

Disambiguation of Taxonomy Markers in Context: Russian Nouns

Olga Lashevskaja

University of Tromsø
Tromsø, Norway

olesar@gmail.com

Olga Mitrofanova

St. Petersburg State University,
Universitetskaya emb. 11,
199034, St. Petersburg, Russia
alkonost-om@yandex.ru

Abstract

The paper presents experimental results on WSD, with focus on disambiguation of Russian nouns that refer to tangible objects and abstract notions. The body of contexts has been extracted from the Russian National Corpus (RNC). The tool used in our experiments is aimed at statistical processing and classification of noun contexts. The WSD procedure takes into account taxonomy markers of word meanings as well as lexical markers and morphological tagsets in the context. A set of experiments allows us to establish preferential conditions for WSD in Russian texts.

1 Introduction

Word sense disambiguation (WSD) plays a crucial role in corpora development and use. A rich variety of reliable WSD techniques such as knowledge- (or rule-) based, statistical corpus-based WSD or their hybrids have been worked out and tested [Agirre & Edmonds 2007; Mihalcea & Pedersen 2005; Navigli 2009]. Knowledge-based WSD is performed with the help of semantic information stored in electronic lexicographic modules (e.g., WordNet, FrameNet). Corpus-based WSD implies extraction and statistical processing of word collocations which makes it possible to distinguish separate meanings of lexical items in context (e.g., [Pedersen 2002; Schütze 1998], etc.). Hybrid WSD brings into action both lexical resources and corpus analysis (e.g., [Leacock et al. 1998; Mihalcea 2002], etc.).

Richly annotated corpora prove to be valuable sources of linguistic evidence necessary for exploring word meanings, their interrelations, extracting lexical-semantic classes, developing taxonomies, etc. Statistical algorithms implemented in contemporary corpora processing tools ensure extraction of information on the frequency distributions of semantic, lexical and morphological markers. These data are indispensable for classification of word contexts and, thus, for proper identification of word senses in contexts [Mitrofanova et al. 2008a, Mitrofanova et al. 2008b].

Major WSD techniques were enabled in experiments on semantic ambiguity resolution in Russian texts. The use of lexical databases for Russian (e.g., an electronic thesaurus RuTes [Lukashevich & Chujko 2007], the RNC semantic dictionary [Rakhilina et al. 2006], RussNet lexical database [Azarova et al. 2008]) provides rather high quality of WSD. If lexicographic information is not available, statistical WSD techniques are indispensable in processing Russian texts. As experimental data have shown, it is possible to identify word meanings in contexts taking into account POS tag distributions [Azarova & Marina 2006] and lexical markers [Kobricov et al. 2005]; hybrid WSD seems to be effective as well [Toldova et al. 2008].

The purpose of the present project is statistical WSD in Russian texts which entails fulfilment of certain research tasks, such as: (1) development of a WSD tool for Russian; (2) experiments on WSD in Russian texts with various parameters; (3) studying preferential conditions for WSD in Russian. It should be noted that the present study is aimed at Targeted WSD (and not All Words WSD).

The scope of the project encompasses statistical WSD procedure in three modes – with regard to three types of contextual information: (1) lexical markers of word meanings in contexts (lemmas of lexical items co-occurring with ambiguous words in contexts); (2) taxonomy markers (semantic tagsets referring to lexical-semantic classes) of context items; (3) grammatical markers (morphological tagsets referring to POS and other grammatical features) of context items – and to compare reliability of these WSD approaches. It should be noted that experiments on WSD based on semantic annotation have no precedent in Russian corpus linguistics.

2 Linguistic data

Contexts for Russian nouns referring to tangible objects and abstract notions serve as an empirical basis of the study (such polysemous and/or homonymic words as *dom* ‘building, private space, family, etc.’, *organ* ‘institution, part of body, musical instrument, etc.’, *luk* ‘onion, bow’, *glava* ‘head, chief, cupola, chapter, etc.’, *vid* ‘view, form, document, image, verbal aspect, kind, species’, *ključ* ‘key, clue, clef, spring, etc.’, *sovet* ‘advice, council, etc.’, *ploščad* ‘square, space, etc.’, *kosa* ‘braid, scythe, peninsula’, etc.). Although the nouns considered in course of experiments belong to different lexical-semantic groups, they reveal regular types of relations between meanings of polysemous words or between homonymic items. That’s why the set of words in question should be regarded as representative of noun class in general.

Sets of contexts were extracted from the Russian National Corpus (RNC, <http://www.ruscorpora.ru/>), the largest annotated corpus of Russian texts containing about 150 M tokens. The texts included in the RNC are supplied with morphological (morphosyntactic) and semantic annotation. The majority of nouns in the RNC are assigned markers according to coarse-grained taxonomy (e.g. ‘concrete’, ‘human’, ‘animal’, ‘space’, ‘construction’, ‘tool’, ‘container’, ‘substance’, ‘movement’, ‘diminutive’, ‘causative’, ‘verbal noun’, and other lexical-semantic classes, cf. <http://www.ruscorpora.ru/en/corpora-sem.html>). Taxonomy markers assigned to a particular lexical item in a context account for the set of its registered meanings, so that a WSD procedure is often required.

WSD has to be performed for nouns with various frequencies of particular meanings (cf. Table 1).

Uses of the given nouns represented in the RNC by 10 or more occurrences for each word sense were analysed. Word senses with fewer contexts in the corpus (such as *dom* ‘common space’ or *dom* ‘dynasty’) were excluded from the study. In course of experiments on Targeted WSD manual disambiguation was performed for a training set of contexts for a particular word, the remaining ambiguous contexts were subjected to statistical WSD.

3 WSD procedure

A Python-based WSD software was developed to perform statistical WSD procedure in three modes, taking into account (1) lexical markers occurring in contexts; (2) taxonomy markers of context elements; and (3) grammatical markers – morphological tagsets assigned to context elements. An automatic word clustering (AWC) tool was adapted [Mitrofanova et al. 2007]. The AWC tool facilitates formation of clusters of similar contexts extracted from the RNC. Adjustment of AWC software for WSD purposes required implementation of machine learning and pattern recognition modules.

WSD procedure is carried out in stages. The first stage implies pre-processing of contexts in experimental set *E*. Semantically and morphologically unambiguous contexts are selected to form a training set *S* required for machine learning, while ambiguous contexts are treated as a trial set *T*. Machine learning is performed at the second stage. For each meaning of a word its statistical pattern is formed taking into account frequencies of taxonomy markers, lexical markers and morphological tagsets of context elements. Further, patterns of meanings, as well as trial contexts, are represented as vectors in a word space model. The third stage implies pattern recognition, i.e. selection of patterns nearest to vectors that correspond to ambiguous contexts. Three similarity measures based on the distance between patterns and vectors of trial contexts are calculated in different ways, so that the user can choose between Hamming measure, Euclidean measure, and Cosine measure. As a result, meanings exposed by particular patterns are automatically assigned to processed contexts.

Table 1. Russian nouns *dom*, *organ*, *luk*, *vid*, *glava*: taxonomy markers and frequencies of meanings (number of contexts in the RNC)

Word meanings and taxonomy markers	Number of contexts in the RNC
dom	3000 (total)
<i>dom</i> 'building' <r:concr t:constr top:contain>	1694
<i>dom</i> 'private space' <r:concr t:space>	95
<i>dom</i> 'family' <r:concr t:group pt:set sc:hum>	72
<i>dom</i> 'common space' <r:concr t:space der:shift der:metaph>	4
<i>dom</i> 'institution' <r:concr t:org>	292
<i>dom</i> 'dynasty' <r:concr pt:set sc:hum>	1
<i>dom</i> (merged meanings)	842
organ	834 (total)
<i>organ</i> 'institution' <r:concr t:org hi:class>	660
<i>organ</i> 'part of body' <r:concr pt:partb pc:hum pc:animal hi:class>	130
<i>organ</i> 'musical instrument' <r:concr t:tool:mus>	27
<i>organ</i> 'means' <r:concr der:shift dt:partb>	9
<i>organ</i> 'publication' <r:concr t:media hi:class>	8
luk	2200 (total)
<i>luk</i> 'onion' <r:concr t:plant t:fruit t:food pt:aggr>	1600
<i>luk</i> 'bow' <r:concr t:tool:weapon top:arc>	600
vid	2866 (total)
<i>vid</i> 'view' <r:abstr t:perc der:v>	1144
<i>vid</i> 'form' <r:abstr der:shift>	1075
<i>vid</i> 'document' <r:concr t:doc>	7
<i>vid</i> 'image' <r:concr t:workart>	10
<i>vid</i> 'expectation' <r:abstr t:ment>	10
<i>vid</i> 'kind, species' <r:abstr r:concr pt:set sc:X>	617
<i>vid</i> 'verbal aspect' <r:abstr>	3
Word meanings and taxonomy markers	Number of contexts in the RNC
glava	1073 (total)
<i>glava</i> 'head, part of body' <r:concr pt:partb pc:hum>	8
<i>glava</i> 'leading position' <r:concr der:shift dt:partb>	140
<i>glava</i> 'cupola' <r:concr pt:part pc:constr>	12
<i>glava</i> 'chief' <r:concr t:hum>	301
<i>glava</i> 'chapter' <r:concr t:text pt:part pc:text> ¹	612

¹ In this table, the following semantic tags are used: 1) top categories *r:concr* (concrete noun), *r:abstr* (abstract noun); 2) taxonomic classes *t:hum* (human beings), *t:org*

Series of tests were performed (1) to evaluate several parameters that can influence test results: context window size, proportional expansion of training sets of contexts for each meaning, etc.; (2) to estimate correlation between taxonomic, lexical and morphological criteria, to compare reliability of these WSD approaches and to ascertain preferential conditions of their application.

Evaluation of WSD quality was performed: results of automatic WSD were compared with results of manual WSD, precision *P* and recall *R* were defined in all series of tests.

4 General results of experiments

Thorough analysis of contexts shows that the appropriate choice of similarity measure (Cosine measure) alongside with expansion of a training set ($S = 100 \dots 500$ contexts) ensures over 85% correct decisions on average ($P \approx 0.85$). Under such conditions, in series of experiments the number of correct decisions turned out to be no less than 50...60% ($P \approx 0.50 \dots 0.60$), in some cases up to 95...100% ($P \approx 0.95 \dots 1$).

The Cosine measure proves to be the most reliable similarity measure as it is the least sensitive to meaning frequencies. Hamming and Euclidean measures provide correspondingly 45% ($P \approx 0.45$) and 65% ($P \approx 0.65$) of correct decisions on average.

WSD experiments were performed with training sets of variable size $S = 10, 15, 55, 75, 100, 200, 500, \dots$ (up to all contexts except for those included in a trial set) and with proportional expansion of a training set S being 10%, 15%, 20% of E . It seems that the training set S should contain at least 100 unambiguous contexts, while 500 contexts provide the best results. In general, to obtain reliable WSD results, the training set size S should be no less than 20% of the experimental set size E . In other cases the amount of correct deci-

(organizations), *t:constr* (buildings/constructions), *t:space* (space/ places), *t:tool:mus* (musical instruments), *t:perc* (perception), *t:ment* (mental sphere), etc.; 3) mereological classes *pt:partb pc:hum pc:animal* (body parts of humans and animals), *pt:part pc:constr* (parts of buildings/constructions), *pt:set sc:hum* (sets of humans); 4) topological classes *top:contain* (containers), *t:arc* (arcs); 5) derivational markers *der:v* (deverbal nouns), *der:shift dt:partb* (semantic shift from the name of a body part), *der:shift der:metaph* (metaphorical shift).

ons may be reduced because statistical patterns for meanings turn out to be rather ‘blurry’.

A series of tests with variable context window size w ($[-i; +k]$, $i, k \leq N$ (N – context length) was carried out, so that the context window could be symmetric or asymmetric, and could be limited to a clause or a syntactic group. Context analysis with regard to syntactic relations showed an increase in WSD precision by $P = 0.05 \dots 0.1$. The best results can be expected if $i \leq 2$, $2 \leq k \leq 4$. In most cases such context window corresponds to noun groups including prepositional (adjectival) and postpositional (nominal, infinitival, etc.) determiners which contain information relevant for meaning disambiguation.

5 WSD based on taxonomy markers, on lexical markers and on morphological tagsets: discussion

Experiments on WSD based on taxonomy markers and on lexical markers gave rather encouraging results. E.g., WSD procedure for the noun *luk* allows to discriminate meanings *luk* ‘onion’ and *luk* ‘bow’ given $P \approx 0.825 \dots 0.85$ on average, cf. Table 2.

Table 2. Results of WSD based on taxonomy markers and on lexical markers for the noun *luk*

	Amount of correct decisions for separate meanings (P)		Average
	<i>luk</i> ‘onion’	<i>luk</i> ‘bow’	
	WSD based on taxonomy markers	0.75	
WSD based on lexical markers	0.75	0.90	0.825

For the most part, WSD based on taxonomy markers and on lexical markers was equally effective: cf. Table 3, e.g. context (c). At the same time, processing of contexts which takes into account taxonomy markers often provides more trustworthy decisions: e.g., the increase of Cosine measure value is noticeable in context (a) where the meaning *luk* ‘onion’ was recognized correctly with the help of both criteria. WSD based on taxonomy markers also helps to evade erroneous interpretations: cf. contexts (b) and (d) where the meaning of *luk* was chosen correctly in case of WSD based on taxonomy markers.

Table 3. Examples of WSD based on taxonomy markers and on lexical markers for the noun *luk*

<i>luk</i>	WSD based on taxonomy markers		WSD based on lexical markers	
	Meaning	Cos	Meaning	Cos
(a) <i>luk</i> ‘onion’ Pomn’u hleb s iz’umom, s <i>lukom</i> , s kakimi-to koren-jami. ([I] remember bread with raisins, with <i>onion</i> , and with some spices.)	<i>luk</i> ‘onion’	0.786	<i>luk</i> ‘onion’	0.572
(b) <i>luk</i> ‘onion’ Nachinajut prini-mat’ <i>luk</i> , kapustu... ([they] begin to eat <i>onion</i> , cabbage...)	<i>luk</i> ‘onion’	0.514	<i>luk</i> ‘weapon’	0.502
(c) <i>luk</i> ‘weapon’ Odni tugije <i>luki</i> , nad kotorymi neskol’ko chelovek spravit’sa ne mogli, ‘igrajuchi’ nat’agival’i... (Some [people] ‘effortlessly’ bent tight bows with which several people couldn’t cope with...)	<i>luk</i> ‘weapon’	0.550	<i>luk</i> ‘weapon’	0.533
(d) <i>luk</i> ‘weapon’ Za spinoj u nego viseli <i>luk</i> i kolchan. (He had a bow on his back.)	<i>luk</i> ‘weapon’	0.517	<i>luk</i> ‘onion’	0.500
...

Comparison of WSD results obtained in three modes shows that in general morphological criteria prove to be more reliable than taxonomic and lexical criteria: average P and R for WSD based on morphological annotation are higher than for WSD based on taxonomy markers and on lexical markers. At the same time, differences in WSD results lead to the conclusion that various types of context-dependent meanings determine preferential conditions for application of WSD approaches (cf. example in Table 4).

The correlation between taxonomic, lexical and morphological criteria for WSD was estimated. The Pearson’s correlation coefficient is quite low: $|Corr| < 0.4$. Thus, criteria in question should be considered as independent. It is expected that WSD based on combinations of criteria (combinations of taxonomy markers and lexical markers, taxonomy

markers and morphological tagsets, etc.) may be more effective.

Table 4. Examples of WSD results obtained in three modes for the noun *vid*: window size w [-5, +5], [-5, +1], [-1, +5]; training set size $S = 20\% E$

p	WSD based on taxonomy			WSD based on lexical markers			WSD based on morphological tagsets		
	<i>vid</i> 'view'	<i>vid</i> 'shape'	<i>vid</i> 'kind'	<i>vid</i> 'view'	<i>vid</i> 'shape'	<i>vid</i> 'kind'	<i>vid</i> 'view'	<i>vid</i> 'shape'	<i>vid</i> 'kind'
[-5,+5]	0.4	0.5	0.7	0.75	0.8	0.5	0.65	0.9	0.8
[-5,+1]	0.95	0.35	0.75	0.7	0.85	0.65	0.6	0.95	0.85
[-1,+5]	0.25	0.8	0.75	0.65	0.7	0.85	0.65	0.9	0.85

6 Additional data for meaning identification

WSD procedure also furnished us with additional information relevant for meaning identification, namely, sets of lexical markers of different meanings deduced from contexts (cf. Table 5). In most cases combinations of a word with its lexical markers should be considered as collocations.

Table 5. Lexical markers of meanings induced from contexts for the noun *organ*

Word meanings	Lexical markers
<i>organ</i> 'institution'	uchrezhdenije 'institution', samoupravlenije 'self-government', nachal'nik 'boss', mestnyj 'local', pravoohranitel'nyj 'law-enforcement', etc.
<i>organ</i> 'part of body'	porok 'defect', vrožd'onnyj 'innate', etc.

7 Analysis of errors

Most errors registered in WSD experiments can be explained by insufficiency of contextual information for meaning identification. WSD results for such contexts often show Cosine measure values about 0.500 (cf. contexts (b) and (d), Table 3). Failures in WSD may also be explained by the use of disambiguated words in constructions and set-expressions, cf. context (e) below:

(e) Poroj Elene kazalos', chto vse javlenija i vse predmety mozno opisat' v treh pozicijah: anfas, profil', *vid* sverhu.

(At times it seemed to Elena that all phenomena and all objects can be described from three positions: front [view], profile, view from above.)

Manual WSD: *vid* 'view'

WSD in three modes: *vid* 'kind'

8 Analysis of merged meanings

It is hardly possible to provide unambiguous analysis of certain contexts for some polysemous nouns revealing merged meanings. For example, a noun *dom* forms pairs of meanings which are almost indistinguishable in certain contexts: *dom* 'building & personal space (home)', *dom* 'personal space & family', etc. Of 3 000 contexts for a noun *dom* there are 842 contexts where ambiguity can't be completely resolved. In such cases WSD results compared with manual analysis make it possible to determine a dominating semantic feature in a pair of merged meanings, cf. contexts (f) and (g), Table 6.

Table 6. Analysis of merged meanings for the noun *dom*: WSD based on lexical markers

<i>dom</i>	Manual analysis	WSD results	<i>Cos</i>
(f) ... v dome u Jozhika topilas' pech... (... in Jozhik's house the stove was burning...)	<i>dom</i> 'building & personal space'	<i>dom</i> 'building'	0.429
(g) Rodstvenniki u Livii... ludi praktichnyje... jedinstvennyj chelovek, kotoryj uvazhajet jejo v etom dome, – eto jejo dvorekij... (Livia's relatives ... are practically-minded people ... the only person who respects her in this house is her butler...)	<i>dom</i> 'personal space & family'	<i>dom</i> 'family'	0.452

In further experiments additional statistical patterns corresponding to merged meanings were introduced to improve the performance of the WSD system.

9 Conclusion

A set of experiments on statistical WSD were successfully carried out for contexts of polysemous and/or homonymic Russian nouns which had been extracted from the RNC.

WSD was performed in three modes – taking into account (1) lexical markers occurring in contexts; (2) taxonomy markers of context elements; and (3) grammatical markers – morphological tagsets assigned to context elements. All these approaches proved to be reliable, although in controversial cases preference should be given to WSD based on taxonomy markers.

Optimal conditions for WSD in Russian texts were discovered: over 85% (in some cases up to 95%) correct decisions may be achieved through the use of Cosine measure, a training set varying from 100 up to 500 contexts that constitutes at least 20% of the experimental set E , context window size $w [-i; +k]$ where $i \leq 2$, $2 \leq k \leq 4$.

Further work implies (1) enrichment of WSD software; (2) experiments on WSD based on complex criteria (combinations of taxonomy markers and lexical markers, taxonomy markers and morphological tagsets, etc.); (3) verification of particular linguistic and statistical hypotheses on WSD in Russian texts. The experiments involving machine learning and pattern recognition put into action the key ideas of cognitive semantics which turn out to be of competitive advantage. It is assumed that words of the same lexical-semantic class (which also share the same place in the taxonomy) reveal similar frequency distributions of context features. Thus, WSD for polysemous words of a certain lexical-semantic class (presumably, its core members) may be performed on the basis of the training set of contexts which was previously formed for monosemous (presumably, peripheral) words of the class. It is expected that this approach to WSD may simplify the procedure of selection and analysis of training data (which is time-consuming).

The work discussed in the paper demonstrates practical application of theoretical cognitive linguistics in NLP. Two hypotheses, on entrenchment of word senses in particular context frames [Brooks et al. 1999] and on center (prototype) – periphery structure of lexical semantic categories [Lakoff 1987], proved to be valid in the course of the verification procedure. It appears that these ideas contribute much to the development of effective WSD techniques.

References

Agirre, E. & Ph. Edmonds (eds.). (2007). *Word Sense Disambiguation: Algorithms and Applications*. Text, Speech and Language Technology, Vol. 33. Springer-Verlag, Berlin, Heidelberg, New York.

Azarova, I.V. & A.S. Marina. (2006). Avtomatizirovannaja klassifikacija kontekstov pri podgotovke dannyh dl'a kompjuternogo tezaurusa RussNet. In: *Kompjuter'naja lingvistika i intellektual'nyje tehnologii: Trudy mezhdunarodnoj konferencii Dialog-2006*. Moscow. P. 13–17.

Azarova, I.V., S.V. Bichineva, & D.T. Vakhitova. (2008). Avtomaticheskije razreshenije leksicheskij neodnoznachnosti chastotnyh suschestvitelnyh (v terminah strukturnykh jedinic RussNet). In: *Proceedings of the International Conference Corpora 2008*, St. Petersburg, Russia, October 6–10, 2008. P. 5–8.

Brooks, P., M. Tomasello et al. (1999). Children's overgeneralization of fixed transitivity verbs: The entrenchment hypothesis. In: *Child Development*, 70. P. 1325–1337.

Kobricov, B.P., O.N. Lashevskaja, & O.Ju. Shemanajeva. (2005). Sn'atije leksiko-semanticheskij omonimii v novostnyh i gazteno-zhurnal'nyh tekstah: poverhnostnyje fil'try i statisticheskaja ocenka. In: *Internet-matematika 2005: Avtomaticheskaja obrabotka web-dannyh*. Moscow. P. 38–57.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago, IL: University of Chicago Press.

Leacock, C., G.A. Miller, & M. Chodorow. (2002). Using Corpus Statistics and WordNet Relations for Sense Identification. In: *Computational Linguistics*, 24(1). P. 147–165.

Lukashevich, N.V. & D.S. Chujko. (2007). Avtomaticheskije razreshenije leksicheskij mnogoznachnosti na baze tezaurusnyh znanij. In: *Internet-matematika 2007*. Ekaterinburg. P. 108–117.

Mihalcea, R. (2002). Word Sense Disambiguation Using Pattern Learning and Automatic Feature Selection. In: *Journal of Natural Language and Engineering (JNLE)*, December 2002.

Mihalcea, R. & T. Pedersen. (2005). Word Sense Disambiguation Tutorial. URL: <http://www.d.umn.edu/~tpederse/WSDTutorial.html>

Mitrofanova, O., P. Panicheva & O. Lashevskaja. (2008a). Statistical Word Sense Disambiguation in Contexts for Russian Nouns Denoting Physical Objects. In: *Text, Speech and Dialogue. Proceedings of the 11th International Conference TSD 2008*, Brno, Czech Republic, September 8–12, 2008. Springer-Verlag. P. 153–159.

Mitrofanova, O., A. Mukhin, P. Panicheva, V. Savitsky. (2007). Automatic Word Clustering in Russian Texts. In: *Matoušek, V., Mautner, P. et al. (eds.): Text, Speech and Dialogue. Proceedings of the Tenth International Conference TSD 2007*, Pilsen, Czech Republic, September 3–7, 2007. Lecture Notes in Artificial Intelligence, Vol. 4629. Springer-Verlag. P. 85–91.

Mitrofanova, O., O. Lashevskaja & P. Panicheva. (2008b). Eksperimenty po statisticheskomu razresheniju leksiko-semanticheskij neodnoznachnosti russkix imen suschestvitel'nyx v korpuse. In: *Proceedings of the International Conference Corpora 2008*, St. Petersburg, Russia, October 6–10, 2008. P. 284–293.

- Navigli, R.* (2009). Word Sense Disambiguation: A Survey. In: ACM Computing Surveys, 41(2), 2009. P. 1–69.
- Pedersen, T.* (2002). A Baseline Methodology for Word Sense Disambiguation. (Pedersen) In: Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, pp. 126-135, February 17-23, 2002, Mexico City. P. 126–135.
- Rahilina, E.V., B.P. Kobricov, G.I. Kustova, O.N. Lashevskaja, & O.Ju. Shemanajeva.* (2006). Mnogoznachnost' kak prikladnaja problema: leksiko-semanticheskaja razmetka v Nacional'nom korpuse russkogo jazyka. In: Kompjuternaja lingvistika i intelektual'nyje tehnologii: Trudy mezhdunarodnoj konferencii Dialog 2006. Moscow. P. 445–450.
- Schütze, H.* (1998). Automatic Word Sense Discrimination. In: Computational Linguistics, 24(1). P. 97–123.
- Toldova, S.Ju., G.I. Kustova, & O.N. Lashevskaja.* (2008). Semanticheskije fil'try dl'a razreshenija mnogoznachnosti v nacional'nom korpuse russkogo jazyka: glagoly. In: Kompjuternaja lingvistika i intelektual'nyje tehnologii: Trudy mezhdunarodnoj konferencii Dialog–2008. Moscow. P. 522–529.