A Corpus-based Approach to Lexicography: Towards a Thesaurus of English Idioms

Guzel Gizatova
Kazan State Agrarian University
e-mail: guzelgizatova@hotmail.com

Abstract

This paper deals with the principles of constructing a new ideographic dictionary of English idioms (DEI) based on corpus data. Idioms are arranged by their figurative meaning rather than alphabetically. The purpose of this paper is to explore English idioms along two revealing lines of inquiry. The first line of inquiry suggests constructing an ideographic dictionary (thesaurus) of English idioms; the second one illustrates advantages of organizing such a dictionary on an analysis of corpus data.

The need for a new dictionary of English idioms is motivated by the fact that at present there is no corpus-based dictionary of English idioms built on the thesaural principle. Ideographic description of idioms enables the reader to find the biggest possible amount of idiomatic word combinations of the language that expresses the given concept.

Due to corpus data, the dictionary presents a range of syntactic patterns, polysemous idioms and unexpected variants which cannot be retrieved from the existing idiomatic and monolingual dictionaries of the English language. Many dictionaries fail to register all meanings of idioms. The corpora help reveal the majority of meanings of polysemous idioms. Apart from its theoretical relevance as an instrument of description of the mental lexicon, a new ideographic dictionary can be used for the purposes of translation and language acquisition.

Keywords: thesaurus**;** corpus-based; English idioms

1 English Idiom-Thesaurus: State of the Art

A work on a project of compiling an ideographic dictionary of English idioms started in 2007. It was inspired by the publication of the *Thesaurus of Present-day Russian Idioms* by Baranov and Dobrovol'skij (2007). Since then the latter has been the first and only dictionary of ideographic description of Russian idioms in international lexicography. My interest to ideographic organization of a dictionary of English idioms was motivated by the fact that at present there is no *corpus-based* dictionary of English idioms built on the *thesaural principle*[1] and exclusively authentic examples. Moreover, there is no *comprehensive* ideographic description of English idioms in the international lexicography. "Thesaurus is a dictionary presenting linguistic information in the direction from 'concept to sign' and explicating relations between the entry concepts" (Dobrovol'skij 1994: 264). Ideographic description of idioms based upon the principle 'from concept to sign' enables the reader

[1] It should be mentioned here that COBUILD Dictionary of Idioms arranged on alphabetical principle is based on all authentic examples drawn from The Bank of English (Sinclair, Fox, Moon 1995); illustrative sentences in Oxford Dictionary of Current Idiomatic English (Cowie, Mackin & McCaig 1984) are mainly *citations* from written or spoken texts and and there are *made-up* examples as well.

to find the biggest possible amount of idiomatic word combinations of the language that express the given concept.

There are two varieties of thematically organized dictionaries of the English language[2]. The first type of references has etymological orientation (Korach 2001; Ostler 2008). Their purpose is to introduce the reader to the fascinating stories behind America's favourite idiomatic expressions. These books make a significant contribution to our knowledge about American history, culture, and everyday life. The aim of the second type of thematic dictionaries is to provide readers with the entire set of idioms organized by common themes and concepts so that they are easy to find and would facilitate an active grasp of the language. These academic dictionaries of American idioms (Spears 1997; Brenner 2003) are based on profound theoretical knowledge and present a perfect tool for scholars, writers and students of English to locate idioms quickly and use the books as a source of reference or a handbook for studying idiomatic expressions. NTC's Thematic dictionary of American idioms (Spears) and American Idioms Handbook (Brenner) are arranged by theme, topic or meaning and contain 5, 500 and 3,500 idioms respectively.

However, these dictionaries have certain shortcomings: the illustrative examples are inconsistent and unpersuasive and most of them do not seem to be authentic. Moreover, in some cases idioms are presented as a set of synonyms series under one topic whereas according to corpora data they have semantic mismatches or connotation differences. That is why the need for a new comprehensive thesaurus of English and American idioms based on authentic examples of idioms drawn from British and American corpora is obvious.

2 Theoretical Concept

The research is based on the main principles of cognitive linguistics and, primarily, on the system organization of structuring of semantic fields. It is assumed that ideographic classifications of different languages *basically* coincide and conceptual sphere covering phraseology of different languages in principle is the same, i.e. extralinguistic, that can be interpreted as a conceptual universal (Dobrovol'skij 1992: 280). However, every language has its own unique semantic structure. Each semantic field segments objective reality in a way, which is specific only to a given language. Moreover, certain linguistic changes within the language belong to the sphere of initial concepts, which have specific linguistic differences in other languages.

Following the principles of the Conventional Figurative Language Theory (CFLT) developed by Dobrovol'skij & Piirainen, we use the term *idiom* in the European tradition of phraseology research and rely on their definition of an idiom as:

> …phrasemes with a high degree of idiomaticity and stability. In other words idioms must be fixed in their lexical structure (however, this does not exclude a certain limited variation), and they must be, at the same time, semantically reinterpreted units (i.e. they do not point to the target concept directly but via a source concept) and/or semantically opaque (2005: 40).

3 Idiom Classification: Analysis and Description

On the preliminary stage of research (in 2007) my database contained about 2000 English and American idioms and by now it comprises about 3700 units and 18000 contexts of idiom use. During

[2] According to the results of our research all existing thematically organized dictionaries were compiled by American scholars on examples of American English idioms. This fact is a motivation for research of the idioms of the British English as well with the purpose of including them in the thesaurus using the illustrative examples from the rich British corpora.

the *first stage,* idioms were drawn from all kinds of monolingual and bilingual idiomatic dictionaries. The thesaurus is based on an inductive method, that is – from idioms to semantic fields and not from an abstract logical outline to idioms. At the first stage, following the strategy of Dobrovol'skij (1994: 264), each idiom under consideration was appended by a certain descriptor (marker). After that, idioms were classified on the basis of their semantic description. Idioms of the same conceptual field were organized under a taxon that is the basic unit of the thesaural representation of idioms and therefore is the main entry-form of the dictionary labelled by a relevant descriptor. For example, a conceptual field *Freedom – Lack of Freedom* has two sub-taxa: 1. *Freedom, No Limitations* and 2. *No Freedom; Limitations, Restrictions.* The second sub-taxon has its own two sub-taxa: 2.1. *Submission* and 2.2. *Violence.* However, besides hierarchical links there are also paradigmatic (horizontal) links in the taxon structure. The zone of paradigmatic references is presented by the sign →. Thus, idioms of the semantic field *Control* are connected semantically with the idea of *freedom* and *lack of freedom*, that is why the reference from the taxon *Control* → is applied to the taxon → *Freedom – Lack of Freedom*. The representation of idioms of this conceptual field coincides in Russian Thesaurus (Baranov & Dobrovol'skij 2007) and in our DEI. This isomorphism can be regarded as conceptual universal. As well as any other subsystem of language, the phraseological system also reveals universal features, as, according to Jacobson, "the languages of the world can actually be approached as manifold variations of one world-wide theme – human language" (Jakobson 1966: 264). However, each semantic field segments objective reality it reflects in a way inherent only to a given language. For instance, my empirical material proves that segmentation of the conceptual field '*Society*' demonstrates the unique semantic structure of the English language. Thus, the semantic field '*Society*' in the Russian Thesaurus (2007) is displayed in a syncretic alliance with the concepts 'Power', 'State', 'Politics', 'Economics'. As far as my data of English idioms is concerned, I assume that the concept '*Society*' should be presented separately as an independent taxon divided into sub-taxa each of which will be included in hierarchical (vertical) links with the taxon and its own sub-taxa[3]. Paradigmatic (horizontal) references are also introduced here. At the moment, the structure of the taxon '*Society*' is the following:

1. Society
   1.1. Political Sphere
      1.1.1. Elections
   1.2. Social Sphere
      1.2.1. National Stereotypes
      1.2.2. Law and Order → Crime
         1.2.2.1. Courtroom → Prison
      1.2.3. Social Conventions
         1.2.3.1. POLITENESS AND MANNERS → Rudeness
         1.2.3.2. SOCIALLY ACCEPTABLE
         1.2.3.3. SOCIAL DISTANCE; FORMALITY
         1.2.3.4. SOCIAL INSINCERITY→ Lie, deception

Idioms referring to national stereotypes represent a typical phenomenon for English phraseology.

[3] I included only those semantic categories that do not enter the Russian thesaurus. This article shows only the process of my work on the dictionary now. The material and structure of many taxa and sub-taxa will be reviewed and changed during the period of further updating of empirical material (the second stage of research).

Thus, comparing Russian and English idioms motivated by national stereotypes, we witness that the quantity of English idioms significantly exceeds the Russian ones referring to ethnic nominations. Indeed, the authors of the Russian Thesaurus Baranov & Dobrovol'skij (2007) give only three idioms referring to national stereotypes: *Russian Ivan, Russian Vanya* (both in the meaning 'common person, a simpleton') and *Uncle Sam*. As for the English idioms, empirical material of our DEI has 184 idioms referring to national stereotypes with authentic examples from different corpora which are not presented in regular English and American idiomatic dictionaries.

The results of research demonstrate that productivity of some semantic fields in the English language differs from those in the Russian language presented in (Baranov & Dobrovol'skij 2007). For example, such semantic fields as *Sport, Gardening, Hunting, Politeness, Law, Privacy, Restraint, Modesty, Shyness* proved to be productive in the English language. Moreover, they seem to be unproductive in Russian phraseology since they are not included in the *Thesaurus of present-day Russian idioms* (Baranov and D. Dobrovol'skij 2007).

The *second stage* of research work involved checking the correctness of the conceptual marking of the idioms under study, which is of primary importance for regrouping the existing stock of idioms.

The *third stage* involved the search of the idioms in corpora with the purpose of verification of the usage of idioms in contemporary discourse. In many cases information about idioms in corpora differed from that in the dictionaries and I had to make some changes, consequently, some idioms fell under different taxa of the thesaurus. For example, traditionally the definition of an idiom *right and left* in the regular dictionaries is: "to both sides, everywhere; in or from every direction". However, comprehensive study of the corpora and retrieval of authentic examples allowed to register three additional meanings of the idiom: 1) not to take into account interests of other people; 2) without a twinge of conscience; 3) anyhow, at haphazard. As a result, this idiom now falls under four taxa in the dictionary: 1) space, place → *everywhere*; 2) morality → *lack of conscience*; 3) traits of character → *egoism*; 4) order → *lack of order*. Each idiom in the dictionary is illustrated by authentic examples drawn from corpora.

The data was retrieved from several corpora: 1) The Bank of English Corpus, a 520-million word corpus of British, North American, and Australian texts, including both written and spoken language; 2) The Corpus of Contemporary American English (COCA) containing 520 million words, equally divided among spoken, fiction, popular magazines, newspapers, and academic texts; 3) British National Corpus (BNC), a 100+ million word collection of samples of written and spoken language originally created by Oxford University Press in the 1980s - early 1990s, and now having its various versions on the web; 4) Corpus of Historical American English (COHA), 1.6 billion words in 7.6 million speeches from the British Parliament, the largest structured corpus of British English for the time period between 1803-2005; 5) The Magazine Corpus of American English, 100 million words of text of American English from 1923 to the present day, as found in Time magazine.

The main important difference of the thesaurus (DEI) from existing English and American idiom dictionaries is in its orientation on modern authentic data drawn from the text corpora, and in some cases, from the English-language Internet. Only those idioms, found in academic, spoken, fiction, newspaper and magazine texts from the 1960-s up to the present day, were included in the thesaurus. Corpus data, obtained in the process of research into the corpora, give evidence of the new perspectives about co-occurrence patterns, semantic and combinatorial properties of idioms and their frequency in various types of corpora (e.g. media versus literary or academic prose). The results of corpus analysis provide an opportunity to reveal new aspects of meanings of English idioms, which have not been yet recorded in dictionaries as well as some cases of their inaccurate definitions. The findings indicate that idioms can only be fully understood if they are considered in a larger context in which they occur.

The entry of the dictionary includes a lemma and an example of idiom use. There is a zone of comments in cases of wordplay and etymological description. One of the peculiarities of idioms is their irregular use, which often manifests itself in variation of an idiom form in the contexts of wordplay. Live and rich image component fixed in the inner form of an idiom is a powerful factor facilitating its irregular behavior, cf. the parodies of well-known Anglo-American proverbs, (1-4) drawnfrom Mieder & Litovkina (2002: 7; 29; 192):

(1) *A barking dog never bites, but a lot of dogs don't know this proverb.*
(2) *A bird in the hand is bad table manners*.
(3) *Absinthe makes the heart grow fonder*.
(4) *The early worm gets eaten by a bird*.

It is essential to include a wordplay in the zone of comments because it allows to define the standard form of an idiom and to register some specifics of the English language that could have remained unnoticed. Moreover, when such idioms continue to be registered in dictionaries "[…] people continue to invent statements based on this structural pattern to parody traditional wisdom" (Dobrovol'skij & Piirainen 2005: 2). As already mentioned, this research is based on the principles of the Conventional Figurative Language Theory (CFLT) developed by Dobrovol'skij & Piirainen where the role of cultural knowledge in idiom motivation is put forward. Idioms have a complex nature that many scholars have tried to clarify; they pose particular problems in language learning and teaching. Opaque idioms are not easily recognized by students so it is crucial to highlight the importance of etymology in their understanding.

According to Dobrovol'skij & Piirainen:

> …research on motivation of figurative units cannot but include etymological description as a constituent part. This does not mean that etymology always influences actual meanings and brings about relevant usage restrictions, but it cannot be excluded a priori (2005: 82).

Referring to the same issue, G. Lakoff points out:

> …the human lexicon will have to take account on the phenomenon of folk etymology – that is, it will have to include an account why expressions with motivating links are easier to learn and remember than random pairings (Lakoff 1987:452).

Let us consider an example:

(5) *Throw your hat in the ring* 'announce your intention to run for office'.

The etymology of this idiom goes back to 1912 when Theodore Roosevelt announced that he intended to run for president against William Taft: "My hat's in the ring. The fight is on and I'm stripped to the buff". Roosevelt was making figurative use of a boxing term that had been around since at least the early 19th century. The phrase commonly appeared in sports magazines that followed boxing matches as in this quote from an 1818 issue of *Sporting Magazine*: "Johnson […] threw his hat in the ring." Professional boxers at this time made money by traveling to fairs, men's clubs, and other venues and challenging all comers. A man who wanted to take up the challenge signaled his interest by throwing his hat into the boxing ring. No doubt, this method was adopted because boxing matches, then, as now, were crowded, noisy events. Just shouting or waving would not be enough to attract attention. Nowadays nearly all candidates for political office are described in the newspapers as having thrown their hats into the ring (Ostler 2008: 89-90).

Since its first use, the idiom has become an active practice to be handled in presidential electoral campaigns. The idiom is used today in 2016 Presidential race in the USA.

(6) Property tycoon Donald Trump, one of America's most flamboyant and outspoken billionaires, *threw his hat* into the 2016 race Tuesday for the White House, promising to make America great again (http://bhcourier.com/donald-trump-throws-his-hat-in-the-ring/).

(7) Hillary Clinton *Throws Hat in Ring* for 2016 Presidential Race
'Everyday Americans need a champion. I want to be that champion,' she said in an Internet video announcing her run (http://www.israelnationalnews.com/News/News.aspx/193945#.VyOhFlSzB0w)

Including etymological information of an idiom in its dictionary description is essential because its "etymological memory" is an important part of its semantics and provides additional information on its actual meaning helping understand the idiom.

4 Conclusion

The research based on theoretical concepts developed by Baranov, Dobrovol'skij & Piirainen enabled us to apply their strategy to construction of a thesaurus of English idioms. The results of analysis demonstrate that many semantic fields productive in the English phraseology are unproductive in Russian phraseology. The use of English and American linguistic material which is the subject of lexicographic description, demonstrates that the structure of taxa and their interaction within the English semantic network differs from the Russian semantic network. This proves the idea that conceptual sphere covering phraseology of different languages is *basically* the same, that can be interpreted as a conceptual universal. However, every language has its own unique semantic structure and each semantic field segments objective reality in a way, which is specific only to a given language.

Due to corpus data, the dictionary presents a range of syntactic patterns, polysemous idioms and unexpected variants which cannot be retrieved from the existing idiomatic, bilingual and monolingual dictionaries of the English language. Apart from its theoretical relevance as an instrument of description of the mental lexicon, a new ideographic dictionary of English idioms based on corpus data can be used for purposes of language acquisition and translation. Most set phrases can be translated correctly only if we take the context into account, something that many dictionaries fail to do in a systematic way. The compilation of a corpus-based thesaurus of English and American idioms based on authentic data is a question of vital importance for modern theoretical phraseology and practical lexicography.

References

Arutz Sheva Israeli Media Network Online. http://www.israelnationalnews.com-/News/News.aspx/193945#.VyOhFlSzB0w. Accessed [30.04.2016].

Baranov, A., Dobrovol'skij D.(2007). *Thesaurus of Present-day Russian Idioms.* – Moscow: Avanta.

Beverly Hills Courier Newspaper Online. http://bhcourier.com/donald-trump-throws-his--hat-in-the-ring Accessed [30/04/2016].

Brenner, G. (2003). Webster's New World American Idioms Handbook. – USA: Wiley Publishing Inc.

Dobrovol’skij, D. (1992). Phraseological universals: theoretical and applied aspects. – In: *Meaning and grammar. Cross-linguistic perspectives*. – Berlin – New York: Mouton de Gruyter, pp. 279-301.

Dobrovol'skij D. O. (1994). Idioms in a Semantic Network: Towards a New Dictionary-Type. In: *Proceedings of the 6th EURALEX International Congress*, Amsterdam, pp.263-270.

Dobrovol’skij, D., Piirainen, E. (2005). Figurative language. Cross-cultural and cross-linguistic perspectives. – Amsterdam: Elsevier.

Korach, M. (2001). Common Phrases and where they come from. – Guildford, Connecticut: The Lyons Press.

Lakoff, G. (1987). Women, Fire and Dangerous Things. – Chicago and London: Chicago University Press.

Mieder, W., Litovkina, A. (2002). Twisted Wisdom. Modern Anti-proverbs. - Hobart, Australia: De Proverbio.

Ostler, R.(2008). Let’s Talk Turkey. – New York: Prometheus Books.

Spears, R. (1997). NTC’s Thematic Dictionary of American Idioms. – USA: NTC Publishing Group.