Constructions in Parallel Corpora: A Quantitative Approach

Dmitrij Dobrovol'skij[1(✉)] and Ludmila Pöppel[2]

[1] Russian Language Institute of the Russian Academy of Sciences, Volkhonka18/2, 119019 Moscow, Russia
dobrovolskij@gmail.com
[2] Department of Slavic and Baltic Studies, Finnish, Dutch and German, Stockholm University, 10691 Stockholm, Sweden
ludmila.poppel@slav.su.se

Abstract. The primary goal of the present study is to find an adequate method for the quantitative analysis of empirical data obtained from parallel corpora. Such a task is particularly important in the case of fixed constructions possessing some degree of idiomaticity and language specificity. Our data consist of the Russian construction *дело в том, что* and its parallels in English, German and Swedish. This construction, which appears to present no difficulty for translation into other languages, is in fact, language-specific when compared with other languages. It displays a large number of different parallels (translation equivalents) in other languages, and possesses a complex semantic structure. The configuration of semantic elements comprising the content plane of this construction is unique. The empirical data have been collected from the corpus query system Sketch Engine, subcorpus OPUS2 Russian, and the Russian National Corpus (RNC). We propose to use the Herfindahl index as a tool for quantitative analysis in order to measure the degree of uniformity in the frequency distribution of the various translations of the construction under investigation. This tool is not universal and does not enable us to answer all the questions that arise in connection with determining the specificity of language units. However, it clearly helps to obtain more objective results and to refine the quantitative analysis of idiomatic constructions on the basis of corpus data.

Keywords: Construction · Contrastive corpus analysis · Parallel corpus · Russian · English · German · Swedish · Language specificity · Quantitative analysis · The Herfindahl Index

This paper is based on work supported by the Russian Science Foundation (RSF) under Grant 16-48-03006 "Semantic Analysis of Translated Texts for Comparative Cultural Studies and Cultural Specificity in Language Learning".

R. Mitkov (Ed.): Europhras 2017, LNAI 10596, pp. 41–53, 2017.
https://doi.org/10.1007/978-3-319-69805-2_4

1 Goals, Methods and Data

The present investigation employs quantitative methods with the goal of enhancing the reliability of findings obtained from parallel corpora. As materials for analysis we use the Russian construction *дело в том, что* (*delo v tom, čto*),[1] which has a great many translation equivalents in other languages. This study will examine its parallels in English, German and Swedish.

Empirical data are taken from the parallel corpora of the Sketch Engine search system, the subcorpus of parallel texts OPUS2 Russian (307 709 872 tokens) and the Russian-English, English-Russian, Russian-German and German-Russian corpora of parallel texts in the Russian National Corpus (RNC). The construction *дело в том, что* was searched in Sketch Engine in the pairs of corpora OPUS 2 Russian and OPUS2 English, OPUS2 Russian and OPUS2 German, OPUS2 Russian and OPUS2 Swedish. None of the Sketch Engine OPUS2 subcorpora mark the direction of the translation – the English-Russian and Russian-English parallels, for example, are in the same corpus – so that this distinction is not indicated in the description of the Sketch Engine data. The quantitative data cited in the present study were obtained in July 2016.

The following methods were used:

- a quantitative research method based on an analysis of parallel text corpora;
- a quantitative method using the Herfindahl index as a statistical tool that allows us to identify the degree of uniformity in the frequency distribution of the various translations of the item under investigation.

Thus our work represents a contribution to the development of contrastive corpus studies and methods for the quantitative analysis of corpus data.

2 Previous Research

We have previously examined the construction *дело в том, что* in (Dobrovol'skij and Pöppel 2016a; 2016b). These works did not use any statistical apparatus, i.e. the analysis was qualitative rather than quantitative.

Dobrovol'skij and Pöppel (2016b) tested the following hypothesis:

The Russian expression *дело в том, что* displays a unique configuration of semantic components; that is, it possesses a certain language-specificity. It has a large number of various parallels in other languages, and the choice of each variant depends on specific contextual conditions.

Dobrovol'skij and Pöppel (2016a) tested the hypothesis that discursive constructions based on the same pattern do not have the same linguistic status. *Дело в том, что*, for example, should be regarded as a unit of the lexicon, whereas the constructions *проблема в том, что* and *правда в том, что*[2] are free co-occurrences.

[1] Literally: *the thing is that.*

[2] Literally: *the problem is that* and *the truth is that*.

Language-specificity is examined in earlier studies such as Wierzbicka (1992; 1996), Zaliznjak et al. (2005; 2012), Zaliznjak (2015), and Šmelev (2002; 2014; 2015). Šmelev (2015) distinguishes three parameters of the phenomenon.

The first is connected with the number of languages, which lack a unit that at least approximately corresponds to the source expression. The more such languages that can be identified, the greater degree to which the expression can be considered language-specific.

The second parameter consists in the specificity of the content aspect of the expression, including connotations, background components of meaning, etc. (Šmelev 2015), from which it follows that the degree of distinctiveness of the semantic configuration of an expression is directly proportional to its degree of language-specificity.

The third parameter is a corollary of the second: the more distinctive the semantic configuration of a lexical unit, the more difficult it is to find an adequate translation equivalent of this unit in another language.

Šmelev (2015) notes that the object of translation is not individual words but texts, so that the translator can deviate from exact equivalence on the lexical level without regard to the language-specificity of the corresponding units. Nevertheless, it is natural to interpret the presence of a large number of different translation equivalents as indicating the absence of a systematic equivalent. This allows us to measure quantitatively the degree of language-specificity in accordance with this third parameter, which is in fact the focus of the present study.

Previous investigations have also pointed to the need for quantitative analysis to identify the degree of language-specificity. Thus Buntman et al. (2014) note that it is necessary to determine how many translation equivalents exist for potentially language-specific lexical units. It then proposed to evaluate their dispersion, but there is no discussion there of any concrete means for such an evaluation. Sitchinava (2016) does suggest such a tool for quantitatively analyzing the degree of language specificity, namely the Herfindahl index.[3] This method is used in the present study.[4]

3 Qualitative Analysis

Analysis of the corpus data allows us to identify only the degree of variety in the means of translating a given expression into other languages. When one or another expression lacks a generally accepted standard context-independent translation equivalent, we can speak of an absence of systematic equivalents, i.e., a kind of non-equivalence. Whether such non-equivalence is connected with the category of language specificity remains an open question.

Our qualitative analysis uses data obtained in Dobrovol'skij and Pöppel (2016b). The following English correlates were found in Sketch Engine:

[3] For more detail see Sitchinava (2016).

[4] The quantitative method for analyzing fixed expressions in monolingual corpora is used in such studies as Zhu and Fellbaum (2015), Steyer (2015).

zero equivalents [154];[5]
the fact is (that) [123];
the thing is (that) [98];
the point is (that) [70];
(it's/this/that is) because/because of [40];
it's just (that)/it's that/just/this is that [27];
in fact [26];
the truth is (that) [26];
however [16];
the fact of the matter is (that) [15];
indeed [13];
the problem is (that) [12];
you see [9];
the reason is (that) [8];
as a matter of fact [5];
for [5];
it's/this is about [5];
it happens that/as it happened/what has happened is/what is happening is [5];
the matter is (that) [5];
but [4];
since [4];
it's a fact that [4];
well [3];
basically [3];
what's true is (that)/it was true (that) [3];
the consequence is (that) [3].

The following parallels occurred twice: *the truth of the matter is (that)*; *the answer is (that)*; *the concern is (that)*; *the crux of the matter is (that)*; *the question is (that)*; *you know*; *look*; *the position is (that)*; *the thing about; in effect.* We also found more than 43 single English correlates: *the situation is; that means that; my story is; the issue is; the reality is; the content is; the explanation is; the fact remained that; the fact that; this is due to; it has everything to do with; what I'm trying to say is that; except that; that is; in reality; actually; in practice; the word is; the plan was; here's the thing; this is the situation; sort of; the point being; the purpose of; it is not that; thus; it should be noted that; in truth; for the reason that; as it was; rather; in that it is; that is; instead; namely; in that connection; in this regard; it is which; to be blunt; here too; it is a matter of; accordingly; the trouble is.* A total of 80 different types of equivalents were found.

The RNC Russian-English parallel corpus contained 26 translation equivalents, among which the zero equivalent was the most frequent:

zero equivalent [27];
the fact is (that) [14];

[5] Figures in brackets indicate total number of hits.

the thing is (that) [14];
the point is (that) [10].
Less common was:
you see [3];
actually [2];
in point of fact [2];
the matter is (that) [2].

18 equivalents occurred only once – *this came about in the following way*; *well*; *for*; *the fact of the matter was that*; *the truth of the matter was that*; *it was exactly that*; *the trouble was that*; *it is that*; *the important point is that*; *the chief thing is that*; *it all lies in the fact that*; *all that matters is that*; *it was true that*; *it was because that*; *the difficulty was that*; *the question is*; *the whole point is*; *the fact remains that.*

These results partly coincide and partly diverge. Four of the most frequent equivalents – zero equivalent, *the fact is (that); the thing is (that) and the point is (that)* – completely coincide, which indicates that the findings are non-random. At the same time, the relatively frequent constructions found in Sketch Engine – *in fact; the truth is (that)* and *however* – do not occur in the RNC, whereas (*it's/this/that is) because/because of; it's just (that)/it's that/just/this is that and the fact of the matter is (that)* – occur only once. These divergences are entirely due to the different sizes of the corpora. Sketch Engine is much larger than the RNC. In addition, the texts in these corpora differ with respect to genre. The RNC contains almost exclusively fictional texts, whereas non-fiction dominates in Sketch Engine.

In Sketch Engine we found 20 German parallels:

zero equivalent [19];
die Sache ist die (dass) [8];
aber [5];
es geht darum, dass [4];
es ist (doch) so, dass [3];
die Wahrheit ist, dass [3];
wissen Sie [2];
nur (dass) [2];
Tatsache ist (nun mal) [2];
es ist nur (dass) [2];
ich meine [2];
der Punkt ist [2];
weil [1];
es ist, was [1];
um die Wahrheit zu sagen [1];
jedoch [1];
das passiert [1];
der Grund dafür ist, dass [1];
das Schlimme ist, dass [1];
wichtig ist nur [1].

The search in the RNC yielded 13 correlates. Some of them coincide with the correlates found in Sketch Engine, some of them not:

die Sache ist die (dass) [18];
zero equivalent [11];
nämlich [9];
es handelt sich darum, dass [3];
die Hauptsache ist, (dass) [3];
doch [2];
der Grund war, (dass) [2];
es kommt (vielmehr/doch nur) darauf an [2];
der Kernpunkt ist vielmehr, dass [1];
die Sache liegt so, dass [1];
es hängt ganz davon ab [1];
es geht darum, dass [1];
weil [1].

Two of the most frequent parallels in Sketch Engine – the zero equivalent and *die Sache ist die (dass)* – coincide with the most frequent ones in the RNC, although in reverse order. The most important difference is the absence of *nämlich* in Sketch Engine, whereas in the RNC it occurs 9 times. This difference is significant because even a superficial analysis of the word *nämlich* shows that its communicative function is very close to that of the Russian construction *дело в том, что*. On the whole, the German parallels display considerable scatter.

The Swedish equivalents are examined only on the basis of the Sketch Engine data, since this is the only text corpus at our disposal. We found 25 Swedish parallels:

zero equivalent [45];
saken är den att [16];
men [8];
problemet är att [7];
faktum är att [4];
det viktiga är (att)/det är viktigt att [4];
det är för att [4];
sanningen är att [3];
grejen är den att [3];
poängen är att [3];
för (att) [3];
det handlar om att [2];
det vad jag vill säga är att [2];
i själva verket [2];
jag/han menar att [2];
det beror på att [1];
det är vad [1];
om [1];
bara [1];
då [1];

faktiskt [1];
det var inte meningen att [1];
oron är att [1];
läget är att [1];
vad jag menar är [1].

The most frequent are the zero equivalent and *saken är den att*. In the intermediate zone (from 10 to 2) there are 13 equivalents, while 10 equivalents are used only once. Here as well we can speak of considerable scatter.

We also consulted the RNC English-Russian and German-Russian parallel corpora, since the objectivity of the findings is increased by testing the hypothesis on materials in which the source texts are not Russian. In the English-Russian corpus we found 54 different English stimuli for the Russian *дело в том, что,* of which 6 equivalents occur more than 10 times each:

zero equivalent [38];
the fact is (that) [36];
for [34];
it's just (that)/it's that/just/this is that [16];
(that is) because [14];
(as) you see [11].

Besides, we found 15 less frequent equivalents, they occur between 10 and 2 times:

well [7];
the thing is (that) [7];
but [5];
it happens (that) [4];
actually [4];
the truth is (that) [4];
the point is (that) [4];
in fact [4];
the reason is (that) [3];
the problem is (that) [3];
I mean [2];
as a matter of fact [2];
I tell you [2];
in truth [2];
to begin with [2].

33 equivalents were found only once: *apparently; it should be understood that; you should understand (that); it appears that; to all appearance; listen; so; I think; it seemed; it depends on; I happen to be; it so happens; it's something in the way; it was the feeling that; the trouble is that; we are asking how; it was due to the fact that; it just amounts to; you know; I may say; it's like this; in the first place; merely; it was a case of; I suppose; that's the proposition; and; you must know; let it suffice to say; now; that's the matter; I believe; nevertheless*.

It is natural to compare these findings with those of the RNC Russian-English parallel corpus, where the corresponding figures are as follows: 3 correlates occur more than 10 times, five range from 10 to 2, and 18 are found only once. Only two equivalents are among the most frequent – the zero equivalent and *the fact is (that)*. This comparison indicates that when translating from Russian to English, translators tend to follow the form of the original, using constructions such as *the fact is (that); the thing is (that)* and *the point is (that)*. Going from English to Russian, however, they are inclined to use the discursive construction *дело в том, что* in places where it is not dictated by form. Thus the most frequent group of English correlates includes lexical units such as *for, just, because, you see*. Actively employed as well are syntactic means such as the cleft. Cf. (1).

(1a) "[…] I'm sorry about this –" My voice was shaking a little, but I couldn't get it under control. "– *it's just that* we can't seem to find Mr. Lagerfeld. [Lauren Weisberger. The Devil Wears Prada]

(1b) […] Я прошу прощения, но… – мой голос слегка дрожал, и я никак не могла унять эту дрожь, – *дело в том, что* мы, кажется, не можем отыскать мистера Лагерфельда.

The following correlates were found in the RNC German-Russian corpus:

nämlich [27];
zero equivalent [11];
die Sache ist die, (dass) [10];
denn [8];
eben [3];
aber [3];
es kommt darauf an [2];
gerade [1];
eigentlich [1];
die Tatsache [1];
doch [1].

A comparison of the RNC German-Russian and Russian-German parallel corpora yields very similar results. The following features stand out. The formal correlate *die Sache ist die, (dass)* dominates in translations from Russian to German, while in the German-Russian corpus the word *nämlich* often correlates with *дело в том, что,* fulfilling the same function even though the two expressions have nothing in common in terms of form. This confirms what was stated earlier. Cf. (2).

(2a) Prinzessin Momo hatte *nämlich* einen Zauberspiegel, der war groß und rund und aus feinstem Silber. (Michael Ende. Momo (1973))

(2b) *Дело в том, что* у принцессы Момо было большое круглое Волшебное Зеркало из чистейшего серебра.

(2c) *You see,* Princess Momo had a magical mirror. It was big and round, and it was made of the finest silver.

Another feature of the German-Russian corpus is that the group of relatively frequent parallels includes the causal conjunction *denn*, which is similar in frequency to the English conjunctions *because* and *for* in the English-Russian corpus.

The empirical data presented in the study indicate the following:

1. The construction *дело в том, что* has many different translation equivalents in English, German and Swedish. Most of these are not mutually synonymous, and choice depends on contextual conditions. This means that *дело в том, что* should be regarded not as a free co-occurrence, but as a unit of the lexicon.
2. The construction *дело в том, что* is characterized by a complex configuration of semantic features. Its semantic structure includes at least the following meanings: substantiation of something stated previously; indication of the reason something has taken place; emphasis on the special significance of the following clause.

Selection of equivalents from the various groups depends on which of these meanings is being highlighted in the utterance. Thus the English equivalent *you see* in the translation of the sentence *Дело в том, что сегодня рождение моей матери – You see, it's my mother's birthday today*; German *nämlich* in *Дело в том, что ночью произошла небольшая катастрофа – In der Nacht nämlich geschah eine kleine Katastrophe* and Swedish *nu är det så* in *Дело в том, принцесса, что у меня есть приказ – Nu är det så, Prinsessan, jag har order* all explain what was stated previously.

In cases where the focus is on the reason or cause, English, German and Swedish translations use causal subordinating conjunctions such as, for example, English *because* in *Ну, дело в том, что у меня есть сюрприз для тебя – Well, because I have a surprise for you*; German *denn* in *Дело в том, что тот, кто заглядывал в, Волшебное Зеркало и видел в нем свое отражение, становился смертным. – Denn wer sein eigenes Spiegelbild darin erblickte, der wurde davon sterblich.* or Swedish *för* in *Дело в том, что если я должен вам, то собрать такую сумму мне будет трудновато. – För att jag är skyldig dig pengar, som jag inte kan få fram.*

When the following clause is emphasized as being especially important, English, German and Swedish employ focusing particles or constructions such as, for example, English *the point is* in *Но дело в том, что я уверен, что это место действительно существует – But the point is, I'm convinced the place definitely exists;* German *der Punkt ist* in *Дело в том, что я влюблен в неё, и это сводит меня с ума – Der Punkt ist, ich bin in sie verliebt und es macht mich wahnsinnig* and Swedish *det viktiga är* in *Но дело в том, что я уверен, что это место действительно существует – Men det viktiga är, jag är övertygad att den platsen verkligen existerar.*

The Russian expression *дело в том, что* simultaneously explains what was said previously, points to the reason something has taken place, and singles out the following statement as especially significant.

4 Quantitative Analysis

The Herfindahl index was used to measure the degree of uniformity in the frequency distribution of the various translations of the construction under investigation. This index is used in economics to indicate the extent of market monopolization. In linguistics its uses include identification of the level of language specificity of various words (Sitchinava 2016). Our study has similar goals. The more uniform the frequency distribution, i.e., the lower the Herfindahl index, the more language-specific the given

unit. The higher the Herfindahl index, the lower the degree of language specificity of the expression, since some particular method of translation dominates and is thus standard.

The non-normalized Herfindahl index (H) is calculated using the following formula:

$$H = \sum_{i=1}^{n} f_i^2 \tag{1}$$

where n is the total number of translation equivalents and f_i^2 is the squared relative frequency of an equivalent.

The normalized Herfindahl index (H*) is calculated as:

$$H^* = \frac{H - 1/n}{1 - 1/n} \tag{2}$$

The Herfindahl index ranges from $1/n$ to 1, the normalized Herfindahl index ranges from 0 to 1.

Our calculations according to the Herfindahl index are presented in Table 1.

Table 1. The Russian construction *дело в том что* in parallel corpora

Subcorpus	H	H*
Sketch Engine Russian-English-Russian	0,1036	0,0922
RNC Russian-English	0,1489	0,1148
Sketch Engine Russian-German-Russian	0,1342	0,0887
RNC Russian-German	0,1855	0,1176
Sketch Engine Russian-Swedish-Russian	0,1798	0,1470
RNC English-Russian	0,0851	0,0678
RNC German-Russian	0,2249	0,1474

All figures are rounded to 4 digits after the comma.

As is evident from Table 1, the non-normalized index (H) and the normalized one (H*) yield different results. Index H depends not only on the degree of uniformity in the frequency distribution, but also on the number of translation equivalents. Index H* allows us to compare the degree of uniformity in the frequency distribution for various language units regardless of the number of different translations of each of them. Thus if it is necessary to compare data obtained from corpora of different sizes, it is preferable to use H*. The H* indices are practically identical, showing that the degree of diversity among translations is the same (rather low in all cases) despite how many different translation approaches are used.

5 Discussion

The data obtained on the degree of translation variety can be meaningfully interpreted only when compared with findings obtained about other language units with the help of similar tools. Sitchinava (2016) uses the Herfindahl index to determine the degree of uniformity in the frequency distribution of translations into English and Ukrainian of words such as *пошлость* [banality/vulgarity], *удаль* [daring/bravado], *тоска* [melancholy/yearning], *пространство* [space], *уют* [coziness/comfort], *страсть* [passion], *простор* [expanse/vastness]. One of the goals of his study was to determine whether this uniformity of frequency distribution corresponds to the degree of language specificity. It was shown that on the whole, such a correspondence exists. A majority of the words analyzed that are traditionally considered to be language-specific display lower H and H* indices than do those which are not regarded as language-specific. This can be demonstrated on the basis of *простор* and *пространство*. *Простор* carries cultural meanings, whereas *пространство* denotes a universal category. Consequently, the Herfindahl index can be expected to be lower for *простор* and higher for *пространство*. Sitchinava's (2016) findings are presented in Tables 2 and 3.

As is evident from the tables, *простор* is language-specific relative to English, but not to Ukrainian, which is due to the proximity of Russian and Ukrainian and shared cultural roots. As for *пространство,* despite the universality of the corresponding concept, the Herfindahl index is lower for the English correspondences than for the Ukrainian ones. From this it can be concluded that even words expressing universal notions possess a certain degree of language specificity when more distant languages are compared. In the present study *дело в том, что* is not compared with equivalents in related languages, which is why Sitchinava's findings based on English materials are of interest to us. The results we have obtained from English, German and Swedish parallel corpora are similar to his findings based on English-Russian and Russian-English parallel corpora. There is reason to assume that *дело в том, что* possesses a high degree of language specificity.

Table 2. The Russian word *простор* in parallel corpora

Subcorpus	H	H*
RNC Russian-English	0,1327	0,0659
RNC English-Russian	0,0718	0,0613
RNC Russian-Ukranian	0,8306	0,8225
RNC Ukranian-Russian	0,7806	0,7795

Table 3. The Russian word *пространство* in parallel corpora

Subcorpus	H	H*
RNC Russian-English	0,3379	0,3217
RNC English-Russian	0,4495	0,4409
RNC Russian-Ukranian	0,6550	0,6494
RNC Ukranian -Russian	0,8611	0,8600

6 Conclusion

We have employed the Herfindahl index as a statistical method of analysis. Our findings show that the normalized Herfindahl index works best for similar linguistic investigations. Comparison with other words demonstrates that the results we obtained tend to resemble earlier findings based on language-specific words. Nevertheless, it cannot be unequivocally asserted that this construction is language-specific, since what the Herfindahl index measures is not the degree of language-specificity, but the degree of uniformity of frequency distribution.

References

Buntman, N.V., Zaliznjak, A.A., Zatsman, I.M., Kruzhkov, M.G., Loshchilova, E.J., Sitchinava, D.V.: Informacionnye texnologii korpusnyx issledovanij: principy postroenija kross-lingvističeskix baz dannyx (Informational technology in corpus-based studies: towards a cross-linguistic database). Inf. Appl. **8**(2), 98–110 (2014)

Dobrovol'skij, D., Pöppel, L.: Diskursivnaja konstrukcija *N в том, что* i ee paralleli v drugix jazykax: kontrastivnoe korpusnoe issledovanie. (The discursive construction *N в том, что* and and its correlates in other languages: A contrastive corpus analysis). Novosibirsk State Pedagogical Univ. Bull. **6**, 164–175 (2016a)

Dobrovol'skij, D.O., Pöppel, L.: The discursive construction *дело в том, что* and its parallels in other languages: A contrastive corpus study. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue 2016", issue 15 (22), pp. 126–137. RGGU, Moscow (2016b)

Günthner, S.: Die "die Sache/das Ding ist"-Konstruktion im gesprochenen Deutsch – eine interaktionale Perspektive auf Konstruktionen im Gebrauch. In: Stefanowitsch, A., Fischer, K. (eds.), Konstruktionsgrammatik II. Von der Konstruktion zur Grammatik, pp. 157–177. Tübingen, Stauffenburg (2008)

Sitchinava, D.: Parallel corpora as a source of defining language-specific lexical items. In: Margalitadze, T., Meladze, G. (eds.) Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, pp. 394–401. Ivane Javakhishvili Tbilisi University Press, Tbilisi (2016)

Šmelev, A.D.: Russkaja jazykovaja model' mira. Materialy k slovarju. (The Russian language picture of the world). Jazyki slavjanskoj kul'tury, Moscow (2002)

Šmelev, A.D.: Jazyk i kul'tura: est' li točki soprikosnovenija? (Language and culture: do they have points of interaction?). In: Proceedings of the V.V. Vinogradov Institute of Russian Language, issue 1, pp. 36–116. Russian Language Institute, Moscow (2014)

Šmelev, A.D.: Russkie lingvospecifičnye leksičeskie edinicy v parallel'nyx korpusax: vozmožnosti issledovanija i "podvodnye kamni" (Russian language-specific lexical units in parallel corpora: prospects of investigation and "pitfalls"). In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue 2015", issue 14(21), vol. 1, pp. 584–594. RGGU, Moscow (2015)

Steyer, K.: Patterns. Phraseology in a state of flux. Int. J. Lexicogr. **28**(3), 279–298 (2015)

Wierzbicka, A.: Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. Oxford University Press, Oxford (1992)

Wierzbicka, A.: Semantics: Primes and Universals. Oxford University Press, Oxford (1996)

Zaliznjak, A.A.: Lingvospecifičnye edinicy russkogo jazyka v svete kontrastivnogo korpusnogo analiza (Russian language-specific words as an object of contrastive corpus analysis). In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue 2015", issue 14(21), vol. 1, pp. 683–695. RGGU, Moscow (2015)

Zaliznjak, A.A., Levontina, I.B., Šmelev, A.D.: Ključevye idei russkoj jazykovoij kartiny mira (Key ideas of the Russian language picture of the world). Jazyki slavjanskoj kul'tury, Moscow (2005)

Zaliznjak, A.A., Levontina, I.B., Šmelev, A.D.: Konstanty i peremennye russkoj jazykovoj kartiny mira (Constants and variables of the Russian language picture of the world). Jazyki slavjanskoj kul'tury, Moscow (2012)

Zhu, F., Fellbaum, C.: Quantifying fixedness and compositionality in chinese idioms. Int. J. Lexicogr. **28**(3), 338–350 (2015)