.

With that end in view, we have compiled a test English-Russian parallel corpus of about 994,000 words. The corpus currently comprises XVIIIth century historical texts. LFAligner [LFAligner] was used for sentence alignment, while word alignment was done by GIZA++ word alignment tool developed within Moses statistical machine translation system [Moses].

The Sketch Engine tool "Thesaurus" was used to create an English distributional thesaurus for the word empire (Fig. 3) comprising 100 lexemes.

Each lexeme from the thesaurus was matched with a list of possible Russian translation equivalents identified by GIZA++ word alignment tool. Table 3 shows Russian translation equivalents for the first ten elements of the English distributional thesaurus for the word *empire* and their percentage of the total number of Russian translation equivalents for the element (percentage values are rounded off). Erroneous translation equivalents resulted from incorrect word alignment were discarded.

These results could be further interpreted using the "Parallel concordance" tool within Sketch Engine. This instrument allows us to see the use of translation equivalents in context (Fig. 4).

Comparing the structure of the semantic field of empire in Russian and Czech, we have made the following observations. If we temporarily exclude from consideration lexemes that mean roughly the same in Russian

Table 3: Summary of the distributional thesaurus for the word *empire* (a fragment)

English distributional thesaurus lexeme	**Lexeme frequency in the test corpus**	**Russian translation equivalent**	**Percentage of the equivalent**
monarchy	131	монархия (‘monarchy’)	75%
		империя (‘empire’)	10%
		владычество (‘dominion’)	3%
		держава (‘power’)	1%
		династия (‘dynasty’)	1%
church	553	церковь (‘church’)	80%
		храм (‘temple’)	2%
province	399	провинция (‘province’)	88%
		владение (‘domain’)	3%
		местность (‘region’)	2%
		страна (‘country’)	1%
		сфера (‘realm’)	1%
city	864	город (‘city’)	87%
		столица (‘capital’)	7%
		горожане (‘city-dwellers’)	1%
monarch	188	монарх (‘monarch’)	92%
kingdom	275	королевство (‘kingdom’)	31%
		царство (‘realm’)	21%
		владение (‘domain’)	19%
		государство (‘state’)	10%
		страна (‘country’)	6%
		владычество (‘dominion’)	3%
East	307	Восток (‘East’)	48%
prince	688	монарх (‘monarch’)	49%
		князь (‘duke/prince’)	31%
		принц (‘prince’)	11%
		владетель (‘owner’)	4%
		государь (‘sovereign’)	2%
		император (‘emperor’)	1%
power	688	власть (‘power’)	49%
		могущество (‘might’)	27%
		права (‘rights’)	4%
		полномочия (‘authority’)	3%
		держава (‘power’)	3%
		влияние (‘influence’	3%
world	265	мир (‘world’)	72%
		свет (‘world/society’)	1%

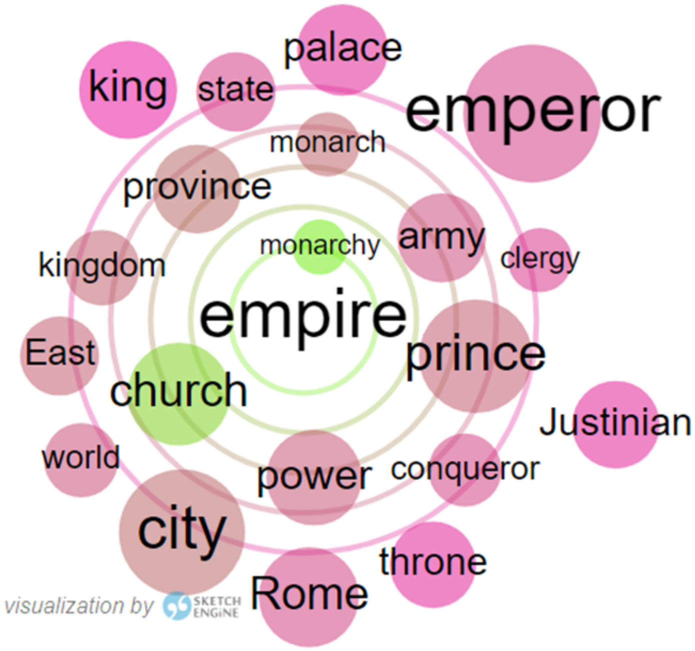


Figure 3: Part of the distributional thesaurus for the word *empire* based on the test corpus

and Czech and lexemes that are present only in one of the semantic fields (*государственность* (statehood), *папа* (pope), *князь* (prince), *цивилизация* (civilization), *generál* (general), *velitel* (commander)), we can see that the remaining lexemes are related to the two microfields.

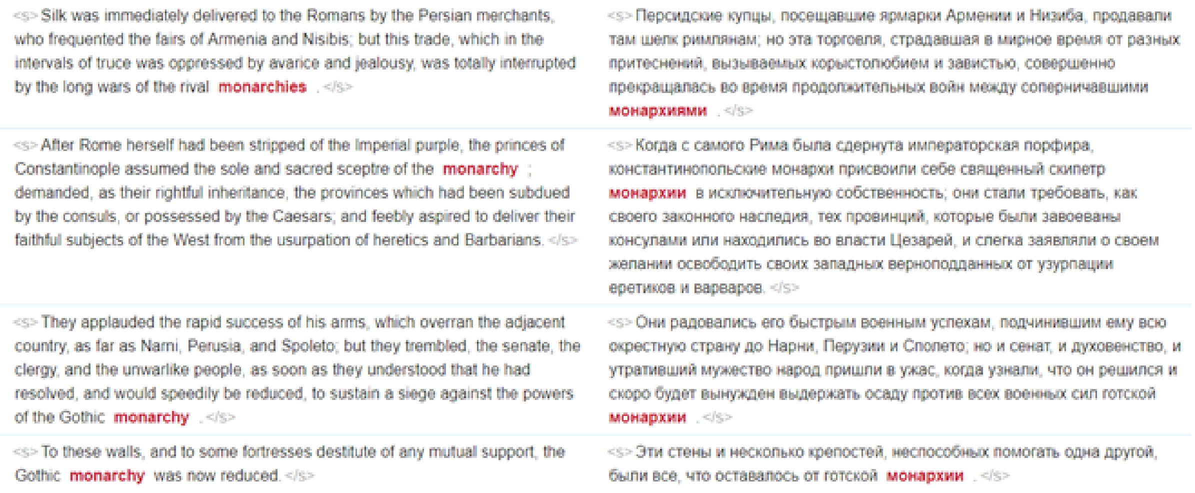
The first microfield contains different names for the concept of empire: in Russian, they are *империя* (empire), *царство* (kingdom), *держава* (power), partly *монархия* (monarchy); in Czech, *impérium* (empire), *říše* (empire), *království* (kingdom), *císařství* (empire), *mocnářství* (monarchy), 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).

The last stage of our research deals with interlanguage equivalents. A preliminary assessment was carried out based on 2-volume dictionaries edited by L.V. Kopecky (Russian-Czech, Czech-Russian). Vocabulary equivalents can be seen in the left column in Table 4. When analysing 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 be translated as *империя* (empire), *королевство* (kingdom), *царство* (kingdom), *рейх* (Reich), *Германия* (Germany). The Russian *империя* (empire) can be translated into Czech as *impérium, říše, císařství, država*, etc. The same applies to other terms, too.

Using the terms from our semantic field as an example, we attempted to evaluate them using the InterCorp parallel corpus that is a part of ČNK [Čermák, Rosen, 2012]. ČNK programmers developed the Treq tool based on the InterCorp [Škrabal, Vavřín, 2017], which retrieves all the translations of a given word and statistics on the frequency of translation equivalents that were found in the corpus.

The results obtained (translations from Czech to Russian) are shown in Table 4. 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). The cells show quantitative characteristics of the translated equivalents: 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 the percentage sum is not always 100%. The most frequent translations are highlighted in bold.



<s> Silk was immediately delivered to the Romans by the Persian merchants, who frequented the fairs of Armenia and Nisibis; but this trade, which in the intervals of truce was oppressed by avarice and jealousy, was totally interrupted by the long wars of the rival **monarchies** . </s>	<s> Персидские купцы, посещавшие ярмарки Армении и Низиба, продавали там шелк римлянам; но эта торговля, страдавшая в мирное время от разных притеснений, вызываемых корыстолюбием и завистью, совершенно прекращалась во время продолжительных войн между соперничавшими **монархиями** . </s>
<s> After Rome herself had been stripped of the Imperial purple, the princes of Constantinople assumed the sole and sacred sceptre of the **monarchy** ; demanded, as their rightful inheritance, the provinces which had been subdued by the consuls, or possessed by the Caesars; and feebly aspired to deliver their faithful subjects of the West from the usurpation of heretics and Barbarians. </s>	<s> Когда с самого Рима была сдернута императорская порфира, константинопольские монархи присвоили себе священный скипетр **монархии** в исключительную собственность; они стали требовать, как своего законного наследия, тех провинций, которые были завоеваны консулами или находились во власти Цезарей, и слегка заявляли о своем желании освободить своих западных верноподданных от узурпации еретиков и варваров. </s>
<s> They applauded the rapid success of his arms, which overran the adjacent country, as far as Narni, Perusia, and Spoleto; but they trembled, the senate, the clergy, and the unwarlike people, as soon as they understood that he had resolved, and would speedily be reduced, to sustain a siege against the powers of the Gothic **monarchy** . </s>	<s> Они радовались его быстрым военным успехам, подчинившим ему всю окрестную страну до Нарни, Перузии и Сполето; но и сенат, и духовенство, и утративший мужество народ пришли в ужас, когда узнали, что он решился и скоро будет вынужден выдержать осаду против всех военных сил готской **монархии** . </s>
<s> To these walls, and to some fortresses destitute of any mutual support, the Gothic **monarchy** was now reduced. </s>	<s> Эти стены и несколько крепостей, неспособных помогать одна другой, были все, что оставалось от готской **монархии** . </s>

Figure 4: “Parallel concordance” tool: Russian translation equivalents for the English lexeme ‘monarchy’ in context

Table 4: Translation equivalents for the words from the core of the semantic field of empire for the Czech language according to 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%			

6 Conclusion

In dictionaries, only the main translation is usually given, and as a rule it is most frequent in the 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.

We can see that text corpora and “smart” corpus tools can be used to identify syntagmatic and paradigmatic relations in an automated mode and fill the term system properly. In our research, we attempted to form the semantic field of empire, and lists of words retrieved expand significantly available lexicographic resources.

Lexemes were extracted using Sketch Engine; the lexemes form and adequately describe the semantic field of empire and could successfully complement data from other sources. A list of empire-related lexemes from WordNet thesaurus, for instance, contains a few items from our semantic fields and therefore could be expanded.

A comparative analysis of the semantic fields of empire for Russian, English and Czech reveals some interesting patterns. English semantic field of empire includes more lexemes that are somehow related to military action (army, conquest, enemy, fleet, force, troop, war etc.), power and statehood (authority, capital, church, city, constitution, court, crown, law, liberty, etc.).

Another peculiarity concerns distributions of lexemes within the core of semantic fields. Lexemes tend to spread equally between the XIXth and XXth centuries in Russian. However, in English, the core lexemes are mainly found in the XVIIIth subcorpus, which might have resulted from an unbalanced corpus of English texts.

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 the automation of semantic processing.

Acknowledgements

This work was financially supported by the Russian Foundation for Basic Research, Project No. 18-012-00474 “Semantic field of empire in Russian, English and Czech”.

References

[Weisgerber, 1993] Weisgerber L. (1993) Muttersprache und Geistesbildung. Translation from German by O. A. Radchenko. Moscow. 1993. 170 p.

[Ufimtseva, 2014] Ufimtseva N. V. (2014) The Associative Dictionary as a Model of the Linguistic Picture of the World. Procedia - Social and Behavioral Sciences 154, 2014. Pp. 36–43

[Ipsen, 1924] Ipsen G. (1924) The Ancient Orient and Indogermans Feast Scipts for W. Streitburg. Heidelberg, 1924. Pp. 30-45

[Shur, 1974] Shur G. S. (1974) Field theories in linguistics: a monography. Moscow, Nauka, 1974. 254 p. (In Rus.) = Shur G. S. Teorii polya v lingvistike: monografiya. M.: Nauka, 1974. 254 s.

[Admoni, 1973] Admoni V. G. (1973) Syntax of modern German language: A system of relations and build system. Leningrad, Nauka, 1973. 366 p. (In Rus.) = Sintaksis sovremennogo nemetskogo iazyka. Sistema otnoshenii i sistema postroeniia. L.: Nauka, 1973. 366 s.

[Novikov, 2011] Novikov A. L. (2011) An Essay on Semantic Field. RUDN journal of language studies, semiotics and semantics. Moscow, 2011. Pp. 7-17. (In Rus.) = Eskiz semanticheskogo polya. Vestnik rossiiskogo universiteta druzhby narodov. Seriya: teoriya yazyka. Semiotika. Semantika. ., 2011. 7-17

[Zakharov, Bogdanova, 2019] Zakharov V. P., Bogdanova S. U. (2019) Corpus Linguistics: Textbook for Students. 3-rd edition, revised and extended. Saint-Petersburg, 2019. 230 p. (In Rus.) = Korpusnaya lingvistika: Uchebnik dlya studentov napravlenya ≪Lingvistika≫ i ≪Pedagogicheskoe obrazovanie≫. 3- izd., pererab. i dopoln. Spb., 2019. 230 s.

[Kilgarriff, 1997] Kilgarriff A. (1997) I don’t believe in word senses. Computers and the Humanities 31 (2). 1997. Pp. 91–113

[Ruge, 1992] Ruge, G. (1992) Experiments on Linguistically Based Term Associations. Information Processing Management 28(3), 1992. Pp. 317–332

[Yartseva, 1976] Yartseva V. N. (1976) Principles and Methods of Semantic Research. Ed. by V. N. Yartseva. Moscow, Nauka, 1976. 380 p. (In Rus.) = Yartseva V. N. (pod red.) Printsipy i metody semanticheskikh issledovanii. M.: Nauka, 1976. 380 s.

[Sketch Engine documentation] Sketch Engine documentation. (2019) Available at https://www.sketchengine.eu/documentation/

[AntConc Software] AntConc Software. (2019) Available at https://www.laurenceanthony.net/software/antconc/

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

[Kilgarriff, Rychlý, 2007] Kilgarriff A., 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 on Interactive Poster and Demonstration Sessions. Czech Republic, June 2007. Pp. 41–44

[Zakharov, 2018] Zakharov V. P. (2018) The Distributive and Statistical Analysis as a Tool to Automate the Formation of Semantic Fields (on the Example of the Linguocultural Concept of ”Empire”). Proc. of Comp. Models in Language and Speech Workshop CMLS, Vol. 2, 2018. Pp. 163-180. (In Rus.) = Zakharov V. P. Distributivno-statistichesky analyz kak instrument avtomatizacii formirovanya semanticheskyh poley (na primere polya ≪imperya≫). V XV Mezhd. konf. po komp. i kogn. lingv. TEL 2018. Sb. trudov, Tom 2, s. 163-180

[LFAligner] LFAligner. Available at https://sourceforge.net/p/aligner/wiki/Home/

[Moses] Moses. (2019) Available at http://www.statmt.org/moses/index.php?n=Main.HomePage

[Čermák, Rosen, 2012] Čermák F., Rosen A. (2012) The case of InterCorp, a multilingual parallel corpus. In: International Journal of Corpus Linguistics 17(3), 2012. Pp. 411–427.

[Škrabal, Vavřín, 2017] Škrabal M., Vavřín M. (2017) 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. Pp. 124-137