EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021
Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

Speech Etymology Lemma Idioms Glossary NLP Corpora Dictionary Meaning Word Lexicon Definition Pronunciation Headword Examples Entry Lexicology Dictionary Use Lexical Resources Neologisms

2020 Edition

Semantic Relations in the Thesaurus of English Idioms: A Corpus-based Study

Giztova G.[1], Ismagilova L.[2]

[1] *Kazan Federal University, Russia*
[2] *Kazan Federal University, Russia*

Abstract

This paper deals with the principles of constructing an Ideographic Dictionary of English Idioms (Thesaurus) based on corpus data. Idioms in the dictionary are arranged by their figurative meaning rather than alphabetically. The need for a new type of dictionary is motivated by the fact that at present there is no corpus-based dictionary of English idioms built on a thesaural principle. Ideographic description of idioms enables a reader to find the largest possible number of idiomatic word combinations of the language that express a given concept. The basic entry of the Thesaurus is called a taxon, consisting of a conceptual descriptor used as a label of a taxon, and a group of idioms expressing the respective taxon. English Web text corpus 2013 (enTenTen13) is used as an empirical basis of the study. The analysis of corpus data presents a range of syntactic patterns, idiom variation, synonymous and polysemous idioms which cannot be retrieved from the existing idiomatic and monolingual dictionaries of the English language, since they fail to register all meanings of an idiom. Today, as lexicography is experiencing "the corpus revolution" (Hanks 2012), this is a question of key importance. The use of corpora provides additional possibilities for compiling the idiom list and structuring entries.

Keywords: thesaurus, idioms, corpus, variation, synonymy

1   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 that 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. The thesaurus is based on an inductive method, that is – from idioms to semantic fields and not from an abstract logical outline to idioms.

Following the principles of the Conventional Figurative Language Theory (CELT) developed by Dobrovol'skij & Piirainen (2005), 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 (Dobrovol'skij & Piirainen 2005:40).

Our work on the English Thesaurus was inspired by publishing of the *Thesaurus of Present-Day Russian Idioms* by Baranov and Dobrovol'skij (2007). It can be called a "lexicographic giant" and had been the first and the only profound, comprehensive and up to date ideographic dictionary of Russian idioms in international lexicography since then. Such types of dictionaries have not been developed in other languages as well. A thorough outline of existing English thematic dictionaries (not ideographic ones, since they do not exist) of idioms is given in (Gizatova 2016). Significance of ideographic dictionaries of idioms lies in their ability to reflect the "naïve" pre-scientific picture of the world, in the study of the history of human knowledge about the world surrounding us, in comparative study of "natural" world picture of different nations. Despite the importance of these issues, few studies have focused on these aspects, so practical lack of research and lexicographic accomplishments in this field motivated our interest to start our work on compiling a corpus-based dictionary of English idioms built on thesaural principle and exclusively authentic examples.

2   Discussion

Lexicographers encounter certain problems in connection with treatment of idioms in dictionaries. In this article the authors approach the issue of idioms in the context of organizing semantic fields of the Thesaurus. High level of variation of idiom components is a typical feature of phraseological system of languages. Very often it is difficult to make a decision whether it is a variant or a synonym of an idiom. The most important conceptual criterion in this case is the identity of inner forms of the idioms under consideration. If the inner form of an idiom is not changed in the case of its component modification, the decision is taken in favour of variants (*shake/quiver/quake in one's boots/shoes*). If the

image components underlying the idioms are different, it is natural to speak of a synonymy (*spill the beans/ let the cat out of the bag*). This leads to the reflection of semantic differences of idioms in structuring of semantic fields of the Thesaurus.

Traditionally, idiomatic variants are units in which partial difference in their constituents does not alter their meaning. In contrast to idiomatic synonyms which are interchangeable only in definite contexts, substitution between idiom variants is possible in any condition of usage. Let us have a closer look at idiom variants:

(1) *to have/bear a grudge* "to have an old resentment for someone; to be continually angry at someone".

The idiom *to have a grudge* is traditionally considered to be more formal than t*o bear a grudge*, the latter being more frequently used in poetry and fiction. But analysis of the corpus material does not prove this notion. It is sufficient to look at their contextual use in the following examples and it is evident that both idioms are used in spoken syntactic constructions, cf.:

(2) But possibly the worst job I ever had was possibly to the worst boss I ever had. Marjorie. I still remember you. Marjorie was my middle-aged upline when I was 17 years old Office Junior. And she *had a grudge* against all humanity, particularly against friendly teenage girls, it turned out (enTenTen13).

(3) There were the key words which only the adults could understand. I grew up and came to understand, and I *bore a grudge* against all men… All of them were treacherous (enTenTen13).

These two idioms have some specific characteristics: both of them are usually used in the construction with preposition *against*, at the same time, the results of corpus analysis show that *to have a grudge* is used almost twice as much with the preposition *against* as the idiom *to bear a grudge* (72% to 38% of occurrences respectively). The idioms also differ in the aspect of their tendency to be used in negative sentences. Thus, *to bear a grudge* can take negative form in 17% of its use, whereas *to have a grudge* practically does not take the negative form (only in 0.7%). These characteristics of the two idioms are of structural nature and refer to their inner organization. Both idioms have identical meanings and functions. That is why we consider them to be idiomatic variants. At the same time, we think it essential to indicate these insignificant structural differences in idioms in this paper. It is important because the specifics of inner organization of an idiom can lead to generation of semantic and other connotations and consequently play a decisive role in transformation of variants into synonyms.

As for arranging idiomatic variants in the Thesaurus taxon, it is important to define which of the variants is the base-form and which is the idiom-variant. According to our data, retrieved from the corpus, the frequency of the idiom *to have a grudge* is twice as high as that of *to bear a grudge* (1456 to 734 occurrences respectively). So, the idiom *to have a grudge* which is more frequently used in the language is the base-form and the less frequent one *to bear a grudge* is the idiom-variant, which follows the base-form after slash, cf.: *to have/bear a grudge*.

As has been mentioned above, one of the challenges in the dictionary-making process is distinguishing between variants and synonyms of idioms. In some dictionaries a range of idiomatic synonyms is traditionally presented as idiom variants. We will show (relying on corpus data) that some of these idioms have different shades of meaning or stylistic colouring and due to that they are regarded as synonymous idioms. Due to that they can fall under different taxa of the Thesaurus. Cf. two synonyms: *to put an end* to and *to put a stop to*. They have a common first component *to put*, but their second components *an end* and *a stop* are synonymous. These idioms have two meanings: 1) to hold back, to restrain; to put something on hold and 2) to destroy, to do away with. In both cases the idiom *to put an end to* has a shade of a more resolute interference, presupposing a significant force and pressure.

In their first meaning "to hold back" these two idioms do not differ in their combinatorial properties, cf.:

(4) We should stop being hypocrites – preaching in public and indulging in private... We now need movements that will *put an end to* corruption and mismanagement which is draining all our efforts and progress (enTenTen13).

(5) A group of women are sitting and talking under the shade of some large mango trees in Juba, South Sudan – a common sight. But today, instead of talking about the weather, the conversation is taking a different tack. One woman asks about the best way *to put a stop to* corruption in local government. She says she wishes their discussion could be heard by all government officials (enTenTen13).

In their second meaning "to destroy; to do away with" both idioms are used with nouns denoting processes, inanimate objects, abstract nouns, e.g.:

(6) In your insane wars you destroy millions of lives and think nothing of it. I am going *to put a stop to* your wholesale destruction of human beings. I want laughter, not slaughter (enTenTen13).

(7) "Mankind must put an end to war before war *puts an end to* mankind" – John F. Kennedy (enTenTen13).

As noted above, the idiom *to put an end* in comparison with the idiom *to put a stop to* denotes terminating something

involving more resolute pressure and interference. For this reason, it expresses not only an uncompromising determined will of a person to do away with something, but his desire to subdue herewith the will of another person. It is easy to see from corpus examples above that both idioms are used with abstract nouns. But due to corpus data we were able to reveal a significant difference in combinatorial properties of these two idioms. The key issue is that the idiom *to put an end to* in the meaning "to do away with" can be used with the names referring to people due to specific characteristics of its semantic functions described above, cf.:

(8) The right-wing opposition once again erred - Confident in the idea that the world crisis and the fall in oil prices would finally bleed dry the revolution and *put an end to* Chavez (enTenTen13).

Sometimes the ability of an idiom to refer to people is transformed into a new meaning: "to kill", cf.:

(9) The jealous Queen, vowing to rid herself of "Miss goody-goody Snow White" once and for all, sends the girl off into the forest with the Huntsman, bidding him to "*put an end to her*". Despite his fear of the Queen's powers, the Huntsman cannot bring himself to obey her command, and takes the Princess to the home of the seven Dwarfs, who vow to protect her (enTenTen13).

In arranging semantic fields of the English Thesaurus, we follow the basic principle underlying the organization of Thesaurus of Present-day Russian Idioms (Baranov & Dobrovol'skij 2007). According to Dobrovol'skij: "As far as idioms are concerned, it seems appropriate to put together not only synonyms proper, but all idioms belonging to the same conceptual domain without taking into account even their word class identity" (1994: 266). Following this concept, for instance, within the domain *Insanity*, *Silliness* such idioms as *go nuts* (verb), *a mare's nest* (noun), *a few cards shy of a full deck* (adjective), *take leave of one's senses* (verb), *not all there* (adverb) can be organized in one taxon. "The reason for this decision is the vagueness of many idioms, their ability to change the formal parameters depending on the context, and the resulting ineffectiveness of postulating artificially precise distinctions" (Dobrovol'skij 1994: 266).

See below an example-taxon *Punishment* with its conceptual variations:
Punishment: Reproach, Blame, Reprimand, Accusation, Criticism
knock s.o.'s block off; give s.o. hell; chew/bawl s.o. out; haul s.o. over the coals; tear s.o.'s head off; finish s.o. off; put and end to s.o.; skin s.o. alive; raise one's voice; lift one's voice; drag s.o. /s.o.'s name through the mud/ the dirt/ the mire/ the muck; blacken/ sully s.o.'s name; give a telling-off to s.o.; give s.o. a sound thrashing; beat the daylights out of s.o; plow s.o. under; teach s.o. a good lesson etc.

284 contexts having direct connection to the taxon *Punishment* have been found in the corpus and we have displayed here only some of them for reasons of space. Let us consider two phraseological units: *lift one's voice* and *raise one's voice*, often being confused as variants of the same idiom. Their direct meaning is 'to speak louder' (physically); the usage of idioms in authentic texts is presented in the following examples.

(10) *lift one's voice*
Turning to face the others, the Weyrleader *lifts his voice* to be heard by all, not that he has to speak loudly. 'No sense in us wasting energy either' (enTenTen13).
(11) *raise one's voice*
Turning, he *raises his voice* so the others can hear too (enTenTen13).

Their secondary figurative meaning according to the New English-Russian Dictionary is 'to raise a voice of protest against somebody/something' (Apresjan 1993:352; 1994:13). But the search in the corpus allowed us to reveal many important semantic differences between these two idioms. First of all, they differ in their semantic functions: if the idiom (10) has a sense of reason, the other idiom (11) is more emotional and often lacks the sense of reason. Consider examples:

(12) I watched some of the commentary after the speech… They swooned over the greatness of Obama's speech. He *lifted his voice* in authority a couple times, but he stumbled on basic words. He came on exited to Accept the Democratic Nomination, then turned into Just Another Politician (enTenTen13).

(13) "In my entire career in the industry I have never had someone react to an evaluation the way Robert did…" the regional manager would later admit. "It's not unusual for people to be disappointed or defensive when given a poor job performance rating, but Robert immediately began *to raise his voice* … He was shouting at me and telling me I didn't know what I was doing. It took me over an hour to calm him down" (enTenTen13).

The second difference is connected to the secondary figurative meaning of these two idioms: 'to raise a voice of protest against somebody/something'. Out of 273 contexts found in the corpus the idiom *lift one's voice* is used only once in the meaning 'to raise a voice of protest against somebody/something'. In other cases, the idiom is used in its direct meaning 'to speak louder'. Let us consider this single example in the corpus, cf.:

(14) This failure was a fresh and yet mortifying disappointment, and his end was a gloomy and somewhat obscure one, but he will always be remembered with gratitude as one of the first who in the Irish Parliament *lifted his voice* against

those restrictions under which the prosperity of the country lay shackled and all but dead (enTenTen13).
The third distinguishing characteristics of the idiom (10) is that in 70 % of its usage in authentic discourse it is practiced in religious texts and mainly in the meaning: 'to chant in the praise of God', cf.:

(15) Wherever stress there is in preparation for the day melts away during the Shabbat morning service as your child leads the congregation in prayer; chants beautifully from Torah and Haftarah and teaches the congregation with his (or her) d'var torah. Samuel is also a 'sweet singer in Israel' so it was a special treat to hear him *lift his voice* to God (enTenTen13).

Analysis of semantic differences of idioms *lift one's voice* (10) and *raise one's voice* (11) helps us in structuring semantic fields of the Thesaurus. Thus, *lift one's voice* falls under three taxa: 1. *Punishment* with its sub-taxa: 'reproach', 'blame', 'reprimand' or 'accusation'. 2. Religion (chanting). 3. Communication (chanting). The idiom *raise one's voice* falls under several taxa as well. We can raise voice under different circumstances: attracting someone's attention, in the process of argument, in the state of irritation, frustration, anger, annoyance and other emotional conditions. The main outstanding semantic difference between idioms (10) and (11) is that the idiom *to raise one's voice* has the meaning 'to raise a voice of protest against somebody/something', whereas the idiom *to lift one's voice* is used only once in this meaning. So, the idiom (11) falls in our Thesaurus under the following taxa: 1. Protest. 2. Punishment in connection with its sub-taxa 'reproach', 'blame', 'reprimand', 'accusation' or 'blame'. 3. Attention and its sub-taxon 'attracting attention'. 4. Behavior with its sub-taxon 'improper argument behavior'. 5. Emotions with its sub-taxa: 'irritation, annoyance', 'frustration', 'anger' 'grief' etc. Definitely, this is not the full picture of presenting idioms in the semantic network, because besides hierarchical links there are as well paradigmatic (horizontal) links in the taxon structure. The zone of paradigmatic references is presented by the sign →. Thus, idioms of the semantic field *Prison* are connected semantically with the idea of *Punishment,* that is why the reference from the taxon *Punishment* → is applied to the taxon → *Prison.* The taxon organization with all its paradigmatic references is presented in the Table 1 in next section of the paper.

3   Idiom Classification: Results

4.1. Stage 1. Collection of data

At *the first stage* of our work (in 2007) the list of English and American idioms consisted of 2000 units collected manually from fiction, academic journals, popular newspapers and magazines. Now it comprises about 6300 idioms and 11500 contexts of their usage in authentic contexts from the corpora. The process of collection of empirical material continued at *the first stage* of research where idioms were drawn from monolingual, bilingual and phraseological dictionaries. The goal of the first stage was to assign a certain descriptor to each idiom under consideration. This was followed by classification of idioms 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 labeled by a relevant descriptor. For example, the following idioms are organized in the dictionary in one group under the taxon *True-Untrue:*

(1) *to talk turkey* "to discuss something directly and honestly", which is regarded as being True;
(2) *to cook the books* "falsify a company's financial records", which is regarded as being Untrue;
(3) *snow job* "a systematic deception; a deceptive story that tries to hide the truth" which is regarded as being Untrue.

Table 1 below displays taxon *True-Untrue,* which is one of the 82 idiom-thesaurus taxa. The sign (→) indicates paradigmatic references to other taxa of the dictionary. We display the contents of the taxon concisely, presenting only two idiom-examples illustrating each taxon and its sub taxa. In general, our empirical material consists of 234 idioms belonging to this taxon.

True-Untrue	*play the dumb; crocodile tears*
1. Truth – Lie, Deception	1.2.2.2. Pseudo-art
1.1. Truth	*soap opera; cock-and-bull-story*
talk turkey; call a spade a spade	1.2.3. Deceit, Verbal lie, Misrepresentation
1.2 Lie, Deception	*a snow job; full of hot air*
→false status →betrayal	1.2.4. Self-delusion, Illusions
1.2.1. Participants, Instruments of situation of lie and deception	*a mare's nest; build castles in the air*
	1.3. Honesty – Dishonesty
decoy duck; stool pigeon	1.3.1. Honesty, Openness, Sincerity
1.2.2.2. Non-verbal lie	→oath
→theft, stealing	*a square deal; an honest Joe*
draw a red herring; take s.o. for a ride; cook the books	1.3.2. dishonesty, insincerity, hypocricy
1.2.2.1. Pretence	→cunning…→prenence→ immorality
→dishonesty, insincerity, hypocricy	*brown nose; play possum*

Table 1

Big size taxa can fall into sub-taxa, e.g. the conceptual field TRUE-UNTRUE has one sub-taxon TRUTH-LIE, DECEPTION. The concept LIE, DECEPTION in its turn has its own four sub-taxa. Each sub-taxon can break down further, thus the concept NON-VERBAL LIE falls into two sub-taxa: PRETENCE and PSEUDO-ART. In some cases, the depth of such division can be up to five levels. Due to space limitations we cannot provide authentic corpora examples in the Table 1.

4.2. Stage 2. Verification of Conceptual Marking

The second stage of classification is connected with checking the correctness of the conceptual marking of the idioms under study, which is of primary importance for regrouping the existing stock of idioms.

4.3. Stage 3. Idioms in Corpora

At present time the third stage of work is being carried out and it is connected with the search of idioms in corpora with the purpose of verification of idiom usage in contemporary discourse. Since the language changes, in many cases information about idioms in corpora differs from that in the dictionaries and for that reason we had to make some changes in classification which had been done on the previous stages of our work. For example, due to corpus analysis new polysemous idioms have been retrieved from authentic contexts of their usage. As a result, some idioms fell under other taxa of the thesaurus than they had been grouped earlier. We will present advantages of corpora in retrieving new polysemous idioms. To illustrate, majority of dictionaries give two meanings of an idiom *on the cuff*, cf.:
1) 'on credit'
His cleaning lady has volunteered to go *on the cuff* when he explained to her about his cash-flow problems (BNC).

2) 'free of charge'
The press agent gets no pay but only a certain amount of drinks *on the cuff* (BNC).

However, the comprehensive study of corpora and retrieving authentic examples allows registering two additional meanings of the idiom:

3) 'confidentially'
But strictly *on the cuff* I'm willing to bet he never did see it and that he never heard of Mildred … (BNC).

4) 'spontaneous, without previous preparation'
We have a little segment here 'They play it and I say it'. They are going to just pick out some things from the speech and I am going to respond to them *on the cuff* (COCA).

As a result, the idiom *on the cuff* now falls under five taxa of the dictionary:

1) In its first meaning 'on credit' it falls under a taxon Money with its sub-taxon Debts;
2) In its second meaning 'free of charge' it also falls under the same taxon Money, but under different sub-taxon, which is Free of Charge;
3) In its third meaning 'confidentially' the idiom falls under a taxon Mystery, Secrecy;
4) In its fourth meaning 'spontaneous' without previous preparation' the idiom falls under three taxa:
a. Behaviour with its sub-taxon Spontaneity;
b. Time and its sub-taxon Spontaneity;
c. Freedom and its sub-taxon Natural Action in the meaning: 'not forced action', 'an action of free will'.

5. Conclusions

The research based on theoretical concepts developed by Baranov, Dobrovol'skij and Piirainen enabled us to apply their strategy to construction of Thesaurus of English Idioms. Results of the study introduce a new approach to phraseography, that is a thesaural principle of structuring of semantic fields of English idioms.
Our analysis clearly demonstrates that advantages of application corpus approach to lexicographic research are evident. Due to corpus data, the dictionary presents a range of syntactic patterns, idiom variation, synonymous and polysemous idioms which cannot be retrieved from the existing idiomatic bilingual and monolingual dictionaries of the English language. Apart from its theoretic 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 idioms based on authentic data is a question of vital importance for modern theoretical phraseology and practical lexicography.

References

Apresjan J. (1993). New English-Russian Dictionary. Vol.1-2.– Moscow: Russkij yazyk.

Apresjan J. (1994). New English-Russian Dictionary.Vol.3. – Moscow: Russkij yazyk.

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

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. (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.

Gizatova, G. (2016). A corpus-based Approach to Lexicography: Towards a Thesaurus of English Idioms. In: Proceedings of the 17th EURALEX International Congress, Tbilisi, pp. 348-354.