International Journal of English Linguistics; Vol. 8, No. 3; 2018
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

A Corpus-Based Approach to Lexicography: A New English-Russian Phraseological Dictionary

Guzel Gizatova[1]

[1] Department of European Languages and Cultures, Kazan Federal University, Russia

Correspondence: Guzel Gizatova, Department of European Languages and Cultures, Kazan Federal University, Russia. E-mail: guzelgizatova@hotmail.com

Received: January 22, 2018 Accepted: February 26, 2018 Online Published: February 28, 2018

doi:10.5539/ijel.v8n3p357 URL: http://doi.org/10.5539/ijel.v8n3p357

Abstract

This paper addresses the principles of constructing the first English-Russian phraseological dictionary based on corpus data. The purpose of the present research is to introduce a methodology for organizing the selected items in a corpus-searchable phraseme list of a dictionary, to discuss linguistic issues presenting difficulties for bilingual lexicography and to analyze semantic asymmetry between English and Russian phrasemes. To achieve this goal, the following methodology has been introduced: analyzing and retrieving idioms from monolingual and bilingual idiomatic dictionaries, determining the degree of frequency of the selected idioms, considering variants of idioms and arranging them in a systematic way, and developing an idiom list. A phraseme is used in this article as a general term for a multi-word phrase with at least one fixed component. The article demonstrates the advantages of compiling a phraseological bilingual dictionary based on an analysis of corpus data and using authentic examples in the lexicographic description of phrasemes. Using corpora provides a new perspective on the contextual behavior of phrasemes and restrictions of their usage. The paper discusses the impact of using parallel English and Russian corpora for analysis of non-trivial features of English phrasemes, in comparison with their Russian equivalents, in the process of constructing an English-Russian phraseological dictionary. After an introduction, the article presents the methodology and data applied in the research and then discusses the results of the study; the author provides evidence of the advantages of using corpora in bilingual lexicography.

Keywords: bilingual, corpus, English, idioms, lexicography, Russian

1. Introduction

Bilingual lexicography encounters certain problems in connection with the treatment of phrasemes in dictionaries. In many cases, the generally accepted equivalent of a phraseme cannot be used to translate authentic texts, which is why our research strategy is to analyze cross-linguistic correlations between English and Russian phrasemes that have strong semantic resemblance, as well as to study semantic asymmetry in phrasemes.

Using relevant lexicographic information, text corpora, and parallel corpora, we shall study the frequency and semantic qualities of phrasemes by empirical methods in order to identify additional specific features that need to be included in the lexicographic description of a phraseme.

The need for a new English-Russian Phraseological Dictionary is indicated by the fact that there is presently no corpus-based bilingual dictionary of these languages. Due to corpus data, the dictionary presents a range of syntactic patterns, polysemous phrasemes, and unexpected variants, which cannot be retrieved from the existing bilingual and monolingual dictionaries of the English and Russian languages. Many dictionaries fail to register all meanings of phrasemes. The corpora help to reveal the specific character of their functional correlations and non-trivial semantic preferences of English phrasemes that do not have standard Russian equivalences.

The primary goal of the research is to conduct a thorough contrastive analysis with the purpose of discovering the unique properties of each phraseme and thus enhance the lexicographical description of

phraseological studies. The author analyzes one of the most curious cases of semantic asymmetry – phraseological "false friends" (Piirainen, 1997). Compare an English phraseme (1) and its Russian pseudo-equivalent (2).

(1) English *to twist (turn, wrap) somebody (a)round one's finger*

"to have the ability to persuade (a person) to do exactly as one wants (usually used to describe wives and daughters who persuade their husbands and fathers)" (Longman, 1979, p. 113).

(2) Russian *обвести вокруг пальца*

"to twist somebody (a)round one's finger"

"to deceive somebody skillfully" (Lubensky, 2004, p. 446).

These two phrasemes are basically equivalent, since they are identical with respect to both their lexicalized meaning and image component. However, it is actually not always possible to translate the English phraseme *to twist (turn, wrap) somebody (a)round one's finger* by the Russian phraseme *обвести вокруг пальца.* Analysis of authentic texts in corpora with the phraseme *to twist (turn, wrap) somebody (a)round one's finger* shows that there are many cases in which this phraseme has to be translated into Russian by the phraseme *вить веревки* (3).

(3) Russian *вить веревки (из кого-либо)*

"to twist the ropes (from someone)"

"compel someone to your will and force him to act your way" (Birikh, 2005, p. 89).

Let us take examples from the corpus query system Sketch Engine, parallel subcorpus OPUS-2.

He has worked there for 38 years and is planning to retire soon. His family includes his wife Linda, a son Jeff and a granddaughter "who absolutely has me *twisted around her little finger*." (enTenTen13).

Он работает там уже 38 лет и скоро планирует выйти на пенсию. У него есть жена Линда, сын Джефф и внучка, которая "беспощадно *вьет из меня веревки*" (enTenTen13).

Thus, despite the intuitively felt similarity of the phrasemes *to twist (turn, wrap) somebody (a)round one's finger* and *обвести вокруг пальца*, this similarity cannot be considered complete. For the lexicographer interested in the maximally precise description of the material, such instances are problematic. The problem is that some dictionaries present these phraseological "false friends" as full equivalents (cf. Kveselyevich, 2002, p. 350), not taking into consideration that between basically similar phrasemes in a source language and in a target language, there are practically always certain semantic, pragmatic, and syntactic differences. Our goal is to discover and describe these linguistic-specific differences in English and Russian phrasemes. We follow the theoretical concept of D. Dobrovol'skij in respect to cross-linguistic correspondence of phrasemes. "What is important for cross-linguistic correspondence… is not "phraseologicalness', but functional equivalence. It is this type of equivalence that is most interesting from the perspective of bilingual lexicography" (Dobrovil'skij, 2013, p. 212).

Apart from its theoretic relevance as an instrument for describing phrasemes of English and Russian languages, a new dictionary can be used for the purposes of translation and language acquisition.

The paper is structured as follows: after an introduction, the author presents the methodology and data used in the research, followed by a theoretical framework. Next, the article gives an overview of results of the study, followed by a discussion. The paper is summed up by conclusions.

2. Methodology and Data

The purpose of the project is to introduce a methodology for organizing the selected items in a corpus-searchable idiom list of the dictionary. To achieve the first goal, the following methodology has been introduced: analyzing and retrieving idioms from monolingual and bilingual idiomatic dictionaries; determining the degree of frequency of the selected idioms; considering variants of idioms and arranging them in a systematic way; and developing an idiom list. The idiom list of the dictionary is based primarily on that of Koonin's English-Russian phraseological dictionary (1996), on the on-line Oxford Dictionary of English Idioms (2016), and on the Cambridge International Dictionary of Idioms (1998). At present, the idiom list consists of approximately 1300 units, including variants, and is going

to expand to up to 3000 idioms. During the first stage of work on the idiom list, 2000 idioms were selected from the abovementioned dictionaries. They were analyzed, and their frequencies were checked in the enTenTen [2013] via Sketch Engine. Approximately 40% of them were excluded because they are seldom if ever used today, and they are thus not registered in the corpora.

The main method of research is the statistical corpus method, which includes the following aspects:

- use of parallel corpora;
- search for all translation equivalents of phrasemes under study;
- data processing using statistical methods;
- analysis of the results obtained.

The empirical data were collected from the corpus query system Sketch Engine, subcorpus [enTenTen13] (19,7 billion tockens), subcorpus [ruTenTen11] (14,5 billion tockens), English-Russian parallel subcorpus OPUS-2, and the parallel subcorpus of the Russian National Corpus (RNC). This made it possible to find instances of English phrasemes and their Russian equivalents under consideration and obtain statistically representative data.

3. Theoretical Concept

The theoretical issue of this research is the lexicographic consideration of the concept of equivalency in the present dictionary, since very often, the generally accepted equivalent of an idiom cannot always be applied to translate authentic texts. To address this issue, the author applies the theoretical concept introduced by D. Doobrovol'skij in his "Studien zur Deutschen Lexic" (2013), in which he argues about the importance of functional equivalence, but not "phraseologicalness," for cross-linguistic correspondence. The author cannot help but agree with this approach, since functional equivalents are parallels that can be used in similar situations "without any information loss".

Some ideas of construction grammar (CXG) are used as the theoretical background of this research work. On the early stages of CXG development, the main attention was focused on bordering compositional fields of the syntax and lexicon. Cf. definition of construction grammar given by A. Goldberg: "Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency" (Goldberg, 1995, p. 5). Similar ideas were expressed earlier in Fillmore et al. (1988).

In recent times, the concept of a construction has been expanding. This point is illustrated by the fact that many regularly used word combinations and models formed from them are used so frequently that they are stored in the memory of language speakers as single blocks, rather than generated according to syntactic, semantic and pragmatic rules (Goldberg, 2006). This point of view becomes more and more widely applicable because of the frequency of study of language units using statistical methods, which are an important constituent of CXG lines of research (Bybee, 2010; Stefanovich, Gries, 2003; Dobrovol'skij, Pöppel, 2016). The theoretical concern of this research is to refine our notions about the structural properties of phraseological constructions (PCs) *on the N of / на N* and identify additional distinctive features of the construction that should be included in this lexicographic description.

4. Results, Discussion

Using the corpus query system Sketch Engine and its English-Russian parallel subcorpus OPUS-2, the Russian National Corpus and its parallel subcorpus, we compared the constructions *on the brink of* and *on the threshold of* in relation to their synonymous proximity. The analysis of their 25 most frequently used co-occurrence partners demonstrate that out of 18,473 cases of usage of the PC *on the brink of*, it is most commonly used with the following nouns: extinction -3,253, collapse -2,815, war -2,524, bankruptcy- 1,923, disaster, destruction, and ruin. All these nouns express extreme, dangerous, critical, and often unpredictable situations, which tend to create problems. The least frequently used noun with the PC *on the brink of* is "millennium" -28 cases.

It is quite interesting that out of 4,881 contexts of usage of the PC *on the threshold of* in the parallel corpus OPUS-2, it is most frequently used with the noun "millennium" -976 cases. The first three high frequency co-occurrence nouns of the PC *on the threshold of* are: *millennium, century* and *era*, which

constitutes 70% of all 25 nouns being examined. Other frequently used nouns are: *reforms, changes, success, etc.*

Analysis of representative empirical data allows us to conclude that the phraseological constructions *on the brink of* / *на грани* and *on the threshold of / на пороге* constitute two groups, in which:

1). Interchange of the PCs in the context is impossible or restricted.

2). Interchange of the PCs in the context is possible.

Let us analyze both groups.

4.1. Interchange of the PCs in the Context Is Impossible or Restricted

In the majority of cases, the PCs *on the brink of* / *на грани* and *on the threshold of* / *на пороге* are not fully synonymous, and therefore, they are not interchangeable. The frequency data from Sketch Engine confirms our observations that *on the brink of* /*на грани* is used with abstract nouns, which expresses negative connotations, and *on the threshold* / *на пороге* is also used with abstract nouns but has a comparably "friendly" surrounding in its context. If concrete nouns are used, which seldom occurs, they are used metaphorically or metonymically in an abstract sense.

Empirical data from corpora shows that most idioms can be translated correctly only if we take the context into account, something that many dictionaries fail to do in a systematic way. As has been mentioned already, the majority of dictionaries postulate a relationship of "full equivalence" between the constructions *on the brink of* / *на грани* and *on the threshold of* / *на пороге.* However, "traditional description… ignores the absence of functional interchangeability between the idioms" (Dobrovol'skij, 2014, p. 872). The following examples for their usage in the parallel subcorpus of the Russian National Corpus in (1) and (2) demonstrate that they are not synonymous, which means that they are not interchangeable in the context, cf.:

(1) Новая школа строилась *на* самом *пороге* века (RNC).

(2) The new school was built *on the threshold of* this century (RNC).

(3) You are *on the threshold of* life, you have only known this girl two months and however deeply you think you love her, I appeal to you to break it off at once (RNC).

(4) Ты стоишь *на пороге* жизни, ты только два месяца знаком с этой девушкой, и какой глубокой ни представляется тебе, твоя любовь к ней, я обращаюсь к тебе с призывом пресечь эту любовь немедленно (RNC).

In contexts (1, 2) and (3, 4), we cannot translate the Russian PC *на пороге* with the English *on the brink of*, nor can we substitute *на пороге* by *на грани*, or *on the threshold* of by *on the brink of.*

It is true that most idioms can be translated correctly only if we take the context into account. However, in some cases, context is not enough, and only culture-specific information about language can help us give a proper interpretation of this or that phrase, cf.:

(5) MEXICAN CULTURAL INSTITUTE "Hina/Jaina: *On the Threshold* of the Mayan Underworld (600-900 A.D.)" [enTenTen13].

In this example, the symbolic meaning of the concept THRESHOLD is taken into account for a better understanding of the PC. According to the Dictionary of symbols, "it symbolizes the potential of friendship, marriage or reconciliation. This potential can be actualized if the individual who comes is greeted at the threshold and invited in" (Chevalier, 1996, p. 997). Thus, THRESHOLD is the symbol of friendship and welcome.

In (5), we are warmly welcomed and invited into Mayan civilization. Here, we cannot substitute the construction *on the threshold of* by the construction *on the brink of* because the noun *brink* has different image components and symbolism. Its image component is based on two ancient oppositions: "us - them" and "life - death". It reflects ancient mythological ideas about boundaries in terms of spatial limits, which are divided into two worlds: "us" - "the world of the living," and "them"- "the world of the dead", and thus, that which is "their" space is threatening "us", which is why it is dangerous.

In (6) the spatial metaphor creates an image of an extreme situation which is recognized as potentially dangerous. In (6) someone had "to stop at the act" due to unpredictable and, most likely, negative

consequences of doing something, cf.:

(6) We must assume, I think, that the forward projection of what imagination he had, stopped at the act, *on the brink of* all its horrible consequences (RNC).

The constructions *on the brink of* and *on the threshold of* are not interchangeable. For instance, we cannot say "on the threshold of horrible consequences" in this particular or any other context. The examples analyzed demonstrate that although PCs are presented as synonyms in almost all dictionaries, in reality, they are not synonymous because each of them has its individual characteristics.

4.2. Interchange of the PCs in the Context Is Possible

Despite the fact that many dictionaries and Sketch Engine graphs treat the phrase *on the brink of* as having negative connotations, this is by no means always the case. The analysis of corpora demonstrates that *on the brink of* in many cases has positive meanings and can sometimes be a semantic synonym to the PC on *the threshold of*. Cf.:

(7) Remember, you are *on the brink of a New Millennium*. Let yourself be inspired, and talk about your inspiration to those who share your life [enTenTen 2013].

(8) Maldives is posed *on the threshold of the New Millennium*, looking forward to it with vigor and enthusiasm. Emboldened with successes of the past and empowered by hope for the future, the Maldives is optimistic in making further strides towards socio-economic development of the country [enTenTen 2013].

Some more examples of the PC *on the brink of* having positive meaning from the British National Corpus BNC (BNC) (9), (10), cf.:

(9) Mr. Ashdown told supporters: "The prize is within our grasp. We stand *on the brink of* an outstanding result" (BNC).

(10) Another Heineken weekend of records and dazzling rugby has convinced fans that Wales is *on the brink of* a new golden era of backplay (BNC).

In many contexts (11, 12) *on the brink of* and *on the threshold of* are interchangeable, cf.:

(11) Goldstein concludes that a parallel growth slowdown in China and the United States, along with deterioration in global financial conditions linked to a disorderly correction of global payment imbalances, could put a group of emerging markets *on the threshold of economic crisis* [enTenTen 2013].

(12) The conflict between the two states placed them both *on the brink of economic crisis*, as South Sudan stopped producing oil and sending it through Sudanese pipelines, accusing the northern state of stealing [enTenTen 2013].

We referred to parallel English-Russian subcorpus OPUS-2 with the purpose of revealing the most frequently used variants of PCs *on the brink of* and *on the threshold of* and their respective translations. The results of the statistical analysis are presented in Tables 1 and 2.

Table 1. Frequency of PCs *on the brink of*/ *на грани* and their variants in OPUS-2

on the brink of	**Q**	**%**	**на грани**	**Q**	**%**
"0" equivalent	25	19%	"0" equivalent	431	57%
на грани	80	61%	on the verge of	181	24%
на пороге	20	15.3%	on the brink of	92	12%
на краю	5	4.7%	on the point of	17	2.2%
			on the edge of	16	2.1%
			on the cusp of	9	1.2%
			on the fringe(s) of	6	0.8%
			on the precipice of	3	0.4%
Total	130	100%		755	100%

Note. Q - Quantity of occurrences.

Table 2. Frequency of PCs *on the threshold/ на пороге* and their variants in OPUS-2

on the threshold	**Q**	**%**	**на пороге**	**Q**	**%**
"0" equivalent	20	17.1%	"0" equivalent	354	30.5%
			literal meaning	342	29.5%
на пороге	79	67.5%	on the threshold of	357	30.8%
в преддверии	9	7.6%	on the verge of	26	22.4%
на рубеже	4	3.4%	on the brink of	24	21%
на границе	3	2.5%	on the doorstep of	21	18.1%
на заре	1	0.95%	on/at the point of	9	0.8%
вот-вот	1	0.95%	on the edge of	9	0.8%
			on the cusp of	7	0.6%
			on the eve of	6	0.5%
			at the gates of	3	0.25%
Total	117	100%		1158	100%

These frequency graphs were processed manually to avoid information noise.

The results of the statistical analysis in the Table 1 show 130 Russian correlates of the PC *on the brink of*: zero equivalent <25>; *на грани* <80>; *на пороге* <20>; *на краю* <5>.

The results of a dictionary analysis indicate that neither general bilingual nor phraseological dictionaries translate the construction *on the brink of* with *на пороге* in their entry, despite the fact that it has a high level of frequency in OPUS-2 and constitutes 15.3% of all its correlates.

The results of statistical analysis in Table 2 show 117 Russian correlates of the PC *on the threshold of*: zero equivalent <20>; *на пороге* <79>; *в преддверии* <9>; *на рубеже* <4>; *на границе* <3>. Two equivalents: *на заре* and *вот-вот* are used only once each. The data in Table 2 indicates that the Russian PC *в преддверии* is used in 7.6% of cases of all its correlated usages. However, dictionaries don't provide such translation for the PC *on the threshold of*.

The evidence suggests that the Russian PCs *на пороге* (Table 1) and *в преддверии* (Table 2) should be included in the dictionary entries of the English PCs *on the brink of* and *on the threshold of* accordingly, since they prove to be more frequent than other PCs which are included in dictionaries, cf.: *на краю, на границе, на рубеже, на заре* (Tables 1, 2).

The frequency graphs demonstrate vividly that there is a great range of variants actually represented in texts which are missed out in the dictionaries. For example, to name but a few of them: *on the cusp of, on the fringe of, on the precipice of*, *on the doorstep of, on the eve of, at the gates of, at the dawn of, at the start of, at the turn of, at the beginning of, at the onset of, at the outset of* and some others. These variants should definitely be included in the dictionary entries of phraseological constructions *on the brink of* and *on the threshold of*.

5. Conclusions

The use of corpora in a dictionary making practice presents a lot of advantages for a lexicographer in expanding the available illustrative materials based on authentic texts. The empirical data presented in the article proves that in the English phrasemes considered here, synonymy is not as complete as it seems at first glance. In many cases, the PCs *on the brink of* and *on the threshold of* are semantically asymmetrical. Often, the generally accepted equivalent of a PC cannot always be used to translate authentic texts, and there are often significant restrictions or impossibilities regarding its substitution with its near-synonyms. However, in some cases, context is not enough, and only culture-specific information about language can help us give proper interpretation of this or that phrase.

Parallel corpus analysis allows us to reveal the full diversity of variants actually represented in texts, which is practically impossible using only monolingual and bilingual dictionaries. Using corpora broadens a phraseographer's resources while arranging the illustrative part of the dictionary entry, and searching translation correlates each unit of the source language under consideration.

References

Birikh, A., Mokienko, V., & Stepanova, L. (2005). *Russian Phraseology: historical etymological dictionary*. Moscow: Astrel AST.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press Publ. https://doi.org/10.1017/CBO9780511750526

Chevalier, J., & Gheerbrant, J. (1996). *Dictionary of symbols*. London: Penguin Books.

Dobrovol'skij, D. (2013). German-Russian idioms online: on a new corpus-based dictionary. In *Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International "Dialogue" Conference* (Bekasovo, May 29-June 2, 2013. Issue 12(19), 210-217). Moscow: RGGU.

Dobrovol'skij, D. (2014). The use of corpora in bilingual lexicography. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: the User in Focus* (pp. 867-884). Bolzana/Bozen: EURAC Research.

Dobrovol'skij, D., & Pöppel, L. (2016). Diskursivnaja konstruktsija N v tom, što i eyo paralleli v drugih jazykah: kontrastivnoje korpusnoje issledovanije. *Vestnik Novosibirskogo Gosudarstvennogo Pedagogicheskogo Universiteta, 6*(34), 164-175.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions. The case of let alone. *Language, 2012, 22*(1), 41-78.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to argument structure.* Chicago and London: The University of Chicago Press Publ.

Goldberg, A. (2006). *Constructions at Work*. Oxford: Oxford University Press.

Kveselevich, D. (2002). *Sovremennyj russko-anglijskij frazeologicheskij slovar.* M.: Astrel.

Longman Group. (1979). *Longman Dictionary of English Idioms*. Harlow and London: Longman Group Limited.

Lubenskaja, S. (2004). *Bol'šoj russko-anglijskij fraseologicheskij slovar'*. M.: Ast-PressKniga.

Piirainen, E. (1997). Da kann man nur die Hände in den Schloß legen. Zur Problematik der Falschen Freunde in niederländischen und Deutschen Phraseologismen. In Nominationsforschung im Deutschen. In I. Barz & V. Schröder (Eds.), *Festschrift für Wolfgang Fleischer zum 75 Geburtstag* (pp. 201-211). Frankfurt/M: Peter Lang.

Stefanovich, A., & Gries, S. (2003). Constructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209-243. https://doi.org/10.1075/ijcl.8.2.03ste