The Taming of the Polysemy: Automated Word Sense Frequency Estimation for Lexicographic Purposes

Anastasiya Lopukhina[a], Konstantin Lopukhin[c], Boris Iomdin[a,b], Grigory Nosyrev[d]

[a]Vinogradov Russian Language Institute, Russian Academy of Sciences,
[b]National Research University Higher School of Economics,
[c]Scrapinghub,[d]Yandex
e-mail: nastya.lopukhina@gmail.com, kostia.lopuhin@gmail.com,
iomdin@ruslang.ru, grigorij-nosyrev@yandex.ru

Abstract

Although word sense frequency information is important for theoretical study of polysemy and practical purposes of lexicography, the problem of sense frequency distribution is a neglected area in linguistics. It is probably because sense frequency is not easy to estimate. In this paper we deal with the problem of automated word sense frequency estimation for Russian nouns. We developed and tested an automated system based on semantic context vectors, supplied with contexts and collocations from the Active Dictionary of Russian – a full-fledged production dictionary that reflects contemporary Russian. The study was performed on RuTenTen11 web-corpus. This allows us to reach a frequency estimation error of 11% without any additional labelled data. We compared sense frequencies obtained automatically with sense ordering in different dictionaries for several words. The method presented in this paper can be applied to any language with a sufficiently large corpus and a good dictionary that provides examples for each sense. The results may enrich language learning resources and help lexicographers order senses within a word according to frequency if needed.

Keywords: semantics; lexicography; word sense frequency; web corpora; polysemy; frequency; semantic vectors; word sense disambiguation; WSD

1 Introduction

Words have many linguistic properties. Normally, not all of these properties are described in the dictionary. The number and the set of the properties depend on the type of dictionary and the goals of its author. Many dictionaries are intended to facilitate text and speech understanding, such dictionaries are sometimes called passive and usually do not go beyond basic grammar information (e.g. word forms), meaning definitions, and examples of how the word is typically used. In contrast, the goal of the so-called production, or active, dictionaries is to facilitate the generation of text and speech. This can be achieved by including all the information speaker may need to use the language correctly: word forms and their meanings, stylistic, syntactic and pragmatic properties, collocational restrictions, synonyms and antonyms, idioms which contain this word, etc (Apresjan 2008; Tarp 2008). Accordingly, passive dictionaries normally include more words, even very rare ones, but do not provide much information about them, while active dictionaries describe in greater detail fewer words that are likely to be used more frequently.

Word frequency can also be considered a property that can be included into the dictionary entry. Kilgarriff (1997) believed in the importance of knowing which words (and word senses) are the most frequent in the language for the purposes of language learning. The information about word frequency was explicitly introduced in the Longman Dictionary of Contemporary English (3rd edition, 1995). The authors selected 3000 most frequent written and spoken words and marked them using special symbols for the first, second and third thousands separately (LDOCE 1995). However important, this information is less illuminating for polysemous words. As shown (Kilgarriff 2004),

within a polysemous word senses are not distributed evenly, and some senses generally occur more frequently than the others. Thus it would be much more useful to present separate information about the frequency of each sense of the word, rather than the total frequency of the word. Apparently no dictionary or learner's resource provide this type of information.

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 & Pustejovsky 2005; Hanks 2008). This project is based on two other projects: (a) Disambiguation of Verbs by Collocation and (b) Corpus Pattern Analysis, both focused on statistical analysis of corpus data and aimed to discover typical usage patterns. The authors emphasize that meanings in the Pattern Dictionary are associated with prototypical sentence contexts (patterns or collocations) and not with word senses from dictionaries. Cf. also (Gries et al. 2010), where frequency distributions of English verbal constructions are discussed. Although Patrick Hanks and colleagues' resource contains the information about the relative frequency of verb patterns, it cannot be easily integrated into explanatory dictionaries; besides, it is focused only on verbs.

The lack of word sense frequency resources is a problem in language learning and teaching. Studies in dictionary user behaviour show that learners often satisfy themselves with the first sense listed in the dictionary, even if it does not fit into the contexts, which leads to incorrect interpretations of the texts they read. See e.g. the results of the experiments described in (Nesi and Haill 2015): "In almost every case it seems that this kind of error arose because subjects unthinkingly selected the first meaning provided for the headword, rather than a more appropriate definition listed later in the entry". Word sense frequency information is also very important in the task of making lists of words to be learned. (Beck at al. 2013: 21) state that any word that has different meanings appears only once in such lists: "Whether bank means financial institution, edge of a river, or angle of an airplane is not taken into account. B-a-n-k appears one time on the list, and its associated frequency represents all the different meanings. In other words, there is no way to get the frequency of the word bank meaning a financial institution". This happens not only with homonyms like *bank*, but also with polysemous words whose senses stand quite far from each other. For example, according to the Active Dictionary of Russian (Apresjan et al. 2014), the Russian word *batareya* can be described as having 4 distinctly different senses: 'several large guns used together', 'a hot water radiator', 'an electric battery', 'a collection of many objects of the same type', . Native speakers would probably agree that the first sense is quite special and rare as compared to the rest. So the information about word sense frequency could help students learn the most relevant sense(s) of the word first.

In this paper we present a method for determining noun sense frequency distributions automatically from raw corpora, the evaluation of this method, and a discussion on its applications to dictionaries. The technique we propose is based on semantic context vectors and uses contexts and collocations from the Active Dictionary of Russian.

2 Word Sense Frequency Estimation

Word sense frequency is not easy to estimate. For the proper estimation we need a source of word senses (a dictionary), a source of word contexts (a corpus), and a sense disambiguation technique. The choice of the dictionary is crucial because it determines the word senses, as a word may have different numbers of senses in different dictionaries. Thus we need a reliable resource with strong theoretical basis that reflects the contemporary language. The Active Dictionary of Russian (ADR), an ongoing project led by Juri Apresjan and the group of researchers from the Russian Language Institute, meets these requirements (Apresjan et al. 2014). ADR is the first attempt at creating a full-fledged production dictionary of the Russian language. Though limited in size, it is now the most up-to-date and the most developed explanatory dictionary of Russian. 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, semantic derivatives, etc.). The lexical entries for all lexemes contain a variety of usage examples based on large corpora (mainly the Russian National Corpus, RNC, https://ruscorpora.ru), which turn out to be crucial for studying sense frequencies.

In this study we use the word senses as defined in this dictionary. Our current research is focused on nouns, because they normally have more distinctly different senses compared to other parts of speech — such as prepositions (*in)*, verbs (*be)* or adjectives (*generous*) — as most of them refer to objects existing in the real world, see similar studies on nouns in (Kilgarriff 2004) and (Iomdin et al. 2014).

The choice of corpus may influence sense frequency, because word sense distributions vary from corpus to corpus. For the purposes of the current study we use the contexts from the RuTenTen11 web-based corpus, the largest Russian corpus consisting of 18 billion tokens integrated into the Sketch Engine system (Kilgarriff et al. 2004). We sample 1000 random contexts for each word, and estimate sense frequency by performing sense disambiguation on these contexts and counting the relative frequency of senses. This sample size yields a statistical error below 3.1%. Of course, other corpora could be used for the same task, first of all RNC, a resource made by a consortium of linguists and developers, and considered to be the best academic corpus for Russian. Kutuzov & Kuzmenko (2015) found that RNC and web-based corpora agree with each other in most cases. Web corpora, however, have more recent data and provide relevant and comparable linguistic evidence for lexicographic purposes (Ferraresi et al. 2010).

3 Method

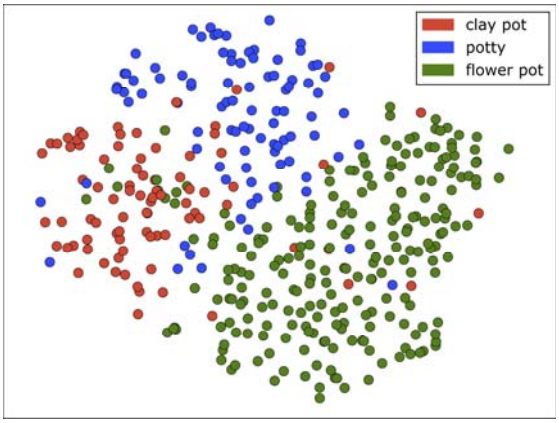
The most precise and reliable way of estimating sense frequency is labelling enough contexts randomly sampled from corpora, but this method is very time-consuming. Another method evaluated (Iomdin et al. 2014) reduces the number of contexts to a label, and instead requires the labelling of collocations. Still, we believe that fully automated methods are preferable if they allow reaching good precision: they reduce the amount of human labour and can be easily applied to different corpora or words. Supervised methods for word sense disambiguation (WSD) were extensively studied, especially during SemEval evaluation series, and reach the accuracy of 85-90% if given hundreds of labelled examples (Navigli 2009), but obtaining enough labelled examples for a large number of words is very processor-intensive. The most promising and robust are fully unsupervised, or sense induction (WSI) methods that solve the knowledge acquisition bottleneck by discovering senses from unlabelled corpora. Such sense discovery can be performed either by building vector representations of contexts and applying conventional clustering methods (Schütze 1998), or by learning multiple vector embeddings for each word (Huang 2012; Neelakantan 2014; Bartunov et al. 2015). Clusters produced by unsupervised methods can be mapped to senses from the dictionary using definition or a small number of examples.

When designing our method, the main constraint we face is the limited availability of sense-annotated data (often just several examples per sense). In other words, the method must be very robust with respect to both quantity and quality of labelled examples. To achieve this, we use large amounts of unlabelled data available from corpora to train a word2vec model that provides us with dense real-valued vectors for each word, called word embeddings, or semantic vectors. We chose semantic vectors as a basic building block because they already capture the most important semantic properties of words in a very compact and easy-to-use way: close vectors correspond to semantically close words. When building representation of context from the words that form this context, we give more weight to words that occur more frequently with the target word than without it, using PMI (Pointwise mutual information) weighting, removing words with negative PMI and assigning small fixed weight to words with unknown PMI. For example, for word *gorshok* (‘clay pot / flowerpot / potty’) words with high weight include *rasteniye* (‘a plant’) and *priuchat’* (‘to accustom’), and words with low weight are e.g. *yesli* (‘if’) and *uchyonyj* (‘scientist’). This

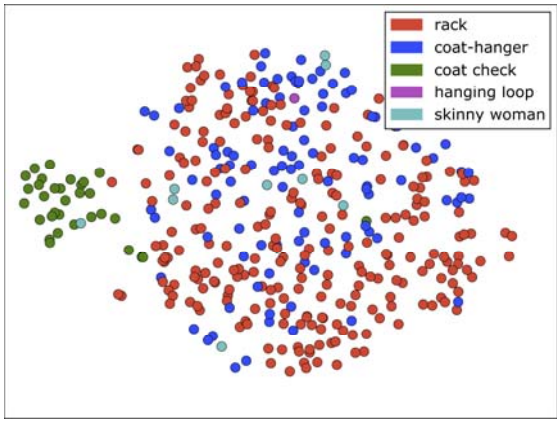
weighting not only removes random or too common words, but also produces context vectors that better capture the sense distinctions characteristic of the disambiguated word. Once we have represented a context as a semantic vector (where close or similar vectors correspond to similar meanings, and hopefully the same sense), we build sense vectors in the following way: we take context vectors for all dictionary examples and collocations for one sense, and average them. During disambiguation, we assign each unknown context to the sense with the closest sense vector.
In order to obtain sense frequencies, we sample a large number of contexts (typically, 1000) from the corpus, perform word sense disambiguation for them, and then estimate sense frequencies from this sample.

4 Evaluation

We evaluated the method on hand-labelled contexts for 20 polysemous words from the Active Dictionary of Russian (sampled from RuTenTen11, at least 100 contexts for each word) and found out that it achieves an average disambiguation accuracy 77% and the maximum frequency error 11%. For many words, the accuracy is close to 90%, but there are some words that are especially troublesome for the method. One such word is *veshalka* ('coat-hanger' / 'rack' / other senses), where 'coat-hanger' and 'rack' are hard to distinguish because the context words are very similar for both senses, although they denote clearly distinct objects.



gorshok



veshalka

On these diagrams, each point is a single human-annotated context coloured according to its sense. Position of the points reflects the way our method represents these contexts: they are projected from semantic space with t-SNE method, where closer points correspond to more similar contexts. We can see that the senses of *gorshok* are much better separated than those of *veshalka*. This happens because the senses of word *gorshok* have very different context words (connected with cooking, children and house plants respectively), but two most popular senses of *veshalka*, 'clothes rack' and 'coat-hanger', have very similar contexts. In the sentence "*povesit' chto-libo na veshalku*" ('to hang something on the clothes rack/coat-hanger'), the most probable sense of the word *veshalka* depends on the type of the object that is put on (*a dress* or *a jacket* for 'coat hanger' and *a hat* or *an anorak* for 'clothes rack'). It is hard to resolve this ambiguity without human knowledge about objects or a lot of labelled contexts.
Another example is *block* with 9 senses in ADR, where many senses are fine-grained and abstract. Human annotators struggled the most with this word, with inter-annotator agreement just above 50%, compared to 88% of average agreement for other words.

5 Discussion: Examples

Accurate word sense frequency data can be used to reconsider sense ordering and depth of description in the dictionaries. We studied the sense ordering in different published dictionaries of Russian; the following examples are based on the sample results already obtained using our method.

The Russian word *batareya*, borrowed from French or German *Batterie* (written as *battereya* in the 18th century), was first used by Peter the Great in 1697 (Vasmer 1986), but is apparently first attested only by (Dahl 1863-1866) as in 'artillery battery'. This sense is given priority in most Russian dictionaries, whereas according to our data, its frequency is 18%. The most frequent sense of the word is now 'a current source providing electric current for a device' (63%). However, the latter sense had not been attested at all by any major Russian monolingual dictionary before ADR (and even there, it is only the third sense of the word). Some dictionaries include a similar, but not identical and a much broader sense 'a combination of several appliances, devices or instruments of the same type within a united system or unit, used for a joint action', with *batareya akkumulyatorov* 'battery of accumulators' given as one of the examples. Note that the electrical battery (which was coined in English by Benjamin Franklin in 1749) can be found in Russian texts in the middle of the 19th century (*leydenskaya batareya* 'Leiden battery'). Oxford, American Heritage, Collins, MacMillan, Merriam-Webster, and Cambridge dictionaries all list this sense as the first one. A historical dictionary of French loanwords borrowed into Russian, published in 2010, gives as many as 14 senses (and many subsenses) for *batareya*, and the sense under discussion is given as one of several subsenses of the 11th sense: 'several joined galvanic cells or accumulators'.

According to our data, the most frequent sense of the word *greben'* is the last one in the list of its six senses presented in ADR: 'the upper edge of a relatively large, tall and normally elongated tall object, or several such objects located near to each other', cf. *greben' volny* 'wave crest', *greben' gory* 'mountain ridge', *greben' kryshi* 'roof crown'. It accounts for 61% of all occurrences of *greben'*, whereas the frequency of the first sense in the list explained as 'a shaft with teeth used for combing hair as well as for fixing and decoration of the hairdo' is apparently only 10%. The now more frequent sense probably results from a metaphorical shift and was attested already in (Dictionary of the Russian Academy 1789-1794), but is still listed as the last one by all major Russian dictionaries. Note that for the first sense, 'comb', another word (*raschyoska*) is now used much more frequently in Russian, as follows from our corpora-based research.

The Russian word *gudok* is explained in (Dictionary of the Russian Academy 1789-1794) as 'an ancient popular musical instrument similar to the violin'; only this sense is attested in (Dahl 1863-1866), too. (Ushakov 1934-1940) gives two senses for the word: 1. 'a large mechanical whistle used for signaling', 2. 'a long drawling sound of a siren or a whistle'. Since then, all dictionaries have been listing the same set of senses for *gudok*, and in the same order. First corpora examples for the second sense (*parokhodnyj gudok* 'steamship whistle', *gudok parovoza* 'train whistle', *gudki zavodov* 'factory whistles') appear in the last quarter of the 19th century. In our data, this sense now accounts for 69% of the examples. According to the RuTenTen11 data provided by the Sketch Engine, the most frequent adjective phrases used with *gudok* are *korotkiye gudki* 'busy tone', literally 'short beeps', and *dlinnye gudki* 'ringing tone, ringback tone', literally 'long beeps'. The correct English translation equivalents for these two phrases cannot be found in any dictionary, including the largest crowdsource dictionary Multitran.ru, nor provided by machine translation (by Google, Yandex, or Bing).

The Russian word *garderob* has a set of senses similar to that of its English cognate *wardrobe.* The sense 'the collection of clothes that someone has' (cf. *garderob delovoj zhenschiny* 'clothes of a businesswoman', *eyo letnij garderob* 'her summer wardrobe') accounts for 76% of all occurrences of *garderob*, and the frequency of the first sense – 'a large piece of furniture where you can hang your clothes' is 12%. This metonymical shift was attested in (Dahl 1863-1866), and the former sense has been listed as the last one by the Russian dictionaries ever since. As for *wardrobe*, English dictionaries differ in ordering the set of its senses: the furniture item sense is the first one in

Oxford, American Heritage, Collins, Macmillan, the second one in Cambridge and Dictionary.com, and the third one in Merriam-Webster.
The information about the most frequent sense of a polysemous word may be important for language learners and thus should be reflected in dictionaries. We intend to apply it to ARD.

6 Conclusions and Future Work

We showed that sense frequency information is important for theoretical and practical purposes, and may enrich language learning resources and help lexicographers order senses within a word according to frequency if needed. We introduced a method for obtaining such information from a large corpus and a good dictionary, and analysed sense frequency results for Russian nouns obtained from RuTenTen11 corpus with ADR.
The frequency distribution presented in this paper was obtained from one web corpus, but different corpora might have different sense distributions, so it is important to compare our results to other corpora of modern Russian – RNC, General Internet-Corpus of Russian (Piperski et al. 2013), web corpus Ruwac (Sharoff 2006) and analyse the difference.
Currently, the method uses dictionary senses verbatim, trying to classify each context, and assumes that there are no senses missing from the dictionary. It should however be possible to modify the method in such a way that it could suggest senses not described in the dictionary, which can help lexicographers in their work.
Sense frequency distribution for a large list of nouns provides exciting opportunities for theoretical studies. It is interesting to study frequency distribution within a word and find patterns that may depend on the type of polysemy. Moreover we have the data to test Kilgarriff's assumption about the dominance of the commonest sense of the word for Russian and compare it to the results for English (Kilgarriff 2004). It is also interesting to study the evolution of the lexical system by counting relative frequencies using the contexts from the historical subcorpora of different periods available in RNC. An example of an NLP task that could be tackled with this information is the problem of disambiguation in the absence of context (outlined in Iomdin 2014). In (Iomdin et al. 2016, in print) we proposed a method of using sense frequency information for comparing the meaning structures of cognates and other similar words in different languages, which might be useful for language learners.
The method we present in this paper can be applied to any language with a sufficiently large corpus and a good dictionary that provides examples for each sense.

7 References

Active Dictionary of Russian. A-G. Edited by Ju. D. Apresjan. (2014). Moscow.
Apresjan, Ju. D. (2008) Theoretical Foundations of Production Dictionaries. RAS General Meeting. Herald of the Russian Academy of Sciences, Volume 78, Issue 3., pp. 203-208.
Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D. (2015). Breaking sticks and ambiguities with adaptive skip-gram. arXiv preprint arXiv:1502.07257, 2015.
Beck, I., McKeown, M., Kucan, L. (2013). Bringing words to life: Robust vocabulary instruction. Guilford Press, 2013.
Dahl, V. (1866). Explanatory Dictionary of the Living Great Russian Language. I-IV. Moscow, St. Petersburg. 1863-1866.
Dictionary of the Russian Academy. I-VI. St. Petersburg, 1789–1794.
Ferraresi, A., Bernardini, S., Picci, G., Baroni, M. (2010). Web corpora for bilingual lexicography: a pilot study of English/French collocation extraction and translation In Using Corpora in Contrastive and Translation Studies. Newcastle: Cambridge Scholars Publishing, 2010.

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.

Hanks P. (2008). Mapping meaning onto use: a Pattern Dictionary of English Verbs. In Proceedings of the AACL, Utah. 2008.

Hanks P. and Pustejovsky J. (2005). A Pattern Dictionary for Natural Language Processing. In Revue Française de linguistique appliquée, 10(2):63-82.

Huang, E., Socher, R., Manning, Ch., Ng, A. (2012). Improving Word Representations via Global Context and Multiple Word Prototypes. 2012.

Iomdin, B. (2014) Polysemous words within and without context. Issues in Linguistics. Vol. 4. Moscow.

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", Bekasovo, pp. 199–212.

Iomdin, B., Lopukhin, K., Lopukhina, A., Nosyrev, G. (2016). Word sense frequency of similar polysemous words in different languages. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016", in print.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. International Journal of Lexicography 10 (2), pp. 135-155.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? In Text, Speech, Dialogue. Lecture Notes in Artificial Intelligence Vol. 3206. Sojka, Kopecek and Pala, Eds. Springer Verlag: 103-112.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). ITRI-04-08 The Sketch Engine. Information Technology, 105, 116.

Kutuzov, A., Kuzmenko, E. (2015). Comparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: The Case for Russian. A. Gelbukh (Ed.): CICLing 2015, Part I, Springer LNCS 9041, 2015, pp. 47–58.

LDOCE 1995 - Longman Dictionary of Contemporary English, 3rd Edition. Edited by Della Summers. Harlow.

Navigli, R. (2009). Word sense disambiguation: A survey. ACM Comput. Surv., 41(2).

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In Conference on Empirical Methods in Natural Language Processing (EMNLP).

Nesi, H., Haill, R. (2015). A study of dictionary use by international students at a British University. Author post-print (accepted) deposited in CURVE February 2015.

Piperski, A., Belikov, V., Kopylov, N., Selegey, V., Sharoff, S. (2013). Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In The 8th Web as Corpus Workshop (Lancaster, July 2013).

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni and Bernardini, pp. 63–98.

Schütze, H. (1998). Automatic word sense discrimination. Computational Linguistics 24, 1, pp. 97–124.

Tarp, S. (2008). Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography. Vol. 134. Walter de Gruyter.

Ushakov, D. N. (ed.). (1934–1940). Explanatory Dictionary of Russian. OGIZ, Moscow.

Vasmer, M. (1986). Etymological dictionary of Russian. I-IV. Moscow.

Acknowledgements

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). Contexts extraction from the RuTenTen corpus and word2vec model training was done by Grigory Nosyrev. The authors thank the anonymous reviewers for the many helpful comments and suggestions they made and Leonid Iomdin for his careful reading of the draft.