Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

WORD SENSE FREQUENCY OF SIMILAR POLYSEMOUS WORDS IN DIFFERENT LANGUAGES[1]

Iomdin B. L. (iomdin@ruslang.ru)
V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences; National Research University "Higher School of Economics", Moscow, Russia

Lopukhin K. A. (kostia.lopuhin@gmail.com)
Scrapinghub, Moscow, Russia

Lopukhina A. A. (nastya-merk@yandex.ru)
V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

Nosyrev G. V. (grigorij-nosyrev@yandex.ru)
Yandex, Moscow, Russia

When words have several senses, it is important to describe them properly in dictionary (a lexicographic task) and to be able to distinguish them in a given context (a computational linguistics task, WSD). Different senses normally have different frequencies in corpora. We introduced several techniques for determining sense frequency based on dictionary entries matched with data from large corpora. Information about word sense frequency is not only useful for explanatory lexicography and WSD, but it also may enrich language learning resources. Learners of a foreign language who encounter a word similar to one of their native language are often tempted to assume that the foreign word and its equivalent have the same meaning structure. Sometimes, however, this is not the case, and the most frequent sense of a word in one language may be much less frequent for its cognate. We proposed

[1] The research of Boris Iomdin, Konstantin Lopukhin and Anastasiya Lopukhina was supported by RSF (project No. 16-18-02054: Semantic, statistic and psycholinguistic analysis of lexical polysemy as a component of Russian linguistic worldview). Parsing the Dante database of English was done by Grigoriy Nosyrev. The authors would also like to thank Leonid Iomdin and Daria Shavarina as well as the anonymous reviwers for their valuable comments.

a method for detecting such cases. Having selected a set of Russian words included into the Active Dictionary of Russian which have more than two dictionary senses and have cognates in English, we estimated the frequencies for English and Russian senses using SemCor and Russian National Corpus respectively, matched the senses in each pair of words and compared their frequencies. Thus we revealed cases in which the most frequent senses and whole meaning structures are, cross-linguistically, substantially different and studied them in more detail. This technique can be applied not only to cognates, but also to pairs of words which are usually offered by the dictionaries as the translation equivalents of each other.

Key words: semantics, lexicography, polysemy, text corpora, experiments, statistical techniques, frequency, meaning frequency

1. Introduction

Polysemy is well known to be one of the key issues both in theoretical and computational linguistics (see e.g. Pustejovsky, 1996; Apresjan, 2000; Lin and Ahrens, 2005; Agirre and Edmonds, 2007; Kwong, 2012; Hanks, 2013; Iomdin, 2014). We know many things about word senses, but very little about their frequency distribution. Indeed, it is quite difficult to obtain such information. Recently, attention has been drawn to the fact that different senses of a word normally have different frequencies in corpora (see e.g. Iomdin, Lopukhina and Nosyrev, 2014). However, there are very few resources that provide such data.

The question of the most frequent (or predominant) sense (MFS) has been discussed for the purpose of automated word sense disambiguation task (WSD), as MFS is considered to be a very powerful heuristic, which is difficult to overcome for many WSD systems (Ide and Véronis, 1998; Navigli, 2009). Various approaches for acquiring predominant senses have been applied to English. Mohammad and Hirst (2006) make use of the category information in the Macquarie Thesaurus. McCarthy et al. (2007) propose an unsupervised approach for finding the predominant senses using a distributional thesaurus and WordNet (Fellbaum 1998). Bhingardive et al. (2015) compare the embedding of a word with all of its sense embeddings (which are produced using various features of WordNet) and obtained the predominant sense with the highest similarity. For Russian, a pilot study of MFS detection was presented by Loukachevitch and Chetviorkin (2015), who used the Thesaurus of Russian Language (RuThes-lite) to determine the most frequent sense of ambiguous nouns, verbs and adjectives with the help of monosemous multiword expressions that are related to those words. Their results are comparable to the state-of-the-art in this field—the highest accuracy rate reaches 57.4%.

The overall sense distribution of a polysemous word is a question that is rarely put in focus. For English nouns, Lau et al. (2014) proposed a topic modeling-based method of estimating word sense distributions, based on Hierarchical Dirichlet Processes and on word sense induction, probabilistically mapping automatically learned topics to senses in a sense inventory. Some information about English verb pattern frequency distributions can be found in the Pattern Dictionary of English Verbs,

developed by Patrick Hanks and colleagues (http://pdev.org.uk/; Hanks and Pustejovsky, 2005; Hanks, 2008). The authors emphasize that senses are associated with patterns (collocations) and not with words and that the Pattern Dictionary provides information about the relative frequency of phraseological patterns rather than dictionary senses. Cf. also Gries et al. 2010, where frequency distributions of English verbal constructions are discussed.

For Russian, a word sense frequency acquisition method and its evaluation for nouns were presented in (Lopukhina, Lopukhin and Nosyrev, in print; Lopukhina et al., 2016). In these articles we reported on our research of this issue and introduced several techniques for determining sense frequency based on dictionary entries matched with data from large corpora. This method is based on building context representations with semantic vectors (Mikolov et al., 2013) and gives robust frequency estimates with little annotated examples available from dictionaries. Supplied with examples and collocations from the Active Dictionary of Russian (Apresjan et al., 2014), the method achieves frequency estimation error of 11–15% without any additional labeled data. It was used to obtain sense frequencies of 440 polysemous Russian nouns.

2. Meaning frequency and foreign language learning

Sense frequency distributions might be used when learning a foreign language. Language teachers can use these data for organizing senses in bilingual dictionaries, composing basic dictionary lists, preparing specific lexicon tests, etc. This information might be especially helpful for studies of cognates and other pairs of similar words in the native language of the learner and the foreign language. Languages can have many such pairs: cognates, borrowings from one of the languages into the other, international words. On the one hand, such words are easier to learn. On the other hand, learners do not always realise that even very similar polysemic words can have very different meaning structures.

In 1928, Koessler and Derocquigny coined the term “faux amis” for pairs of identical or similar cognate words with different senses, emphasizing the importance of such pairs to be identified and properly described so that translators could avoid using such words incorrectly (Koessler and Derocquigny, 1961). For English and Russian, examples such as English *artist* ‘painter’ vs. Russian *artist* ‘performer’, *English box* ‘container’ vs. Russian *boks* ‘boxing’, English *clay* ‘wet soil’ vs. Russian *klej* ‘glue’ are often given. Ample literature is devoted to such cases in different languages (see e.g. Darbelnet, 1970; Walter, 2002; Szpila, 2006; Vrbinc and Vrbinc, 2014, to name a few). In many cases, cognates in a language pair do have some senses in common, but the meaning structure is different. An important case for language learners is when the most frequent sense of a word in one language turns out to be much less frequent for its cognate.

The new techniques for counting word sense fequency on large corpora provide data for resources where similar words in different languages can be analyzed according to the comparative frequency of their senses. Bilingual comparative lexical entries sorted according to sense frequency will give learners new opportunities for better mastering the lexicon and the differences in using the words of both languages.

3. Data and Method

Our primary word sense inventory and source of materials for Russian was the Active Dictionary of Russian (ADR, see Apresjan et al., 2014). This dictionary was chosen for three reasons. First, it is the most up-to-date and most developed explanatory dictionary of Russian. Second, ADR uses a systematic approach to polysemy. The main unit of the ADR, the lexeme, is a well-established word sense identified by a set of its unique properties (syntactic, semantic and pragmatic features, sets of synonyms, analogues, antonyms and semantic derivatives etc.). Third, for each lexeme the ADR provides many sample phrases and sentences, which contributes to the precision of our word sense frequency counting technique. We selected 66 Russian nouns having phonetically similar counterparts in English. These pairs included authentic cognates of the same Indo-European origin (such as *brat—brother*, *gus'—goose*), English borrowings into Russian (such as *bar* < *bar*, *biznes* < *business*), French borrowings into both English and Russian (such as *anekdot—anecdote*, *batareja—battery*) and words of Latin and Greek origin, mostly borrowed into Russian through French or German (such as *advokat* 'lawyer'—*advocate*, *al'bom—album*, *garmonija—harmony*). Since etymology is not relevant for our purposes, we refer to all word pairs under discussion as cognates for the sake of simplicity.

Sense frequencies of the Russian nouns were estimated automatically by performing word sense disambiguation on contexts sampled from the corpus and then calculating relative sense frequencies in the sample. For the purpose of the current study, we have sampled 1000 random contexts for each word from the domain-neutral Russian National Corpus (RNC, ruscorpora.ru, 230 million tokens in the main corpus).

The method consists of two parts: a context representation technique and a disambiguation method. We used semantic vectors as a basis for context representation and averaged them. We gave more weight to words that occur more frequently with the target word than without it. This weighting allows to capture the most important context words and build a better sense representation. Disambiguation was performed in the following way: for each sense we took all examples and collocations from the ADR and averaged their context vectors, obtaining a sense vector. During disambiguation, the context was assigned to the sense with the closest sense vector using cosine similarity measure.

Word vectors were produced using the word2vec skip-gram model with negaive sampling (Mikolov et al., 2013) from a large corpus (about 2 billion words from RuWac (Sharoff, 2006), lib.ru and Russian Wikipedia) with lemmatization, which is especially important for Russian because of its rich morphology. The dimension of word vectors was set to 300. Context vectors were calculated as a weighted average of word vectors in the fixed window of 10 words before and after the target word, where weights were proportional to PMI of context words.

The method was evaluated on hand-tagged contexts for 20 polysemous words from the ADR. It reaches an average disambiguation accuracy above 75% and a maximum frequency error below 16%. The evaluation showed that our method can perform sense frequency estimation with high accuracy (details in Lopukhina, Lopukhin and Nosyrev, in print; Lopukhin and Lopukhina 2016; Lopukhina et al., 2016).

Sense frequencies of the English counterparts were obtained from the largest sense tagged SemCor 3.0 corpus (Miller et al., 1993). SemCor is composed of 220,000

words taken from the Brown corpus (Francis and Kučera, 1979). Approximately half of the words in this corpus are open-class words (nouns, verbs, adjectives and adverbs) which have been linked to WordNet 3.0 senses (Fellbaum, 1998) by human taggers using a software interface. SemCor (and WordNet-like resources, in general) is often criticized for its excessively fine-grained sense distinction that is not supported by syntactic, syntagmatic or semantic criteria, and is neither really needed for NLP tasks (Hanks and Pustejovsky, 2005; Navigli, 2006; Snow et al., 2007) nor reflects the way people represent word meaning (Ide and Wilks, 2007; Brown, 2008). Nonetheless, SemCor remains the state-of-the-art resource in most WSD experiments. For this study, we selected a subset of words that occur at least 20 times in SemCor 3.0 and calculated the frequencies of their senses directly from labeled SemCor contexts, with an estimated maximum frequency error of 15–20%.

4. Results and Discussion

Comparing the senses of cognates by taking into account their frequency, one can detect various cases of cognates whose meaning structures are dissimilar. For some words, no senses have a match in the other language at all (these are authentic "faux amis"). In our data, typical examples of such pairs would be *arka* (most frequent sense 'a structure with a curved top and two straight sides that you can walk through, an arch') vs. *arc* 'a curved shape', *vagon* 'railway carriage' vs. *wagon* 'any of the various kinds of wheeled vehicles drawn by an animal or a tractor', *gradus* 'degree' vs. *grade* 'a level of school'. In other cases, there is one matching sense (or more), but the most frequent senses differ drastically. Cf. *avtoritet* 'the property of someone such that people see it proper to take into account his/her opinions because of his/her knowledge and experience' vs. *authority* 'the power or right to give orders or make decisions', *akcija* 'one of the equal parts of a company that you can buy, a share' vs. *action* 'something done', *artist* 'someone who performs in plays and films' vs. *artist* 'someone who makes paintings, sculptures etc', *banda* 'a group of criminals acting together' vs. *band* 'a group of musicians'.

In many other pairs, several senses match but others do not, and learners naturally tend to use them incorrectly. The Russian noun *blok* has nine senses in ADR, whereas the English noun *block* has fourteen senses in the MacMillan dictionary. Some of the senses these words share differ significantly in their frequency, e.g. 'a solid piece of something (usually having flat rectangular sides)' (more frequent in English), 'a number or quantity of related things dealt with as a unit' (more frequent in Russian), 'a pulley' (more frequent in Russian). Some other senses of the Russian word absent in its English counterpart are 'a set of tightly packed or fastened homogenous objects' (cf. English *carton*), 'a group of organizations sharing a joint purpose' (cf. English *bloc*). The senses of the English word absent in its Russian counterpart include 'a rectangular area in a city surrounded by streets and usually containing several buildings' (cf. Russian *kvartal*), 'a three-dimensional shape with six square or rectangular sides' (cf. Russian *parallelepiped*), 'a large building with a lot of different levels' (cf. Russian *mnogoaetazhka*), 'a building that is part of a larger building or group of buildings' (cf. Russian *korpus*). Another interesting example is the pair *baza* (7 senses in ADR) vs.

base (12 senses in MacMillan, only 4 of which occur in SemCor, only 2 of the latter having correspondent senses in the Russian word).

Some of the mismatches exemplified above are not mentioned in dictionaries, and lack of knowledge thereof often leads to erroneous translations or usage. Cf. the expression *criminal authority,* clearly a calque of Russian *kriminal'nyj avtoritet*, which can be found in texts translated into English from Russian: *The Russian mass media informed that Alexander Matusov nicknamed Basmach, head of the Shelkov criminal gang, a* ***criminal authority****, has been arrested in Thailand* (Lragir.am). It occurs, however infrequently, in genuine English texts, meaning 'power to make decisions with regard to crime': *The state had no civil or* ***criminal authority*** *to force the surrender of revoked permits* (Gun control. A report by United States General Accounting Office). Another expression that reveals the Russian origin of English texts is *touristic base* meaning 'a camping site': *The* ***touristic base*** *built in Erzhei for the 'Oktai' ensemble is now open not only to the young singers but to everybody* (TuvaOnline.ru). It very rarely occurs in genuine English texts, meaning 'a base for tourism': *Such issues need to be recognized, addressed and appropriately dealt with if a precinct is to form a truly sustainable part of a city's* ***touristic base*** (Bruce Hayllar, Tony Griffin, Deborah Edwards. City Spaces—Tourist Places); *tourist base* is slightly more frequent for this meaning.

It has to be taken into account that the usage of such words (and probably their image in the mental lexicon) may differ in bilinguals (see e.g. Schreuder and Weltens, 1993; Jiang, 2004; Dong et al., 2005, Degani and Tokowicz, 2013). By distinguishing senses that are not shared in cognate pairs in standard language, we can more easily reveal cases when they are mixed up and include them as examples of non-standard usage into standard language learning manuals or dictionaries. For example, the word *blok* is widely used by Russian immigrants in the USA in the sense of 'the distance along a city street from where one road crosses it to the next road' (absent in standard Russian, where its equivalent is the word *kvartal*): *Kogda segodnja cheloveku proshche sest' v mashinu i proexat' dva* ***bloka*** *v magazin, to emu nado pomnit' o svoem zdorov'e* 'Nowadays, when it is easier for one to drive two blocks to the store, one has to consider one's health' (Chajka, a Russian magazine published in the USA); *Tak net zhe, nado bylo emu imenno takoe mesto dlja obeda vybrat', chto by perekryt' chut' li ni samyj glavnyj vyezd iz dauntauna. Dvadcat' minut v ocheredi, chto by vyexat' iz garazha i eshche sorok, chto by proexat' dva* ***bloka*** 'But no, he had to choose precisely such a place for lunch to block probably the most imporant exit from downtown. Twenty minutes waiting in the line line to leave the garage and forty more to drive two blocks' (Livejournal.com). Many other interesting cases can be found when immigrants use a word in a sense that is absent in the standard language because it has acquired a borrowing for this sense; cf. *akcija* vs. *aekshn* < *action*: *Fil'm nudnyj, bez akcii* 'The film is boring, no action' (immigrant usage) vs. *Mogli by razbavit' specaeffektami, no i zdes' ix malo, aekshn otsutstvuet* 'They could have include special effects, but there are few of them, there is no action' (from Internet movie discussion websites).

Online dictionaries and translation memories dealing with parallel corpora sometimes contain non-genuine texts, which can result in misleading their users. E.g. the recently launched resource linguee.com, a powerful translation tool combining an editorial dictionary and a search engine for parallel corpora, provides 28 examples

of parallel Russian-English texts for the Russian word *vagon*, in 5 of which it is rendered as *wagon* in English (all of them taken from Russian or Czech websites and clearly representing translations into English rather than genuine English texts), cf. *Once, when they were travelling by train, a* ***wagon*** *accidentally disconnected from the train and began to roll slowly down a slope* (from Skolkovo.ru). An inverted example: in the same dictionary we can see the English word *arc* translated into Russian as *arka* in the following sentence: *The result will be an* ***arc*** *defined by three points—V rezul'tate poluchitsja* ***arka****, postroennaja po trem tochkam* (in this geometrical context, one should use *duga* rather than *arka*). Such infelicitous translations can be found in Google Translate, too; cf. *kriminal'nyj avtoritet—criminal authority*, *sistema blokov—blocks system*, *v vagone—in the wagon* (checked on February 17, 2016).

In February 2016 we performed an online experiment with participants claiming to be (1) native speakers of English, (2) native speakers of Russian, and (3) native speakers of other languages. The respondents were to find mistakes in fifteen English sentences taken from real texts (sometimes slightly shortened). Besides ten filler sentences (taken either from explanatory dictionaries of English or from various articles), there were five sentences containing one of the words under study (*arc*, *authority*, *base*, *block*, *wagon*):

(1) *To do so it takes a wide arc to the east, via the villages of Pilton and Croscombe.*

(2) *The tribunal had no criminal authority except over soldiers.*

(3) *The touristic base of the region expanded greatly in the 1950s.*

(4) *The band regularly played at the Mad Frog bar located just a couple of blocks from the campus.*

(5) *Families were walking beside wagons pulled by teams of oxen.*

(6) *The outhouse with an arc was built much later than the main building, in 1867.*

(7) *The head of the Shelkov gang, a criminal authority, has been arrested in Thailand.*

(8) *The touristic base built for the ensemble is now open not only to the young singers but to everybody.*

(9) *Caucasian men representing NATO block nations accused Eritrea of human rights violations.*

(10) *Once, when they were travelling, a wagon accidentally disconnected from their train.*

No further information about the nature or number of the mistakes was available. As it appeared, English speakers reported significantly less mistakes (p-value 0.03) in sentences (1–5) where the words under discussion were used in their dictionary

meanings. Russian speakers reported less mistakes in sentences (6–10) where these words were used with meanings absent in English dictionaries but natural to the Russian cognates of these words. Sentences (7) and (8) were least accepted by native English speakers, many of them claimed that *criminal authority* and *touristic base* did not make sense at all. As for native Russian speakers, they had most problems in understanding sentences (2) and (3).

Table 1. Percent of mistakes in the usage of the words *arc*, *authority*, *base*, *block*, *wagon* reported by English and Russian speakers

	Sentences (1–5)	**Sentences (6–10)**
English speakers	1%	21%
Russian speakers	8%	10%

The participants of the experiment did not have to prove their proficiency in English, but for most of them it could be estimated as quite high judging by their answers where they corrected the mistakes deliberately made in the test sentences and not related to the words in question.

Hence, the results seem promising: we can obtain data that may prove useful for learners of Russian or English as well as for lexicographers and computational linguists dealing with machine translation or deep semantic analysis.

Furthermore, this technique can be applied not only to cognates, but also to pairs of words which are usually offered by dictionaries as translation equivalents of each other (see e.g. Dobrovol'skij 2007, where sets of senses of polysemous words in Russian and German are compared, for interesting observations). In order to elaborate this idea, we took a random sample of highly polysemous (more than 5 senses in ADR) and relatively frequent Russian nouns (*verx, veshch', vid, volna, vstrecha, glubina, golos*) whose most obvious English equivalents (*top, thing, view, wave, meeting, depth, voice*) are attested in SemCor, and compared their sense frequency distributions. For every word, there were senses of the Russian word that had no counterparts in the English equivalent, and vice versa. Some corresponding senses showed significant difference in frequency. Cf. *veshch'* (most common sense 'an artifact', frequency of 0.36 in Russian, cf. 0.11 for the same sense of *thing* in English) vs. *thing* (most common sense 'a special situation', frequency of 0.20, cf. 0.06 for a similar sense of *veshch'* in Russian); *vstrecha* (most common sense 'an encounter', frequency of 0.49 in Russian, cf. 0.05 for the same sense of *meeting* in English) vs. *meeting* (most common sense 'a formally arranged gathering', frequency of 0.79, cf. 0.30 for the same sense of *vstrecha* in Russian). Interestingly enough, the English word *meeting* in its most frequent sense was borrowed into Russian as *miting* in a narrower sense ('political gathering'), whereas the Russian word *vstrecha* recently developed a new sense 'an appointment'. One of the senses of *vstrecha* is 'celebration', which is apparently absent in English but can be found on Russian websites, cf. *On the chair the evenings and the celebrations devoted to the Victory Day, a meeting of New Year, to a farewell to winter (Shrovetide) etc. are spent* (Moscow State Pedagogical University, Chair of Russian as a foreign language, English website).

5. Conclusion

The idea of this study was to perform a pilot experiment determining sense frequencies for cognate Russian and English words in corpora and, taking this information into account, compare their meaning structures. The main issue is the lack of large semantically annotated corpora and dictionaries with a sufficient number of examples for each word. This limits the possibilities of automatic techniques for calculating meaning frequency. Our future plans include the following:

- applying the method of estimating word sense frequencies used for Russian on the base of the examples provided in the Active Dictionary of Russian to English, by using the data from the MacMillan dictionary and the Dante database (www.webdante.com);
- studying parallel Russian-English corpora (primarily subcorpora of the Russian National Corpus) to investigate possible differences in meaning structures of cognates used by native and non-native speakers, in Russian-to-English and English-to-Russian translations;
- creating and updating a database of Russian-English cognates comparing their senses according to their frequency;
- expanding expanding the method to other language pairs; inter alia, compare closely related languages (such as Russian and Polish) to provide material for comparative semantic and lexicological studies.

References

1. *Apresjan Ju. D.* (1974/1995). Lexical semantics. The synonymical means of language [Leksičeskaja semantika. Sinonimičeskie sredstva jazyka]. Nauka, Moscow.
2. *Apresjan Ju. D.* (ed.). (2014). Active Dictionary of Russian. A-G [Aktivnyj slovar' russkogo jazyka. A-G]. Jazyki slavjanskih kul'tur, Moscow.
3. *Agirre E. and Edmonds P.* (Eds.). (2007). Word sense disambiguation: Algorithms and applications (Vol. 33). Springer Science & Business Media.
4. *Brown S. W.* (2008). Choosing sense distinctions for WSD: Psycholinguistic evidence. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics, pages pp. 249–252.
5. *Darbelnet J.* (1970). Dictionnaires bilingues et lexicologie différentielle. Languages, 19:92–102.
6. *Degani T. and Tokowicz N.* (2013). Cross-language influences: Translation status affects intraword sense relatedness. Memory & cognition, 41(7):1046–1064.
7. *Dobrovol'skij D. O.* (2007). Polysemy structure in cross-linguistic perspectives (verbs of motion in Russian and German). In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2007", pages 162–166.
8. *Dong Y., Gui S. and MacWhinney B.* (2005). Shared and separate meanings in the bilingual mental lexicon. Bilingualism: Language and Cognition, 8(03):221–238.
9. *Fellbaum C.* (ed.) (1998). WordNet, An Electronic Lexical Database. The MIT Press, Cambridge, MA.

10. *Francis W. N. and Kučera H.* (1979). A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). Brown University. Providence, Rhode Island.
11. *Gries S. T., Hampe B. and Schönefeld D.* (2010). Converging evidence II: More on the association of verbs and constructions. In: Empirical and experimental methods in cognitive/functional research, CSLI Publications, pages 59–72.
12. *Hanks P.* (2008). Mapping meaning onto use: a Pattern Dictionary of English Verbs. In Proceedings of the AACL, Utah.
13. *Hanks P.* (2013). Lexical analysis: Norms and exploitations. Boston: MIT Press.
14. *Hanks P. and Pustejovsky J.* (2005). A Pattern Dictionary for Natural Language Processing. Revue Française de linguistique appliquée, 10(2):63–82.
15. *Ide N. and Véronis J.* (1998). Introduction to the special issue on word sense disambiguation: the state of the art. Computational linguistics, 24(1):2–40.
16. *Ide N. and Wilks Y.* (2007). Making sense about sense. In Word sense disambiguation. Springer Netherlands: 47–73.
17. *Iomdin B.* (2014). Polysemous words in and out of the context. [Mnogoznachnyje slova v kontekste i vne konteksta]. Voprosy jazykoznanija [Issues in Linguistics]. Vol. 4. Moscow.
18. *Iomdin B., Lopukhina A., Nosyrev G.* (2014). Towards a word sense frequency dictionary. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014", pages 205–219.
19. *Jiang N.* (2004). Semantic transfer and development in adult L2 vocabulary acquisition. In Vocabulary in a second language: Selection, acquisition, and testing, pages 101–126.
20. *Kilgarriff A., Rychly P., Smrz P. and Tugwell D.* (2004). The Sketch Engine. In Information Technology Research Institute Technical Report Series, pages 105–116.
21. *Koessler M. and Derocquigny J.* (1961). Les faux amis ou les pièges du vocabulaire anglais. Paris: Vuibert.
22. *Kusal K.* (2006). Russian-Polish interlinguistic homonymy and paronymy. A doctoral dissertation.
23. *Kwong O. Y.* (2012). New perspectives on computational and cognitive strategies for word sense disambiguation. New York: Springer Science & Business Media.
24. *Lau J. H., Cook P., McCarthy D., Gella S. and Baldwin T.* (2014). Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In Proceedings of ACL, pages 259–270.
25. *Lin C. C. and Ahrens K.* (2005) How many meanings does a word have? Meaning estimation in Chinese and English. In Language acquisition, change and emergence: Essays in evolutionary linguistics. Hong Kong, pages 437–464.
26. *Lopukhin K., Lopukhina A.* (2016). Word sense disambiguation for Russian verbs using semantic vectors and dictionary entries. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016".
27. *Lopukhina A., Lopukhin K. and Nosyrev G.* Automated word sense frequency estimation for Russian nouns. In Quantitative Approaches to the Russian Language. In print. (Available online: sensefreq.ruslang.ru/download/Automated_Word_Sense_Frequency_Estimation_for_Russian_Nouns_Lopukhina_et_al.pdf)

28. *Lopukhina A., Lopukhin K., Iomdin B. and Nosyrev G.* (2016). The taming of the polysemy: automated word sense frequency estimation for lexicographic purposes. In Proceedings of EURALEX-2016. In print.
29. *Loukachevitch N. and Chetviorkin I.* (2015). Determining the Most Frequent Senses Using Russian Linguistic Ontology RuThes. In Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA.
30. *McCarthy D., Koeling R., Weeds J. and Carroll J.* (2007). Unsupervised acquisition of predominant word senses. Computational Linguistics, 33(4):553–590.
31. *Mikolov T, Chen K., Corrado G. and Dean J.* (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
32. *Miller G. A., Leacock C., Tengi R. and Bunker R. T.* (1993). A semantic concordance. In Proceedings of the workshop on Human Language Technology, Association for Computational Linguistics, pages 303–308.
33. *Mohammad S. and Hirst G.* (2006). Determining Word Sense Dominance Using a Thesaurus. In EACL.
34. *Naumov V. G.* (2014). The Ruthenian-Russian interlingual formal-semantic similarity in lexicographical representation: principles of a learner's dictionary. In: Rusin, 4(38).
35. *Navigli R.* (2006). Meaningful clustering of senses helps boost word sense disambiguation performance. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pages 105–112.
36. *Navigli R.* (2009). Word sense disambiguation: A survey. ACM Computing Surveys (CSUR) 41.2: 10.
37. *Pustejovsky J.* (1996). Lexical semantics: The problem of polysemy. Oxford.
38. *Schreuder R. and Weltens B.* (Eds.). (1993). The bilingual lexicon (Vol. 6). John Benjamins Publishing.
39. *Sharoff, S.* (2006). Creating general-purpose corpora using automated search engine queries. In Baroni and Bernardini, 2006, pp. 63–98.
40. *Snow R., Prakash S., Jurafsky D. and Ng A. Y.* (2007). Learning to merge word senses. In Proceedings of EMNLP. Prague, Czech Republic.
41. *Szpila G.* (2006). False friends in dictionaries. Bilingual false cognates lexicography in Poland. International Journal of Lexicography, 19(1):73–97.
42. *Vrbinc M. and Vrbinc A.* (2014). Friends or Foes? Phraseological False Friends in English and Slovene. AAA: Arbeiten aus Anglistik und Amerikanistik, pages 71–87.
43. *Walter H.* (2002). Les "faux amis" anglais et l'autre côté du miroir. La linguistique, 37(2):101–112.
44. *Zalizniak Anna* (2006), Polysemy in language and its representation [Mnogoznačnost' v jazyke i sposoby ee predstavlenija]. Jazyki slavjanskih kul'tur. Moscow.