Semantic Annotation Layer in Russian National Corpus: Lexical Classes of Nouns and Adjectives

Olga N. Lashevskaja, Olga Yu. Shemanaeva

Institute for Scientific and Technical Information (RAS), Russian State University for Humanities

125190, Moscow, Usievicha st., 20

olesar@gmail.com, shemanaeva@yandex.ru

Abstract

The paper describes the project held within Russian National Corpus (http://www.ruscorpora.ru). Beside such obligatory constituents of a linguistic corpus as POS (parts of speech) and morphological tagging RNC contains semantic annotation. Six classifications are involved in the tagging: category, taxonomy, mereology, topology, evaluation and derivational classes. The operating of the context semantic rules is shown by applying them to various polysemous nouns and adjectives. Our results demonstrate semantic tags incorporated in the context to be highly effective for WSD.

1. Introduction

Russian National Corpus is a collection of written and spoken texts (since XVIII c.) that currently contains over 150 million tokens. RNC bears a variety of annotation layers, among those are parts of speech (POS), morphological, accentological, morphosyntactic tagging, but what is striking for a corpus of such size is its semantic markup.

The semantic annotation runs in word by word mode and implies that each word in the lexicon is semantically classified and is given several tags, corresponding to a certain lexical class (e.g. 'motion', 'time', 'sound', 'colour', 'parts of the body', etc.).

The working annotation of this kind provides a wide range of advanced possibilities beyond lexical and grammatical search. For example, it allows queries for lexical constructions, co-occurrence patterns and government of semantically characterized classes of verbs.

Unlike M. Davies's customized lists of words relating to a certain topic (Davies 2005) we offer ready-to-use classes that are traditionally involved in semantic researches. Some lexical classes consist of two thousand elements (for example, verbs of motion) and even more. Our approach is close to the USAS system (Piao et al. 2005) and to the lexical classification of FrameNet (Ruppenhofer 2006) elaborated for English corpora. Another alternative approach to WSD, where machine learning software for WSD is developed on statistical processing and classification of noun contexts, is presented in (Mitrofanova et al. forthcoming).

In this paper, we describe semantic annotation of nouns and qualitative adjectives. In Section 2 we present the variety of classifications in the semantic database, their sources and principles. In Section 3 we deal with polysemy and describe semantic filters used in word-sense disambiguation.

2. Lexical resources and principles of classification

Our semantic tool provides full-text annotation, so the semantic database (that currently contains 375 000 elements) is being constantly extended. Each meaning of polysemous words has separate entry in the database and is classified manually on the base of definitions given in explanatory dictionaries of Russian (Ozhegov 1992, Evgenjeva 1999). Besides that, data from Babenko's (2005) and Shvedova's (2004) semantic dictionaries have been analyzed, although they have other structure of lexical classes.

There is so far no such database as WordNet for Russian, though the research is being done in that domain, cf. (Azarova 2006).

The principles of lexical classification in RNC are based on the project "Lexicograph" (http://www.lexicograph.ru) supervised by E. Paducheva and E. Rakhilina. On the theoretical ideas of the lexicon hierarchy see Kustova et al. (forthcoming). It should be underlined that the classification follows the multi-facet principle: there are several classifications (some of them hierarchical) independent of one another. At present, 6 classifications are involved in the annotation:

- **Category** (prime lexical divisions that determine main semantic features: concrete, abstract, proper nouns; qualitative, relative, possessive adjectives);
- **Taxonomy** (e.g. *luk* 'bow': «weapon», *radost'* 'joy': «emotion», *bystryj* 'quick': «speed», *staryj* 'old': «age»);
- **Mereology** (e.g. *rukav* 'sleeve': «parts of clothes», *buket* 'bunch': «sets and aggregates», *kaplja* 'drop': «quanta and portions of stuff»);

- **Topology** (e.g. *kastrjula* ‘pot’: «container», *nos* ‘nose’ «juts», *zmeja* ‘snake’ «ropes»);
- **Evaluation** (e.g. *blagouxanije* ‘odor’: «positive», *presmykat’sja* ‘lick the boots’: «negative»);
- **Derivational classes** (e.g. *knizhechka* ‘little book’: «diminutives», *sosnovyj* ‘piny’: «adjectives derived from nouns».

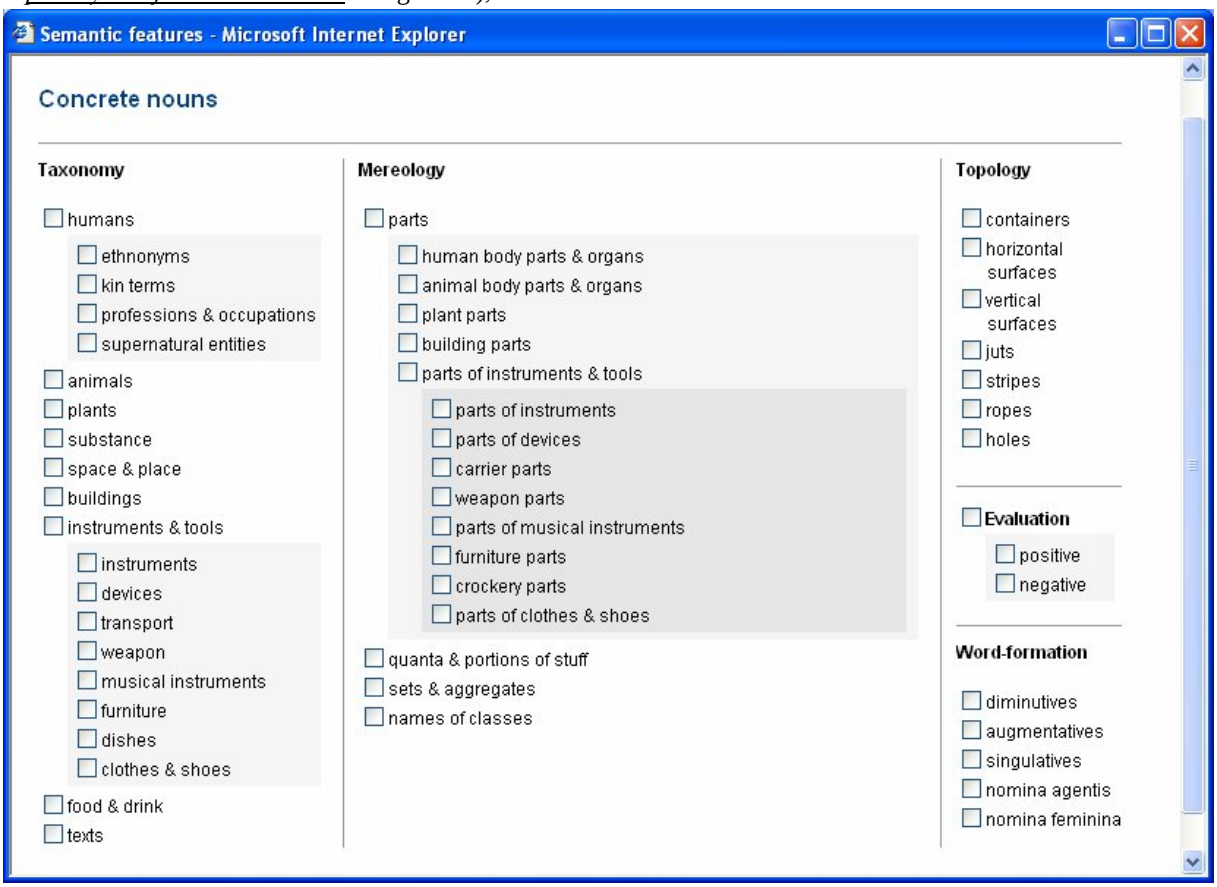

Fig. 1. Semantic classes of concrete nouns.

Classes of concrete nouns (i. e. names that refer to physical objects) which can now be obtained from the corpus are shown in Fig.1.

The system of taxonomic classes is rather elaborated. It includes size, distance, quantity, time, physical and human properties for adjectives; people, animals, plants, buildings, devices, stuff, texts, food and drinks for concrete nouns; first and last names, patronymic names and toponyms for proper nouns; classes of abstract nouns are inherited mainly from verb and adjective hierarchies and include movement, impact, speech, human properties, colour, temperature, diseases, sports, parameters, etc.

Mereological annotation gives distinction between parts of the body, parts of instruments, clothes and other things as well as quanta & portions of stuff and phases of processes. The feature of sets and aggregates are used for such words as *set, bunch, furniture, mankind*. Nouns like *animal, fruit, instrument, name* that denote categories of the world belong to the «names of classes» group.

The notion of topological types was put forward by L.Talmy (1983), who has demonstrated their significance for the understanding of linguistic structures that describe space and shape as well as undoubted cross-linguistic relevance of geometric features. Names of physical objects associated to such topological types as «horizontal spaces», «containers», «juts», «ropes», etc. occur to be sensible to space operators, such as adjectives of form and size, prepositions, verbs and nouns which refer to form, location, and motion.

Lexical meanings that have positive or negative connotations form two classes in the category of Evaluation. Derivational classes include words in which semantic components are introduced by a certain prefix or suffix or words derived from other parts of speech and what is more, from a particular semantic class of a particular POS (e.g. nouns derived verbs; adjectives derived from names of substance).

Though the features are organized hierarchically they are not inherited but assigned according to meaning of the word in the dictionary.

3. Polysemy and word-sense disambiguation

Each use of a given word in the corpus is automatically assigned all the tags that the word has in the dictionary. For example, adjective *protivnyj* (‘contrary’ & ‘unpleasant’ & ‘opposite’) has the following tag sets:

{«relative», «place»/«direction»} >>> *protivnyj veter* ‘contrary wind’
{«qualitative», «negative»} >>> *protivnyj zapax* ‘unpleasant smell’
{«relative»} >>> *protivnyje storony* ‘opposite parties’
Adjective *bol’shoj* (‘big’ & ‘grown-up’ & ‘important’) has the following tag sets:
{«qualitative», «size»} >>> *bol’shoj muzhchina* ‘big man’
{«qualitative», «age»} >>> *bol’shoj mal’chik* ‘grown-up child’
{«qualitative», «metaphor_size»} >>> *bol’shoj chelovek* ‘big | important person’.

To avoid the polysemy in RNC we formulate special rules, the so-called filters, which assign to the word the only meaning appropriate in the corresponding context. After the filter has been applied, we have three semantic fields ascribed to the word: SEM (tag set that characterizes the first meaning listed in the dictionary), SEM2 (tag sets associated with other meanings) and SEMF (tag set(s) of disambiguated meaning).

The rules are formulated manually on the n-grams database (2 and 3 unique word clusters with associated frequency, POS and semantic tags).

The disambiguating filters deal with the following information:

1) grammatical tags of the target word (case; number; full or short, comparative, superlative forms of adjectives);
2) POS and grammatical tags of elements in the context (animate vs. non-animate nouns, prepositions, participles);
3) semantic tags of neighbour words (e.g. «motion», «time», «sound», «colour», «place», «emotions», «parts of the body», «hair», «animals», «plants», «texts», «relatives», «professions», «stuff», etc.);
4) lemmas and word forms in the context (for very frequent collocations which can not be formulated in terms of grammatical and semantic classes);
5) punctuation marks (comma, hyphen, etc.);
6) word order and distance between the target word and other words that constitute a collocation;

A database of Russian multi-word expressions is used as well to discriminate the meaning or to establish its bleaching in stable prepositional collocations, idioms and so on.

The main theory which is used to deal with the changing word meanings in the corpus is Construction Grammar (cf. Fillmore et al. 1987, Goldberg 1995). According to the Construction Grammar, the speakers use constructions rather than combine words into constructions ad hoc. Constructions can lead to the meaning shift of the lexemes: the given meaning of the lexemes is coerced by the construction (see Rakhilina et al. 2007).

10,7% tokens in RNC are adjectives, and half of them is ambiguous. Semantic filters cover 500 000 occurrences of the most widely spread adjectives.

otlichnyj ‘excellent’, ‘different’

target word	conditions	WSD
otlichnyj	+ *ot* ‘from’, cf. *otlichnyj ot drugix* ‘different from the others’	«relative»
otlichnyj	by default, cf. *otlichnyj den’* ‘excellent day’	«qualitative», «positive»

grubyj ‘coarse’, ‘rude’

grubyj	+ «stuff», cf. *grubaja tkan’* ‘coarse fabric’	«relative», «physical property»
grubyj	+ «humans», cf. *grubyj čelovek* ‘rude man’	«qualitative», «human property», «negative»

As for semantic filters for nouns, let us consider their mechanisms in the example of the names of shape. This is a new class in the semantic database, and it can be viewed as a kind of mereology. Similar to names of quanta, names of shape pose themselves as transparent nouns (Ruppenhofer et al. 2006: 66), cf. *Ugol zat’anut **mnogougol’nikom** pautiny* ‘The corner is laced with **a polygon** of cobweb’ = ‘laced with cobweb’.

These nouns are usually polysemous, their donating domains are instruments, tools, natural objects and other things and parts with prominent shape, cf. *kol’co* ‘ring’, *oblako* ‘cloud’, *sigara* ‘cigar’. Below we provide several semantic filters for the noun *kol’co* in such contexts as *kol’ca dyma*, *dym kol’cami ‘rings of smoke*’, *v kol’ce ulic* ‘in the ring of lanes’, *zhivoje kol’co* ‘ring of people, lit. live ring’:

target word	conditions	WSD
kol’co	+ S&gen&«stuff»	«form», «quanta & portions»
kol’co	sg + S&gen&pl&inan&«concrete noun»	«form»
kol’co	zhivoj +	«form», «multiword exp.»

Due to the fact that the frequency of semantic tags in collocations is higher than the frequency of lemmas, semantic information allows us to reduce number of context rules.

4. Conclusion

The aim of our work is to distinguish the different meanings of words thus providing the users of the corpus with the semantically disambiguated texts. As a result users will have a possibility to organize the search by the

first meaning of the word, by the other meanings listed in the dictionary, or by the disambiguated meaning. In our research we prove semantic tags useful for word-sense disambiguation and organization of lexicon in terms of Construction Grammar. In the future we plan to compare two approaches to disambiguation, pure lexical and lexico-semantic, in statistical WSD tool.

Acknowledgements

The project of semantic annotation and word-sense disambiguation in RNC is supported by Russian Academy of Sciences and RFFI foundation (ref. 05-06-80396). The work on topological classes of nouns is funded under RGNF project (ref. 07-04-00240).

References

1. Archer, D., P. Rayson, S. Piao, T. McEnery (2004). Comparing the UCREL semantic annotation scheme with lexicographical taxonomies. In G. Williams & S. Vessier (eds.) *Proceedings of the 11th EURALEX*, Lorient, France, 6-10 July 2004. Université de Bretagne Sud. Vol. III, 817-827.
2. Azarova, I.V., Marina, A.S.: Avtomatizirovannaja klassifikacija kontekstov pri podgotovke dannyh dl'a kompjuternogo tezaurusa RussNet. In: Kompjuternaja lingvistika i intellektual'nyje tehnologii: Trudy meždunarodnoj konferencii "Dialog–2006". Moscow (2006) 13–17
3. Babenko, L. G. (2005). *Bol'shoj Tolkovyj Slovar' Russkix Suscestvitel'nyx.* Moscow.
4. Davies, Mark (2005). The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics* 10: 301-328.
5. Evgen'jeva, A. P. (ed.) (1999). *Slovar' Russkogo Jazyka v 4-x t.* Moscow. – http://feb-web.ru/feb/mas/mas-abc/default.asp
6. Fillmore, Charles J., Paul Kay & M.C. O'Connor (1987). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64, 501-538.
7. Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.
8. Grishina E.A. & Rakhilina E.V. (2005). Russian National Corpus (RNC): an overview and perspectives. In *AATSEEL 2005,* Washington, 27-30 December 2005.
9. Kilgarriff, A. (1991). Corpus word usages and dictionary word senses: What is the match? An empirical study. In *Proceedings of the 7th conference, UW Centre for the New OED & Text Research Using Corpora*.
10. Kustova, G. I., O. N. Lashevskaja, E. V. Paducheva & E. V. Rakhilina (forth-coming). Verb taxonomy: From theoretical lexical semantics to practice of corpus tagging. In B. Lewandowska-Tomaszczyk, K. Dziwirek (eds.), *Cognitive Corpus Linguistics*. Frankfurt (forthcoming).
11. Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago etc.: and London: Chicago University Press.
12. Lashevskaja O. (2006). Corpus-aided construction grammar: Semantic tools in the Russian National Corpus. In *Proceedings of the Second International Conference of the German Cognitive Linguistics Association*, Munich, 5-7 October 2006.
13. Mitrofanova O., Panicheva P., Lashevskaja O. Statistical Word Sense Disambiguation in Contexts for Russian Nouns Denoting Physical Objects. // TSD-2008 (forthcoming).
14. Ozhegov, S. I. & N. Ju. Shvedova (eds.) (1992). *Slovar' Russkogo Jazyka.* Moscow.
15. Piao, Scott S.L., Dawn Archer, Olga Mudraya, Paul Rayson, Roger Garside, Tony McEnery, Andrew Wilson (2005). A large semantic lexicon for corpus annotation. In *Proceedings of the COLING 2005*, July 14-17, Birmingham, UK.
16. Rakhilina, E.V., T. I. Reznikova, O. Yu. Shemanaeva (2007). Dealing with polysemy in Russian National Corpus: the case of adjectives. In *Tbilisi 2007: Seventh International Tbilisi Symposium on Language, Logic and Computation*.
17. Rakhilina, E. V., B. P. Kobritsov, G. I. Kustova, O. N. Lashevskaja & O. Yu. Shemanaeva. Mnogoznachnost' kak prikladnaja problema: Leksiko-semanticheskaja razmetka v NKRJa. In N. I. Laufer, A. S. Narin'jani, V. P. Selegej (eds.). *Kompjuternaja lingvistika i intellektual'nyje texnologii: Trudy mezhdunarodnoj konferencii Dialog'2006*, 445-450.
18. Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson, J. Scheffchyk. *FrameNet II: Extended Theory and Practice.* Available from http://framenet.icsi.berkeley.edu/book/book.pdf.
19. Talmy, Leonard (1983). How language structures space. In H. Pick and L. Acredolo (eds.), *Spatial Orientation: Theory, Research, and Application*. New York: Plenum Press, 225-282.