SLAVIC PHRASEOLOGY: A VIEW THROUGH CORPORA

VICTOR ZAKHAROV

Saint-Petersburg State University, Russia

Abstract: The study of word collocability is one of the main tasks of linguistics. The combinatory ability of language units, collocability, is one of the linguistic syntagmatic laws. This phenomenon is the main object of the phraseology and lexicography. The article deals with set phrases of different types in Russian, Czech and Slovak from the point of view of their quantitative evaluation. Corpus linguistics understand set phrases as statistically determined unities. This approach is the basic point of different automatic ways to extract idioms and collocations. The paper describes experiments which show how text corpora and corpus methods and tools can be used to expand the entries in existing dictionaries and how set phrases could be evaluated quantitatively. It is shown and maintained that corpus linguistics methods and tools allow to create dictionaries of new type which have to include a larger amount of set phrases and collocations than before.

Keywords: Slavic phraseology, phraseological units, set phrases, idioms, collocations, corpus, lexicography

1   INTRODUCTION

One of the popular topic in the science of language are set phrases. They are studied under different sections of linguistics and from various standpoints. The classical name for a set phrase in the linguistic terminology is a phraseological unit. Various scientists interpret this notion and its properties in various ways, and there are many different classifications of set phrases. However, when comparing them, we will see that the list of these properties and the very classification are often alike and have a lot in common.

The commonly adopted interpretation of phraseological units defines them as set, reproducible, expressive word combinations (*беречь пуще глаза* (literally, protect more than eyes); *драть по три шкуры* (literally, to skin three skins); *sedí na uších (literally, sit on ears), jako by do moře padl* (as it fell into the sea)). However, we can often find in translation dictionaries such units as, *и ничего больше* (and nothing else), *to je to* (that's just it)*, a tečka* (and period)*,* where the main property is reproducibility. And this is right, since such combinations are often stable, and can be compared to words by their frequency. In this article, we will consider phraseological units as "stable word complexes" [4].

However, despite the close attention of linguists to the phraseology, we can state that the computer-based methods of research have hardly been used to study phraseology. The dictionary description of phraseological units shall be elaborated. The phraseological reserve of a language is scattered in various lexicographic

publications, and no dictionary can be regarded as covering the phraseological lexicon in full, taking into account also the fact that it always expands.

Today, we can improve the situation by means of corpora. Due to the availability of large text corpora, including those that cover a long period of time, and the software tools that allow to estimate the compatibility quantitatively, all pre-requisite conditions have appeared for the creation of a large combinatory dictionary, obviously, in electronic form, based on corpora, and having quantitative parametrization inside.

It should be noted that the methodology of a larger understanding of phraseology has formed in linguistics, and the boundaries of phraseology have been significantly enlarged (or blurred) due to new approaches that have the notion of "statistical stability" in common. I.A. Melčuk spoke about it as early as in 1960. "The stability of a combination relative to this element is measured by the probability of this element forecasting the combined occurrence of the other elements of the combination (in a certain order relative to the forecasting element)" [7, p. 73].

In corpus linguistics, frequency characteristics and structural and syntactic models form the basis of the methods of calculating the strength of syntagmatic association between the word combination elements. Based on them, the association measure score, or, in other words, the uniqueness of this combination, is calculated.

As it is known, language is a dynamic system, which shall be reflected in dictionaries and grammars. However, maybE this chronological aspect has been less studied in terms of phraseological units and other stable combinations. One of the reasons is the absence of large historical (diachronic) corpora till our times.

The objective of this research is to show how phraseological and combinatory dictionaries can be improved using corpus-based methods. The basis of our approach are the analytic tools for phraseological units and stock-taking of the language material. Furthermore, we show how similar phraseological units correlate with each other in related Slavic languages. We also consider the issue of translation of phraseological units from one language to another. The Slavic phraseology is represented by Russian, Czech, and Slovak.

2   RESEARCH MATERIAL AND TOOLS

A family of comparable corpora Aranea of the Comenius University in Bratislava (`http://unesco.uniba.sk/`) and the Russian corpus of Google books Ngram Viewer (`https://books.google.com/ngrams`) have been used as the material and the tool. The corpora of the Aranea family [2] operate under the NoSketch Engine corpus manager. They are represented by the Maximum type in the research, and they have the following volume: Russian corpus is made up of 13.7 billion tokens, the Czech one – 5.17 billion tokens, and Slovak – 2.68 billion tokens. Also, some data were retrieved in [6], [10] and [13]. Google books Ngram Viewer system [8, p. 14] was used for the diachronic research. The system allows to plot graphs of word occurrences and collocations for a certain period of time. It also allows to select the most frequent collocations with such word form, both from the right and from the left, using wild cards. There is also the possibility of setting the part of speech of another components of the collocation.

Our chosen phraseological units for the research are the set phrases of two types: classical phraseological units (idioms) and syntactic idioms (idiomatic constructions).

3   REPRODUCIBILITY AND VARIABILITY OF PHRASEOLOGICAL UNITS

3.1   Idioms

Idioms are not only strongly reproducible, but they are, at the same time, very variable, and it is an important task of the phraseology to present this variability in the dictionaries. Many idioms and set phrases have lexical-syntactic variants when either the lexical meaning within a certain structural formula or the formula changes. There are a lot of examples of this phenomenon: *беречь (хранить) как зеницу ока* (keep as the apple of one's eye)*; беречь пуще глаза* [3]; *dát si na někoho/před někým majzla* (pay attention to somebody); *dát/uložit něco k ledu/z ruky* (literally, let smth to ice/from hand) [9]. Let us consider some of them.

Example 1:
Rus. *на сердце (на душе) кошки скребут* (literally, cats scratch the heart/soul) [5];
Czech. *je mi těžko (úzko) u srdce (v duši)* (be sick at heart/soul) [11];
Slov. *je mi ťažko (úzko) okolo srdca (na duši)* (near the heart/at soul) [12].

	на … кошки скребут	кошки скребут на ….	Total
на сердце	27	5	32
на душе	383	85	**468**
Total	**410**	90	500

Tab. 1. The frequency of occurrences of combinations with the phrase «кошки скребут» in the Araneum Russicum Maximum corpus

	je mi těžko …	je mi úzko …	Total
u srdce	17	3	**20**
v duši	0	1	1
Total	**17**	4	21

Tab. 2. The frequency of the occurrences of the combinations with the phrase «je mi těžko (úzko)» in the Araneum Bohemicum Maximum corpus

	je mi ťažko …	je mi úzko …	Total
okolo srdce	0	0	0
na duši	21	10	**31**
Total	**21**	10	31

Tab. 3. The frequency of the occurrences of the combinations with the phrase «je mi ťažko (úzko)» in the Araneum Slovacum Maximum corpus

What can be derived from this example? In Russian, this idiom is more commonly used than in the Czech and Slovak languages, and «на душе» (soul) is

used with «кошки скребут» more often than «на сердце» (heart), i.e. 468 occurrences against 32 (Table 1). It is an interesting fact that an object noun usually precedes the verb (410 against 90). As contrasted with Russian, Czech prefers the "heart" variant (Table 2). It is possible that this is the influence of the German language. In the pair *těžko/úzko,* the variant *těžko* is used most often. The "object" (noun) virtually always stands after the verb. The same goes for the Slovak language: *ťažko* – 21 occurrences, and *úzko* – 10 occurrences, while the dictionary variant *okolo srdce* does not occur in the corpus at all.

Meanwhile, the Czech corpus returns word combinations with preposition "na" (on), which is not present in the dictionary [9]: *je mi těžko (úzko) na srdci (na duši)* (Table 4). And the combination *na duši* is far more often.

	je mi těžko …	je mi úzko …	Total
na srdci	4	0	4
na duši	15	6	**21**
Total	**19**	6	25

Tab. 4. Frequency of the occurrence of combinations with the phrase «je mi těžko (úzko) na ...» in the Araneum Bohemicum Maximum corpus

Similarly, the Slovak corpus adds to the dictionary data the following combinations: *na srdci* and *v duši* (Table 5).

	je mi ťažko …	je mi úzko …	Total
na srdci	29	1	**30**
v duši	1	2	3
Total	**30**	3	33

Tab. 5. Frequency of the occurrence of additional combinations with the phrase «je mi ťažko (úzko) ...» in the Araneum Slovacum Maximum corpus

For Russian, the dictionary [4] returns also a synonymic expression *камень лежит на сердце (на душе)* (literally, a stone lies on the hear/soul). Once again, the corpus analysis makes another correction: this expression occurs more often in the form ***камнем*** *что-то лежит на сердце (на душе)* (65 occurrences versus 15).

Example 2. Let us consider the phraseological units that are rather literal:
Rus. *камень преткновения* (literally, a stone of obstacle) [5];
Czech. *kámen úrazu* (literally, a stone of injury) [11];
Slov. *kameň úrazu* Id. [12];
Rus. *нашла коса на камень* (literally, a meak stumbled on a stone) [5];
Czech. *přišla/trefila/padla kosa na kámen* (came/dropped …) [11];
Slov. *padla kosa na kameň* Id. [12].

Idiom	frequency	ipm	variants
Rus. *камень преткновения*	16710	1,20	
Czech. *kámen úrazu*	11003	**2,10**	
Slov. *kameň úrazu*	5425	**2,00**	

Rus. *коса на камень*	1171	0,10	***найти – 945*** *пойти – 13* *налететь – 3*
Czech. *kosa na kámen*	649	**0,13**	*přijít – 9* *trefit – 5* ***padnout – 335*** *narazit – 201*
Slov. *kosa na kameň*	406	**0,15**	*prísť – 5* *trafiť – 24* *padnúť – 149* ***naraziť – 180***

Tab. 6. Frequency of occurrence of the combinations «камень преткновения» and «нашла коса на камень» and their equivalents in Czech and Slovak in three corpora

From Table 6 it is seen that in Czech and Slovak these expressions are used more often. While in the Russian corpus this expression is almost always found with the verb «найти» (*нашла*) (find), in the Czech and Slovak corpora, alongside with the word "padnout" ("padnúť"), the verb "narazit" ("narazit") (injure oneself, stumble upon smth) is used frequently, too, that is not given in dictionaries.

Example 3:
Rus. *делать (сделать) из мухи слона* (literally, make an elephant out of a fly) [5];
Czech. *dělat (udělat) z komára velblouda* (make a camel out of a mosquito) [11];
Slov. *robiť (urobiť) z komára somára* (make a donkey out of a mosquito) [12].

Idiom	frequency	ipm	variants
Rus. *из мухи **слона***	2343	**0.20**	*толстяка 1* *барсука 1*
делать/сделать	1098		
раздувать/раздуть	791		
Czech. *z komára **velblouda***	634	0.10	*vola (bull) 14* *slona 5 (3 of them appeared in texts in Slovak!)*
dělat/udělat	602		
Slov. *z komára **somára***	265	0.10	***slona 56*** *velblouda 8 (4 of them appeared in texts in Czech!)* *vola - 4*
robiť/urobiť	240		

Tab. 7. Frequency of occurrence of the word combination «делать из мухи слона» and its equivalents in Czech and Slovak in three corpora

What is the essence of Table 7? First of all, different frequency of occurrences: it is twice more often in Russia; secondly, slightly different variations and, what is more interesting, the Russian corpus returns a lot of combinations with the verb «раздувать»

(blow), which is not given in dictionaries. In the Slovak language, it is often an elephant that is made out of a mosquito, and the dictionary has no evidence of that.

3.2 Idiomatic Constructions

Set syntactic constructions that have variable lexical elements are singled out as a separate type of phraseological units. They are, so to say, syntactic patterns that are filled depending on the context, communicative aim of the author, and – as we will show – on the language. They include such expression as «Х как Х» (X as X), «тоже мне Х» (some X), «всем Х-ам Х» (X to all Xs), «Х Х-ов» (X of Xs), «Х – он и в Африке Х» (X is the same in Africa), etc. Idiomatic constructions are studied within the framework of the construction grammar developed by Ch. Fillmore, A. Goldberg, and others. It is considered that "blank spaces" can be filled with anything at all. However, there are semantic limitations for the filling, and there is the language usage that changes in different languages. Let us consider several cases.

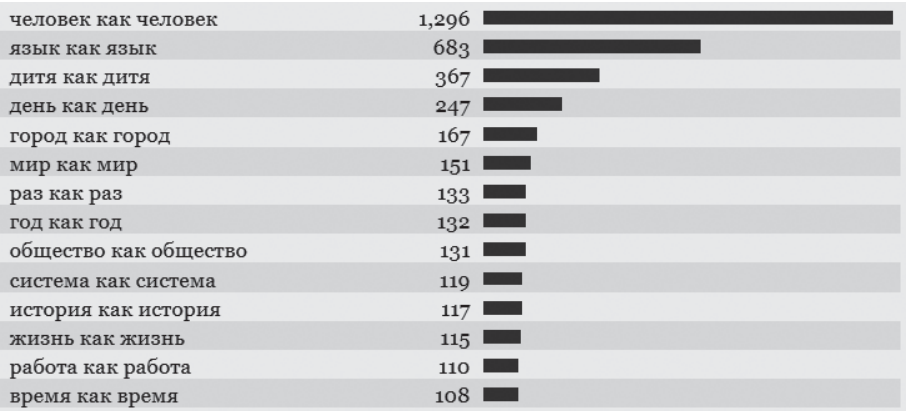Example 4: Х как Х (X jako X, X ako X) (X as X).

человек как человек	1,296
язык как язык	683
дитя как дитя	367
день как день	247
город как город	167
мир как мир	151
раз как раз	133
год как год	132
общество как общество	131
система как система	119
история как история	117
жизнь как жизнь	115
работа как работа	110
время как время	108

Fig. 1. Frequency of occurrence of expressions like «Х как Х» in the Araneum Russicum Maximum corpus

hra jako hra	245
soud jako soud	220
člověk jako člověk	171
přání jako přání	143
den jako den	128
voda jako voda	112
práce jako práce	112
trh jako trh	110
tuk jako tuk	97
právo jako právo	93
les jako les	78
škola jako škola	77
jazyk jako jazyk	76
olej jako olej	67

Fig. 2. Frequency of occurrence of expressions like «Х jako Х» in the Araneum Bohemicum Maximum corpus

súd ako súd	250	
človek ako človek	111	
deň ako deň	104	
škola ako škola	89	
voda ako voda	88	
tuk ako tuk	88	
osoba ako osoba	70	
práca ako práca	62	
jazyk ako jazyk	62	
káva ako káva	57	
život ako život	54	
poistenie ako poistenie	47	
láska ako láska	46	
právo ako právo	45	

Fig. 3. Frequency of occurrence of expressions like «X ako X» in the Araneum Slovacum Maximum corpus

Based on the data in Figures 1-3, it is clear that there are frequent set combinations of such type (for Russian ipm=1.20, for Czech ipm=3.60, for Slovak ipm=4.40), and that they differ in different languages and that Czech and Slovak are more close between themselves.

<u>Example 5</u>: Все X суть Y (všichni X jsou Y, všetci X sú Y) (All X are Y).

Russian	freq.	Czech	freq.	Slovak	freq.
все люди братья	447	všichni muslimové jsou teroristé	33	všetci moslimovia sú teroristi	19
все мужики сволочи	167	všichni teroristé jsou muslimové	123	všetci ľudia sú bratia	15
все мужики козлы	163	všichni lidé jsou bratři	113	všetci politici sú zlodeji	9
все люди взрослые	90	všichni politici jsou zloději	69	všetci Rómovia sú zlodeji	6
все ребята молодцы	84	všichni muži jsou násilníci	61	všetci ľudia sú hriešnici	5

Tab. 7. Frequency of occurrence of the like-expressions in the three corpora

As in Figures 1-3, Table 7 shows different filling of variables X and Y for different Slavic languages. It is fair to say that these data to a certain extent reflect the linguistic consciousness of the native speakers. These data would have been hard to obtain without a large representative corpus.

4   COMPARABLE PHRASEOLOGY AND ISSUES OF TRANSLATION

4.1   Idiom Equivalency

The subject of comparable phraseology in the narrow sense of the word are phraseological units of different languages that have similar semantic or structural

characteristics. One of the issues of a comparable analysis is the typology of interlingual equivalency. We have considered such idioms and constructions (without going into much detail) in Section 3.

The interlingual equivalency, including that of phraseological units, is also the equivalency in the system of a language (a dictionary problem) and the issue of translation of phraseological units in an individual text. While the translation of a set phrase of the "denotative" type is not much of a challenge (*silný čaj – крепкий чай* (strong tea), *nevlastní otec – nevlastný otec – неродной отец* (step-father)), the translation of other phraseological units is far from simple. This is explained both by their "non-denotativeness" and the absence of full, semantically organized collections of foreign idioms. Sometimes, such translation can be found quite easily (*делать из мухи слона – dělat z komára velblouda – robiť z komára somára*), and sometimes it requires a creative approach.

One should remember that the majority of idioms do not have direct equivalents in the compared language. Then, they can be translated descriptively or you can try and find one of the semantic equivalents. For example: *дуракам закон не писан* (literally, the law is not written to fools) – Czech. *hloupa kůže všechno může* (a stupid skin can do anything)*, на безрыбье и рак рыба* (literally, when there is no fish, a crawfish is fish) – Czech. *z nouze Franta dobrý* (even František is good when there is need) [11].

Besides, there is a partial equivalency that can be called the problem of the relation "many to many". For example, in Russian, one can say *молчит как ... рыба, партизан, истукан, вода, пень ... как язык проглотил* (silent as … fish, guerilla, statue, water, stub… as if has swallowed his tongue) [5]. In Czech, it will be *mlčí jako ... dub (*oak)*, sfinga (*sphinx)*, hrob (*grave)*, pěna (*foam)*, ryba (*fish)*, jako zařezaný (*slaughtered)*, jako by mu přimrzl jazyk (*as if his tongue froze down)* [9]. In Slovak, it will be *mlčí ako ... hrob, kameň', peň (*stub)*, ryba, sfinga, zarezaný, ako voš pod chrastou* (as a louse under blemish) [1]. This raises a question of what variant to choose in translation or to cite in a dictionary.

Example 6:
Rus. *Пьяный* (drunk) *как … сапожник* (as a shoemaker)*, свинья* (pig)*, скотина* (cattle)*, в доску* (in board)*, в стельку* (in insole)*, в дым* (in smoke)*, вдрызг, вдребезги* (into smithereens) [5];
Czech. *opilý (vožralý, nalitý, zpitý)* (drunk, tight) *jako Dán* (as a Dutch)*, dráteník* (potter)*, zvíře* (animal)*, Holandr* (Dutchman)*, duha* (rainbow)*, námořník* (sailor)*, pod obraz (boží)* (unlike God's image) [9];
Slov. *opitý (spitý)* (drunk) *ako cepelín* (airship)*, snop* (sheaf)*, čík* (misgurnus (fish))*, prasa* (pig)*, na mol* (in flat note)*, do nemoty* (to muteness)*, pod obraz boží* (unlike God's image) [1].

Which of the adjectives (or respective verbs) are used more often? Which of the comparisons are used more often? Which of the word combination have more association strength? A large corpus can give at least preliminary answers to these questions. However, naturally, those answers would not be exhaustive (see Table 8).

Russian	freq. (ipm)	Czech	freq. (ipm)	Slovak	freq. (ipm)
пьяный как …	*1163 (0,10)*	*opilý jako …*	*165 (0,03)*	*opitý ako …*	*345 (0,10)*
свинья	79	prase	21	čík	26
сапожник	53	Dán	10	doga	19
стелька	7	doga	6	delo	15
скотина	7	žok	4	prasa	14
собака	5	dělo	4	teľa	6
фортепьян, швед, ямщик, лорд, зюзя,извозчик...	…	slíva, konev, štoudev, kára, kráva, hovado, zeppelín...	…	sviňa, Dán, snop, činka, vôl, Rus, Poliak...	…

Tab. 8. Frequency of the combinations with «пьяный как …» and their equivalents in Czech and Slovak in three corpora

Table 8 gives only some examples whose aim is to show the differences in word combinations between the dictionaries and large corpora of a "live" language. Some of the expressions which are present in corpora and absent in dictionaries shall, without doubt, be included in them. For example, for the Czech language, those are *opilý na mol* (45 occurrences) and, possibly, *opilý na plech* (7) and *na šrot* (4). Besides, frequencies obtained from representative and balanced corpora allow to specify the sequence of phraseological units in dictionaries.

4.2 Parallel Corpora

One of the sources that can help in solving the tasks of comparable phraseology and translation are parallel corpora. Let us give several examples from the Intercorp which is a part of the Czech National Corpus [10].

Example 7: *pepka klepne* (apoplexy)

Otcenasek-Kulhavy_Orfe	Vo pár důstojnejch tatíků se pokoušela **pepka**.	Нескольких почтенных дядюшек чуть **кондрашка** не хватила.
Otcenasek-Kulhavy_Orfe	A nech toho chlastání, nebo mamulu klepne **pepk**a, až tě uvidí takhle zhulákanýho .	И перестань хлестать , а то мамулю **кондрашка** хватит, как увидит тебя такого развеселого.
Doncova-Manikura_pro	Říká se, že můj otec byl úplně stejný, do své smrti běhal za ženskými a pak ho klepla **pepka**.	Говорят , отец мой такой был, до смерти по бабам бегал и от **инфаркта** умер.

Tab. 9. Examples of translation of the «pepka klepne» word combination in the Intercorp corpus

Example 8: *obrátit vniveč, přijít vniveč* (render null, go down the drain)

Eco-Jmeno_ruze	A je - li upálen člověk, shoří i jeho individuální substance a s ní je **vniveč** obráceno i konkrétní bytí, skutečné, a už jen proto dobré, aspoň v očích Boha, který je naživu držel.	Между тем, когда человек сгорает, раньше всего сгорает его индивидуальная субстанция, и при этом **аннулируется** то, что прежде составляло конкретный акт существования - очевидно, благой по своей идее, хотя бы на взгляд Господа Бога, который для чего-то потворствовал сему существованию.
Wells-Stroj_casu	Podle mého by byla škoda, kdyby to jídlo přišlo **vniveč**, poznamenal redaktor chvalně známého deníku	Досадно, если обед будет **испорчен**, - сказал Редактор одной известной газеты.
Granin-Krasna_Uta	Kdepak, dobro nepřichází **vniveč**, spíš zlo může zmizet, ztratit se v něčí duši, zlo lze odpustit, zapomenout, ale dobro se podle všeho neodpouští.	Нет, нет, добро не пропадает, скорее зло может **пропасть**, сгинуть в чьей-то душе, зло можно простить, забыть, а добро, оказывается, не прощают.
Wells-Valka_svetu	Všechna naše práce **vniveč**, všechna ta práce...	Все наши труды **пропали**, все труды...

Tab. 10. Examples of translation of the word combinations with the word «vniveč» in the Intercorp corpus

Use of parallel corpora, on the one hand, shows which of the multi-variant phraseological equivalents are most often used by translators, and, on the other hand, it can enrich dictionaries and text books.

5   ANALYSIS OF PHRASEOLOGICAL UNITS USAGE IN DIACHRONY

It is common knowledge that the language is a dynamic system, which shall be reflected in dictionaries and grammars. However, this chronological aspect is far less common for the study of phraseological units and other set phrases. One of the reasons is the absence (till the recent times) of large historical (diachronic) corpora.

When we tried to study the behaviour of some phraseological units in time based on the National Russian Corpus (`http://ruscorpora.ru`), the experiments showed that its volume (283 million tokens) was too small for such tasks.

Fortunately, there is a large diachronic corpus Google Books (books for the period from 1800 to 2008) that exists for the Russian language (and eight others). The volume of the Russian corpus is 67 billion tokens. Let us show the results which can be obtained using it. Unfortunately, neither Czech, nor Slovak are included.

Example 9: *кошки скребут*

Fig. 4. The curves of occurrence of word combinations with the expression «кошки скребут» in the corpus Google books Ngram Viewer

The curves at Fig. 4 show that approximately up until the end of the 1940s the phraseological unit «на сердце кошки скребут» was used more often. Then, the situation changed drastically. And we see that the majority of the occurrences account for hard 1980–1990s.

Example 10: *ничтоже сумняшеся* (without a moment's hesitation)

The dictionaries present this phraseological unit in two forms: «ничтоже сумняшеся» and «ничтоже сумняся». Historians are well acquainted with this expression. But it is the corpus that will tell us how this expression "lived" in the language for centuries.

Fig. 5. The curves of occurrence of the bigram «ничтоже сумняшеся» in the corpus of Google books Ngram Viewer

We see in Fig. 5 that for quite a long time, the main form was «ничтоже сумняся». For example, in 1889 this form was used in literature 4 times more often.

It is no coincidence that we see this very form in Chekhov's books. «Для нее ясна была эта красивая смелость современного человека, с какою он, не задумываясь и ничтоже сумняся, решает большие вопросы и строит окончательные выводы» (A.P. Chekhov. «Несчастье» (Misfortune)). The variant «ничтоже сумняшеся» became preferred as late as in the second decade of the 20th century.

A search in the Google Ngram Viewer allows to identify other word combinations with the word «ничтоже», which can also be of interest for linguists: «ничтоже есть», «ничтоже суть», «ничтоже бысть».

Example 11: *перебиваться с ... на ...* (in the meaning "live in great poverty").

Fig. 6. The curves of occurrence of set phrases with the word «перебиваться» in the corpus of Google books Ngram Viewer

The curves of Fig. 6 show us that the most frequent set phrases with the verb «перебиваться» are those that are given in phraseological dictionaries (*с хлеба на квас, с хлеба на воду*) and that became used actively as late as in the 20th century.

6   CONCLUSION AND FURTHER WORK

The phraseology of any language is rich and variable. However, in order to see all this variability, we need large corpora, taking into account the relatively low frequency of usage of phraseological units in texts. Fortunately, for Russian, Czech, and Slovak, such corpora exist.

The research has shown that the corpus linguistics tools and corpora allow to identify and significantly enlarge the lexicon of set phrases of various types and peculiarities of their functioning. Based on corpora, linguists can create dictionaries and text books of a new generation, including phraseological dictionaries where the collocability will be represented far more widely than ever before. It is desirable that such dictionaries had such quantitative characteristics as the association strength in synchrony, and the history of usage in diachrony.

It is feasible to continue the research by choosing for experiments various types of phraseological units and, possibly, including other Slavic language. During the

research, we also found repeatedly that in order to make credible conclusions based on corpus data one should be aware of the disadvantages and the limitations of the tools used.

ACKNOWLEDGEMENTS

References

[1]   Avramovová, M et al. (2006–...). Slovník súčasného slovenského jazyka. Jarošová, A., editor, Veda, vydavateľstvo SAV, Bratislava.

[2]   Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. In *Proceedings of the 17th International Conference Text, Speech and Dialogue,* pages 257–264, Springer International Publishing Switzerland (LNCS 8655).

[3]   Birikh, A. K., Mokiyenko, V. M., and Stepanova, L. I. (1997). *Slovar' frazeologicheskikh sinonimov russkogo yazyka*. [Dictionary of phraseological synonyms of the Russian language.] Rostov-on-Don.

[4]   Chernysheva, I. I. (1970). *Frazeologiya sovremennogo nemetskogo yazyka*. [Phraseology of the modern German.] Moscow.

[5]   Denisov, P. N. and Morkovkin, V. V., editors (1983). *Slovar' sochetayemosti slov russkogo yazyka.* [Collocability dictionary of Russian language words.] Russkiy yazyk, Moscow.

[6]   Čermák, F. and Hronek, J. (1994). *Slovník české frazeologie a idiomatiky. Výrazy slovesné*. Academia, Praha.

[7]   Melčuk, I. A. (1960). *O terminakh 'ustoyvhivost'' i 'idiomatichnost''*. [About the terms steadiness and idiomaticity.] *Voprosy yazykoznaniya* [Questions of Linguistics], 4:73–80.

[8]   Michel, J-B. et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331:176; DOI 1126/Science. 1199644. Accessible at: `http://www.sciencemag.org/content/331/6014/176.full.html`, retrieved 2017-01-30.

[9]   Mokienko, V. and Wurm, A. (2002). *Česko-ruský frazeologický slovník*. Olomouc.

[10] Rajnochová, N., Runštuková, N., and Vavřín, M. (2016). Korpus InterCorp – ruština. Verze 9.9. Ústav Českého národního korpusu FF UK. Praha. Accessible at: `http://www.korpus.cz/`.

[11] *Russko-cheshskiy slovar*. [Russian-Czech dictionary.] (1978). Moskva – Praha.

[12] *Slovatsko-russkiy slovar*. [Slovak-Russian dictionary.] (1976). Bratislava – Moskva.

[13] Slovenský národný korpus – prim-7.0-public-all. Jazykovedný ústav Ľ. Štúra SAV, Bratislava. 2015. Accessible at: `http://korpus.juls.savba.sk/`.

[14] Zakharov, V. P. and Masevich, A. Ts. (2014). Diakhronicheskiye issledovaniya na osnove korpusa russkikh tekstov Google books Ngram Viewer [Diachronic researches on the base of the Russian Google books Ngram Viewer text corpus.] *Strtructural and Applied Linguistics* [Strukturnaya i prikladnaya lingvistika], 10:303–327.