Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2018"

Moscow, May 30—June 2, 2018

LEXICAL VARIATION: WORD KNOWLEDGE AND POLYSEMY IN RUSSIAN EVERYDAY LIFE LEXICON[1]

Levin I. (levinivanse@gmail.com),
Andriyanets V. (blindedbysunshine@gmail.com)
National Research University "Higher School of Economics"

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

Ambartsumian A. (anna.ambr@yandex.ru)
Russian State University for the Humanities

Many words that according to the dictionaries have just one meaning are in fact understood in different ways by different speakers. In this article we deal with Russian nouns denoting everyday life objects which are subject to much variation by age, gender, and region and are poorly described by the existing dictionaries. We report the results of a multilevel survey, propose some possible metrics of word knowledge and show to what extent the words we studied are known among a certain population. We also claim that different speakers possess different sets of meanings for each word, propose ways to discover the distribution patterns for these sets and introduce the notion of disperse polysemy. We believe that our findings may be useful in lexicography (providing detailed information on current word usage in different social groups), lexical semantics (researching meaning shifts and patterns of its distribution among speakers), and language testing (more precise detection of the vocabulary sizes both in native speakers and in language learners).

Key words: semantics, polysemy, lexicography, lexical variation, word knowledge

[1] The research of Boris Iomdin was supported by RSF (project No. 16-18-02054: Semantic, statistic and psycholinguistic analysis of lexical polysemy as a component of Russian linguistic worldview).

0. Introduction

Linguists and lexicographers often deal with polysemy. In natural language processing, in particular, a lot of research is aimed at word sense disambiguation, normally context-dependent [Ide and Veronis 1998, Navigli 2012, Chandra and Dwivedi 2014, Iomdin 2014, Iacobacci et al. 2016]. However, many words that according to the dictionaries have just one meaning are in fact understood in different ways by different speakers. We are currently researching this issue as part of our work on the Thesaurus of Russian Everyday Life Lexicon [Iomdin 2011] and within the project "Semantic, statistic and psycholinguistic analysis of lexical polysemy as a component of Russian linguistic worldview" funded by RSF.

Here we deal with nouns denoting everyday life objects which are subject to much variation by age, gender, and region [Iomdin 2014] and are very poorly described by the existing dictionaries, so we have to obtain the necessary data from sociolinguistic surveys. Corpora are less useful here, because in the texts that they incorporate artifacts are rarely described in detail sufficient to distinguish between similar objects and to provide accurate and distinctive definitions. Since we started working on the thesaurus we have conducted many surveys to this aim. In some of them, we asked the respondents whether they knew certain Russian words, and the analysis of the results clearly shows that the answer to this question cannot be binary.

Various experiments dedicated to detecting the vocabulary size of native speakers were conducted based on the idea that one can obtain a specific number of words known by the respondent because each word can be assumed either known or unknown. In one of the earliest experiments of this type, [Hartman 1946] related the idea of *knowing* the word with the ability to give a definition to it. One word was selected from every fortieth page of Merriam Webster's New International Dictionary, and so a list of 50 words was created. On average, the students were able to define 26.9 out of the 50 words, a proportion that gave the impression that they had a vocabulary of 215,000 English words. [Goulden 1990] used the same dictionary to choose the words for their experiment, but the final list was reduced by excluding proper nouns, derived nouns and compounds. He presented lists of 50 words each to 20 university graduates who had to indicate whether they knew the word without proving it. The result was 17,200 known English words, lower than in previous experiments. In [Milton and Treffers-Daller 2013] some parts of the two previous studies were united: they took the reduced list of words from [Goulden 1990] and asked first-year university students to provide either a definition or a synonym for each word they knew. In this experiment, the resulting figure was still lower: 9,800 word families (morphologically related groups of lemmas) known to an average respondent. These three experiments show that the more complicated the selection of lemmas for the final list and the structure of the experiment are, the less is the number of known words that is received as a result.

Along with the selection of words, the fact that a word can have multiple senses or meanings must be taken into account. [Rodd, Gaskell, Marslen-Wilson 2002] studied the response time in a lexical decision task and its correlation to the numbers of senses or meanings. The results of three experiments represented "an important challenge to accepted views of how semantic ambiguity affects recognition of isolated

words. Ambiguity between multiple meanings produces a disadvantage, while multiple senses produce faster responses". [Brysbaert and Stevens 2016] in their work dedicated to the same issue (how many words a native speaker knows) note the importance of qualitative estimation of the result: "our assessment says nearly nothing about how well the participants know the various words".

The present work is based on the hypothesis that word knowledge may have a more complicated structure which includes various levels. In order to test this hypothesis, we conducted a multilevel survey, for which we selected seven Russian nouns denoting everyday objects. In **Section 1** we deal with the history and semantic development of these nouns. In **Section 2** we describe the design of the experiment. In **Section 3** we propose possible approaches to defining word knowledge and provide corresponding data from our experiment. In **Section 4** we discuss the question whether we deal with polysemy in the cases we studied.

1. Material

Having analyzed Russian text corpora (mainly Russian National Corpus and RuTenTen11) as well as the Google Books collection and various online resources, we selected several less frequent Russian words that apparently are understood differently by different speakers.

Сланцы [*slancy*] 'flip-flops, jandals'. The etymology of this word can be traced to a proper name. The factors contributing to its meaning development were probably (1) the plural form of the word, characteristic for all kinds of shoes, (2) its similarity to an older word *šljopancy* 'sliders, jandals', and (3) the novel nature of this kind of shoes and lack of a conventional name for it (another frequent colloquial term for them are *v'etnamki*, lit. 'Vietnamese', absent in dictionaries but appearing in published texts since the 1970s). The meaning shift is an example of metonymy (a label on the object → the object itself).

Барсетка [*barsetka*] 'man bag, man purse, murse'. This word is in all probability borrowed from Italian, where *borsetta* (and *borsetto*) is a diminutive form of *borsa* 'bag'. The meaning, however, differs from the Italian word *borsetta*, which means 'women purse', whereas the meaning of the Russian word is similar to that of *borsello* 'a small bag for men with the function similar to that of the female handbag, often with a strap that allows one to hang it on the shoulder'. Apparently, it was not the name of this very object, but rather the label applied to various kinds of leather handbags that was used as the basis for the Russian word.

Креманка [*kremanka*] 'ice-cream bowl, dessert bowl'. This word is generally considered to be a derivative of *krem* 'cream, hard sauce', which has a common meaning component ('dessert'). However, the word *kremanka* is now normally used for a bowl for ice cream and other desserts but not for hard sauce served separately. Moreover, the suffix *-ank(a)* is common for animated nouns rather than inanimate nouns derived from names of objects or substances. We believe that *kremanka*

is a derivation not from *krem*, but from *kreman*, a now obsolete word borrowed from the French *crémant* 'sparkling wine'. Examples of *kremanka* used in this sense can be found in texts published in late 19th century and early 20th century. The term was used for champagne coupes, which at some point started to be used as ice-cream bowls. Here we deal with a metaphorical shift: an object got its name from another object with a different purpose, but of the same form; its phonetic resemblance to the word *krem* associated with desserts contributed, too.

Тренч [*trenč*] 'trench, trench coat'. This word apparently was borrowed from English twice. English dictionaries list two senses for *trench coat*: (1) usually double-breasted raincoat with deep pockets, wide belt, and often straps on the shoulders, and (2) a waterproof overcoat with a removable lining designed for wear in trenches". The word *trench coat* was borrowed as a whole, occurring in Russian texts in the 1930s (first *трэнчкот*, then *тренчкот*). Then, according to RNC and Google Books, it was rarely used until a rebirth at the beginning of the 2000s, normally as just *тренч*. It is now associated with youth fashion rather than military style.

Манто [*manto*] 'fur opera cloak'. This word was borrowed from French *manteau* at the beginning of the 19th century (at first it was masculine, then neuter). It used to mean a coat in general, particularly a light one. Then its meaning narrowed down to women fur opera cloaks.

Душегрейка [*dušegrejka*] 'a warm women jacket', literally 'soul warmer'. This word used to describe a traditional women outer garment, generally sleeveless and warm. Another word of similar structure, *telogrejka*, literally 'body warmer', according to several dictionaries, was used as a synonym to *dušegrejka*. Later, however, the meanings of both words started to differ, and now *dušegrejka* often describes a fashionable women garment, whereas *telogrejka* is used to describe a warm cotton quilted jacket used in the Soviet army and labor camps. This divergence of meanings may be connected with different associations of *duša* 'soul' as something fragile vs. *telo* 'body' as something earth-bound.

Трюмо [*trjumo*] 'console mirror, three-leaved mirror'. This word was borrowed from the French *trumeau*. The Russian dictionaries list two senses: (1) 'console mirror, standalone mirror', (2) 'trumeau, pillar' (in architecture). However, the word frequently refers to a three-leaved mirror. Here, again, at least two factors contributed to this meaning development: (1) the advancement of three-leaved mirrors and the lack of a one-word nomination for such an object, (2) the phonetic similarity to the word *tri* 'three'. The latter factor also influenced the development of another word, *trel'jaž*, which is now a close synonym to *trjumo*. *Trel'jaž*, too, was borrowed from the French *treillage*. The French word means 'trellis, latticework', and one of the senses of the Russian word is close to it. However, no later than in the 1930s the word acquired a new, now much more frequent sense 'console mirror, three-leaved mirror'. The factors contributing to this meaning development are probably exactly the same as in the case of *trjumo*, even though the reason of its closeness to the root

tri is entirely different. For a given speaker of contemporary Russian, the two types of mirrors can be expressed by these two words (*trjumo* and *trel'jaž*), the former meaning 'a console mirror' and the latter 'a three-leaved mirror', or vice versa: a distinction not quite described by the dictionaries.

We can see here different kinds of meaning shifts, various situations of borrowings, paronymic attraction and influence of official stock lists, inventories and industrial naming practices, resulting in rather complex and varying sets of meanings, which we decided to investigate further through a sociolinguistic survey.

2. Experiment design and participants

The questionnaire[2] was organized as follows. For each of the seven words, the participants were first asked to identify the only existing word among four possible options. The options for the word *barsetka*, e.g., were *barsetka*, *barfetka*, *baržetka*, and *barzetka*. Afterward, the participant had to choose which semantic field the word belonged to (they were presented with four options such as clothing, food, crockery, etc.), and then to choose the nearest hypernym from four given options. These three stages were used to detect how familiar the words in question are to the participants. One could proceed to the following stage only if they chose the correct option in the previous one.

Finally, in the fourth stage, the participants were presented with four pictures that represented different variants of the objects in question. Every picture was accompanied by a short description. Unlike the previous stages, this one was a multiple-choice task and had no presupposed correct answers.

1706 people participated in the study, including 1297 (76%) women, 404 (24%) men, and 5 people who did not submit the information on their gender. The median age of the participants was 32 years. We grouped the places of residence indicated by the participants into the following regions (see the map on **Figure 1**):

- Russia: Moscow (917 participants; 53,8%), Moscow Oblast (4,6%), Saint Petersburg (9,2%), the Center (11,1%), the East (9,5%), the South (1,4%), the North-West (1,8%).
- Ukraine (2,3%), Belarus (1,2%), and other countries (2,3%).

Fig. 1

[2] Available at https://goo.gl/forms/fYqG0aHHW2hOC1w73

3. Defining word knowledge

A person may know a word passively or actively, understand its meaning with certain precision and be familiar with a certain set of meanings if a word is polysemous. Here we propose some possible metrics of word knowledge that can be identified based on the data from the first three stages of our experiment and show to what extent a word is known among the participants.

These metrics are presented in Table 1. The second column represents the percentage of people who can identify a word as the only existing one among several similar strings of letters, i.e. the percentage of people who at least know that such a word exists in the language. The word *slancy* has the highest score while *kremanka* has the lowest one. The third column shows the percentage of people who know to what semantic field the word belongs, with *slancy* having again the highest score and *trenč* having the lowest score. Finally, the percentage of people who can give the nearest hypernym is presented in the fourth column. Again, *slancy* has the highest score and *trenč* and *kremanka* have the lowest scores.

While the general ranking of the seven words is more or less the same, the differences between different metrics vary to a remarkable degree. These differences are given in the last two columns of Table 1. N1–N2 shows the number of participants who correctly identified the word among similar nonce words but have no idea what it actually means. N2–N3 shows the number of participants who only know the meaning roughly.

Table 1. Different degrees of word knowledge

Word	Word identification among similar nonce words (N1)	Correct semantic field (N2)	Correct closest hypernym (N3)	N1–N2	N2–N3
slancy	99,4%	94,4%	93,7%	4,0%	0,7%
trjumo	97,4%	94,2%	84,6%	3,2%	9,6%
barsetka	95,2%	91,2%	83,0%	4,0%	8,2%
manto	91,5%	89,5%	74,4%	2,0%	15,1%
dušegrejka	87,1%	83,3%	72,6%	3,8%	10,7%
trenč	83,1%	74,6%	70,7%	8,5%	3,9%
kremanka	84,1%	78,4%	70,6%	5,7%	7,8%
mean	**91,1%**	**86,5%**	**78,5%**	**4,5%**	**8,0%**

A sociolinguistic dimension can make this picture even more complex. Below we show how our third knowledge metric (N3) depends on the social variables (gender, age, and region) using the words *kremanka* and *trjumo* as examples.

The decision tree for the knowledge of the word *kremanka* is presented in Figure 2. It divides the participants of the experiment into several groups based on the social variables. Figure 3 shows the analogous data for the word *trjumo*. The variation among the groups is not so striking. Still the difference is statistically significant.

This kind of data is missing from traditional dictionaries and can be useful for understanding that the prevalence of a particular word is not a simple quantitative variable but rather a complex entity that involves different social factors. As a whole, based on our findings, at least in the domain of everyday life lexicon, females and older people are more likely to know the words in question.

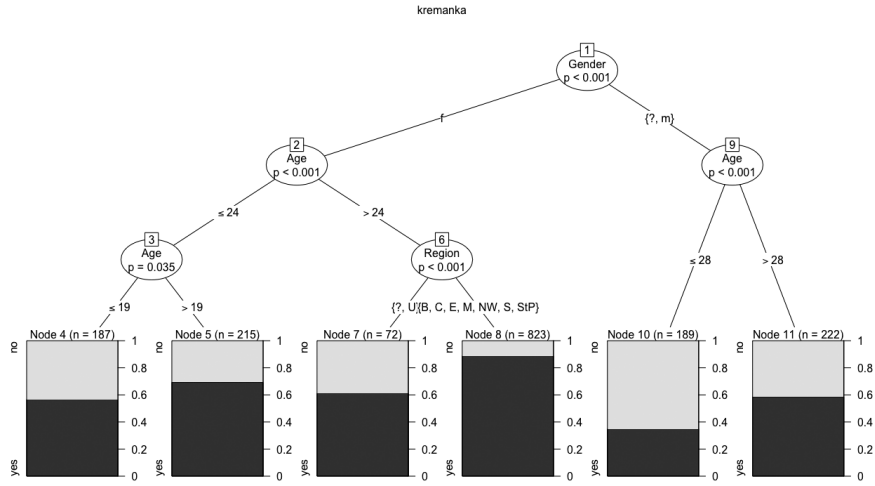

Fig. 2. Decision tree for the knowledge of the word *kremanka*

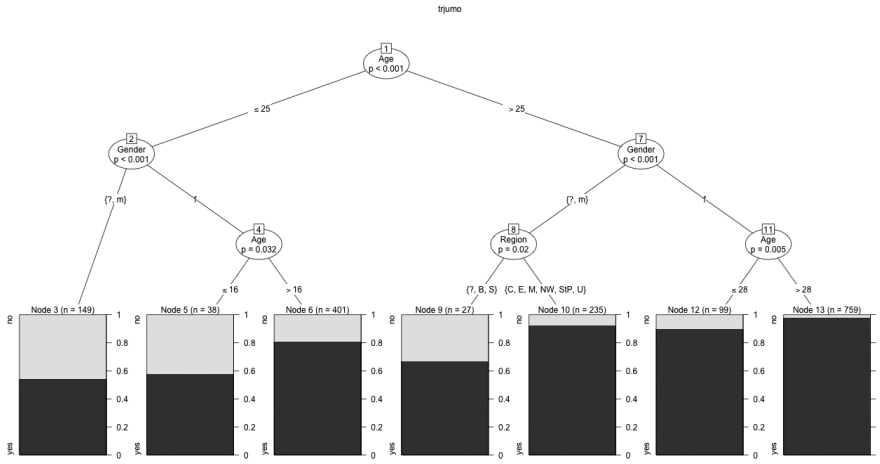

Fig. 3. Decision tree for the knowledge of the word *trjumo*

4. Is it polysemy?

One and the same word denoting an everyday life object can refer to objects different in shape, dimensions, or function. In each case, lexicographers have to decide whether to describe these differences as different dictionary senses. Here are the English translations of the sets of descriptions proposed in our experiment for the objects in question.

Slancy: (1) beach footwear with a strap between the toes; (2) beach footwear with a strap across the foot; (3) street footwear with a strap between the toes; (4) street footwear with a strap across the foot.

Barsetka: (1) man purse with a loop handle around the wrist; (2) small briefcase; (3) wallet with many pockets; (4) belt bag.

Kremanka: (1) small dessert bowl with a stem; (2) small dessert bowl without a stem; (3) small salad bowl with a stem; (4) small salad bowl without a stem.

Trenč: (1) city coat with pockets and a belt; (2) military style coat with a wide belt; (3) raincoat; (4) military coat without a belt.

Manto: (1) short elegant sleeveless fur coat; (2) long elegant fur coat with sleeves; (3) fur mantle; (3) light coat.

Dušegrejka: (1) ethnic Russian women jacket; (2) fur waistcoat; (3) cotton quilted jacket; (4) jacket with a fur collar.

Trjumo: (1) alone standing mirror; (2) dresser with a mirror; (3) dresser with three mirrors; (4) table with three mirrors.

Participants of our experiment were given these descriptions along with the pictures of these objects and were free to choose any set of them, including none or all of them. In most cases, participants chose only one option for each object, and these choices were significantly different. This may mean that most speakers have clearly defined mental images, rather than fuzzy concepts, behind these words, but these images differ across the pool of participants. This can hardly be considered true polysemy because the meanings are quite close to each other: in most cases, they have the same genus proximum, the same hypernyms and nearly the same synonyms. On the other hand, we could not provide a common definition for each meaning set in each group, because it would invariably be way too vague and broad. We would call this a case of *disperse polysemy*: a situation where several close but distinct definitions can be assigned to a word, which hardly ever coexist in a single speaker's mind, but rather in the speakers' population as a whole. Upon analyzing the distribution of these meanings, we can list them in the dictionary entry assigning labels with sociolinguistic information.

As an example of a possible analysis of how the four meanings of the word *trjumo* are organized, we provide a decision tree model that takes into account all social variables and the subsets of usages ascribed. This particular model (see **Figure 4**) shows with what probability people acknowledge that a dresser with three mirrors can be called *trjumo*. The lowest percentage (Node 8) corresponds to people older

than 26 years that acknowledge that an alone standing mirror can be called *trjumo*. It can be explained by the fact that this meaning is diachronically and semantically the most remote one from the one in question for this model. It also means that these two meanings are almost incompatible for the word *trjumo* within the lexicon of a single person (at least among the corresponding age group). The highest percentage (Node 5) is found among the people not older than 26 years who acknowledge also 'dresser with one mirror' and 'table with three mirrors', i.e. closely related meanings to the meaning in question, as possible meanings of *trjumo*.

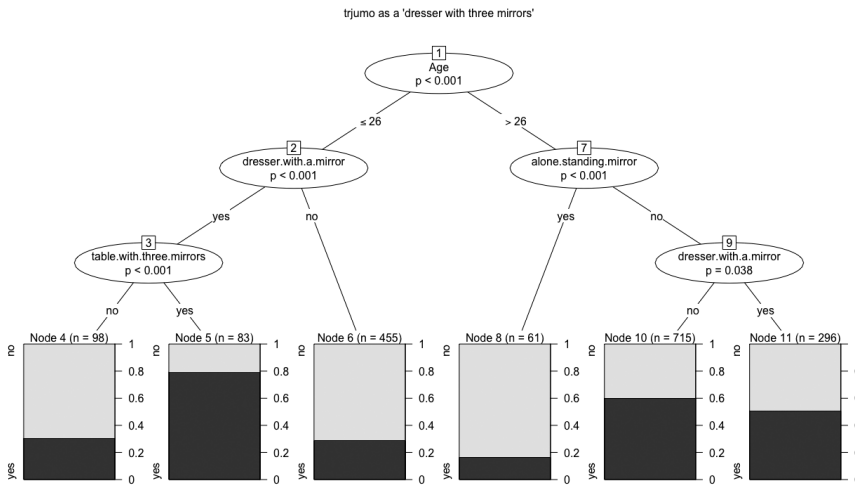

Fig. 4. Decision tree for *trjumo* as 'a dresser with three mirrors'

5. Conclusion

We believe that our findings may be useful in lexicography, lexical semantics, and language testing.

The existing dictionary entries are often insufficient and too narrow. For the purposes of our thesaurus, we intend to take into account the lexical variation and include different descriptions into the lexical entries, thus providing the dictionary users with more accurate and detailed information on current word usage in different social groups.

While we believe that the proposed notion of disperse polysemy is most characteristic for concrete nouns referring to artifacts, it can be further verified on various kinds of lexemes.

What we have shown in this paper are just several excerpts of the vast data that we collected. A further analysis can be conducted to see whether there is a correlation between participant profiles and the number of meanings they know; we could hypothesize that certain groups of respondents are better at handling polysemy than others.

The design of the multilevel survey that we created can be used for more precise testing of the vocabulary sizes both in native speakers and in language learners, if applied to mass lexical material, including much more frequent words.

References

1. *Brysbaert M., Stevens M., Mandera P. and Keuleers E.* (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. Front. Psychol. 7:1116.
2. *Chandra, G., & Dwivedi, S. K.* (2014). A literature survey on various approaches of word sense disambiguation. In Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium on (pp. 106–109). IEEE.
3. *Goulden, R., Nation, I. S. P., and Read, J.* (1990). How large can a receptive vocabulary be? Appl. Linguist. 11, 341–363
4. *Hartmann, G. W.* (1946). Further evidence on the unexpected large size of recognition vocabularies among college students. J. Educ. Psychol. 37, 436–439
5. *Iacobacci, I., Pilehvar, M. T., & Navigli, R.* (2016). Embeddings for word sense disambiguation: An evaluation study. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers, pp. 897–907.
6. *Iomdin B. L.* (2011). Materials for the thesaurus of Russian everyday life terminology. *SWEATER*: a sample dictionary entry [Materialy k slovarju-tezaurusu bytovoj terminologii. *SVITER*: obrazets slovarnoj stat'i]. Slovo i jazyk. Sbornik statej k vos'midesiatiletiju akademika Ju. D. Apresjana [The word and the language. A collection of papers to commemorate Academician Apresjan's 80th annivesrary]. Jazyki slavjanskih kul'tur, Moscow, pp. 392–406.
7. *Iomdin B. L.* (2014). Polysemous words in and out of the context. [Mnogoznachnyje slova v kontekste i vne konteksta]. Voprosy jazykoznanija [Issues in Linguistics]. Vol. 4. Moscow.
8. *Milton, J., and Treffers-Daller, J.* (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. Appl. Linguist. Rev. 4, 151–172
9. *Navigli, Roberto.* (2012). A quick tour of word sense disambiguation, induction and related approaches. In: International Conference on Current Trends in Theory and Practice of Computer Science. Springer Berlin Heidelberg.
10. *Rodd, Gaskell, Marslen-Wilson.* (2002). Making Sense of Semantic Ambiguity: Semantic competition in Lexical Access // Journal of Memory and Language 46, 245–266.