RussNet: Building a Lexical Database for the Russian Language

Irina Azarova, Olga Mitrofanova, Anna Sinopalnikova, Maria Yavorskaya, Ilya Oparin

Applied Linguistics Department, Philological Faculty, Saint-Petersburg University
Universitetskaya nab. 11, Saint-Petersburg, Russia
azic@bsr.spb.ru , asinopalnikova@yahoo.com, yav_mas@hotmail.com

Absract

The paper describes the on-going work on creating the WordNet-type lexicon for Russian, so called RussNet. The project started 3 years ago; preliminary results will be available at www.phil.pu.ru. The existing database contains verbs, nouns, and adjectives, the number of senses amounting to 2500.

The Top Ontology of RussNet is under construction, it will be co-ordinated with that of EuroWN. RussNet has inherited EuroWN language-internal relations. Several types of derivational links are added to describe Cross-Part-Of-Speech relations as well as Inner-Part-Of-Speech ones. Adjective-to-noun and verb-to-noun relations of words in collocations are described in details.

An overview of methods used for construction of the Russian WordNet is presented; the procedure of sense definition generation is also discussed.

1. RussNet Structure.

1.1. Vocabulary

For the RussNet structure we accepted the general approach, presenting only **Standard Russian** lexis, as opposed to various terminological subsets. The position doesn't prevent us from including those terms that were incorporated into the common language.

On the one hand this approach follows Russian lexicography tradition and on the other hand allows us to provide first and foremost **frequently-used current** vocabulary, that will be exploited by the majority of users. The main sources for such words are newspaper and magazine articles.

1.2. Inherited Features in RussNet

- RussNet is structured along the same lines as Princeton WN, EWN (Vossen, 1998, Miller et al, 1993) and other wordnets: words are grouped into synonym sets (**synsets**), each representing one underlying concept.
- Synsets in their turn are linked by means of various **Language Internal Relations** (LIR), such as hyponymy/hyperonymy, antonymy, meronymy/holonymy, entailment, causation, etc., hyponymy/hyperonymy being the most important one.
- RussNet consists of **4 interrelated files** for basic POS: nouns, verbs, adjectives and adverbs. So far we dealt only with 3 of them, but later we are going to add adverbs as well.
- Each of the 4 files contains a number of hyperonymy trees, with concepts of top levels constituting so called **Top Ontology**.
- Now, we are elaborating mainly internal structure of Russian wordnet and are not dealing with **Inter-Lingual-Index** (ILI).

2. Synset Formation

There are two different ways to define synonymy:

- in terms of substitution
- in terms of semantic similarity.

Although in EWN the weaker notion of synonymy is adopted: «two words are synonyms if there is a statement (class of statements) in which they can be interchanged without affecting truth value», we have to combine substitution method with that of semantic similarity. The reason for such a decision is as follows: in Russian there are many words which are not interchangeable in a context because of the syntactic, stylistic, expressive differences, but they are considered by native speakers as having similar meanings, denoting the same objects, entities, etc., e.g. aspect opposition for verbs.

There are two types of synonymy dictionaries for Russian:

- New Explanatory Dictionary of Russian Synonyms (Apresjan et al) is following the substitution strategy. The first issue of this dictionary was published in 1999, but so far it includes 132 entries only.
- Dictionary of Russian Synonyms (Evgenjeva,1970) & Explanatory Dictionary of Russian Verbs (Babenko, 1999) are based on semantic similarity.

Unfortunately, conventional Russian lexical resources may be used only partially because they don't cover all the lexis, the words definitions provided are made according to inconsistent patterns, and they may even obscure real semantic relations between words. That's why we can't simply import the data from those resources into RussNet without correcting it by means of our own lexical research procedures.

We begin with the collection of word senses for particular semantic groups of Russian words such as emotional verbs, nouns denoting the social relations and so on. The words realising the hyperlexeme sense were picked out from the sample of fiction or newspaper texts. A mean sample size ranges from 200 to 400 thousand word occurrences, from which about 150 core words and 70 peripheral words with appropriate senses were usually chosen. Having examined the synonymic relation in such groups we saw that words with the most abstract sense were encountered with relatively higher frequency and they would have synonymic equivalents. The hyponyms of the group were rare and may have derivational synonyms, but quite a few synonyms with different roots. So the collected words may be considered to be dominant representatives for respective synsets. Afterwards, extending the sample size or using synonymic information given in a conventional dictionary, we may expand synonymic sets with extra members.

3. Problems and discussion

3.1. Derivation

The Russian vocabulary, in particular verbs and nouns, is characterised by the high degree of derivation motivation. For example, dealing with verbs of thinking in Russian and English, we can see that there is about dozen of verbs with different roots in English (*to think, to contemplate, to consider, to regard, to reflect, to muse, to ponder, to cogitate, to meditate, , to conceive, to imagine, to picture etc*), and only 3 such items in Russian (*думать, мыслить, мозговать*), with a number of affixed derivatives amounting to 30 resultant verbs. Thus the total number of lexemes in Russian may be twice as much as that in English, while the situation with roots may be quite the opposite (Mitrofanova, 1999). From the point of view of frequency this causes specific distribution of lexical items in texts: it is rather flat in English in comparison with Russian sharp peak of frequencies for a hyperonym of this group *думать (think)* see Table 1, Table 2.

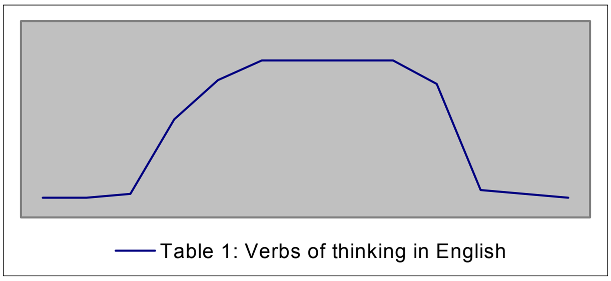
Table 1: Verbs of thinking in English

In many cases semantic relations between stem word and its derivatives couldn't be treated in terms of EWN Language Internal Relations (Vossen, 1998). They are more complicated: the main difficulty is that they are relations **between lexical items**, not synsets. Other reasons why we have to introduce new links are as follows:

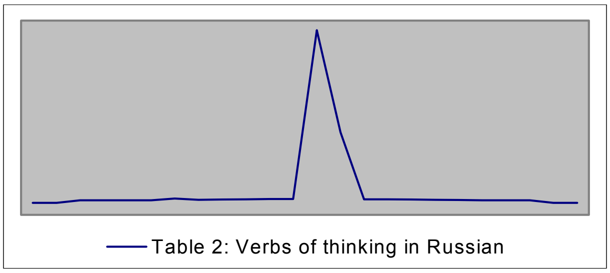
Table 2: Verbs of thinking in Russian

- There are many almost **unlimited** derivational chains: verb denoting process => noun denoting the process => attribute denoting the relevance to the process => adverb denoting the changing quality and so on, e.g. *удивлять (to astonish, to surprise) - удивление (astonishment) - удивленный (surprised) - удивленно (surprisingly).*
- The important traits of these chains are, that derivatives may be used freely in **paraphrases**: the motivating item may substitute the motivated ones in syntactic transformations. For example, a Russian noun *проверка* (*a check*) is paraphrased as a denotation of the process expressed by Russian verbs *проверить, проверять, провериться, проверяться (to check, to be checked)*. These links may be useful for syntactic analysis.
- **Lexical meaning** of derivatives is determined by that of the stem word.
- We would like to stress that verbal nouns inherit also the **syntactical** features of the motivating words. So if we describe the complex system of verb valences, they would be reproduced with little (and well known) changes by nouns denoting the same action or quality, on the one hand, and participants of action, on the other hand.

In those cases when it is possible we regard derivational relations in terms of LIR:

- **SYNONYMY -** relations between words which have the same root and different sets of affixes. They are not expressive and their senses differ so slightly that not every native speaker (researcher) is able to explain the distinction between them. Those words are also rarely interchangeable in the same context. *Семья – семейство (family), зло (malice) – злоба (malice, anger) – злость (malicious anger), бунтарь – бунтовщик (rebel, insurgent, mutineer, rioter), беда (misfortune, calamity) – бедствие (calamity, disaster).*
- **NEAR_SYNONYMY -** relations between
  - ➢ verb and abstract nouns, denoting processes of the same nature, e.g. *двигаться => движение (move => movement)*,
  - ➢ adjectives and abstract nouns, denoting characteristics and qualities, e.g. *красный => краснота (red => redness)*,
  - ➢ adjectives and nouns, e.g. гриб => грибной (fungus => relative to fungi)
  - ➢ verbs and adjectives, e.g. гнить => гнилой (rot => rotten).

In other cases we have to introduce a set of Derivational analogues of LIR, such as:

- **DERIVATIONAL_SYNONYMY** – relation between neutral words and their expressive derivatives. As those words differ from their stem word in style, they are not interchangeable in context, e. g. *старик (old man) => старикан, старикашка* (*impolite appeal to an old man), дом (house) – домик (house to which the speaker has positive emotions)*. Here we follow the idea, offered in Czech WordNet, of special attributes introduction. Thus *домик* will have X_EXPRESSES_ POSITIVE_EMOTION, while *старикашка* – X_EXPRESSES_IMPOLITE .
- **DERIVATIONAL_HYPONYMY** – verb-to-verb, noun-to-noun, adjective-to-adjective relations of following types. For verbs we may use
  - ➢ specific attributes X_HAS_INCHOATIVE or X_HAS_SPECIFIED_DURATION for actions restricted in time duration (inchoatives), e.g. *петь => запеть (to sing => to begin to sing), сидеть => посидеть (to sit => to sit for a while), сидеть => просидеть (to sit => to sit for a long time)*;
  - ➢ an attribute X_HAS_SPECIFIED_RECURRENCY for actions repeated only once or several times, e.g. *кричать => крикнуть, покрикивать (to shout*

=> to shout out once, to shout not aloud many times);

- ➢ an attribute X_HAS_SPECIFIED_NUMBER for actions, having many objects involved, e.g. *думать - раздумывать (to think - to ponder about many things for a long time), резать - вырезать (to cut - to cut out some part from many things)*, and so on.

These special verbal derivatives interacting in a complex manner with an aspect category of verbs and having semi-grammatical nature. We still don't know in which manner to treat them, on the one hand, aspect pairs look like very close synonyms, though on the other hand, they realise a very important semantic opposition, such as activity ⬄ action. We may introduce specific attributes, as follows: X_HAS_IMPERFECT, X_HAS_ PERFECT.

- ➢ For nouns and adjectives we may add attributes X_IS_SMALL and X_IS_BIG, and possibly several others, when the clear sense component is added by some affixes to the stem word meaning, and the resultant word couldn't be regarded as purely expressive variants; this why we should treat such pairs as *стол => столик (table => small table), дом => домишко (house => small house), пожар => пожарище (fire => big fire), громадный => громаднейший (huge => very huge)* as derivational hyperonym - hyponym.

We should note that the majority of these derivational variants doesn't belong to the core of Russian lexis because of their infrequency in texts. However, the highly inflected nature of Russian may turn any potential derivative into common and frequently used one, that's why all derivational regular models should be taken into account.

Moreover, we may find several cases when an expressive shade may disappear, then a word would change expressive synonym status for a synonym position. Another example of extending the sphere of usage for diminutives may be seen in the Russian spoken language (usually by women), when these words function as oral equivalents for their neutral motivating counterparts, so we may expect that in future they have a chance to become colourless synonyms.

Expressive synonyms and hyponyms may exist beyond the derivational scope, but in these cases they are rather few, irregular, and disputable, that's why it would be adequate to include them into the synset with a proper attribute.

- **DERIVATIONAL_ROLE_RELATIONS** are established to link a verb to its derivatives, designating action participants, such as ROLE_DERIVED_AGENT, ROLE_DERIVED_ OBJECT, ROLE_DERIVED_INSTRUMENT, ROLE_DERIVED_ LOCATION and so on, e.g. *сеять => сеятель, сеянец, сеялка (to sow => sower, seedling, seeding-machine*). The link in the opposite direction is a realisation of the semantic link INVOLVED_IN_ACTION. We are inclined to treat such cases as a specific derivational relation because the semantic link usually has wider scope, e.g. *принимать => приемник (receive => radio set = receiver)*, the object is involved in the first place into the situation *слушать (listen)*. This is usual for complex activity nomination, which as a rule is designated with regard to one action varying from one language to another, e.g. *шить => швея (to sew => seamstress)*. Above we have mentioned the inheritance of syntactic features, moreover, the collocation restrictions of stem verbs may be inherited by their derivatives.

3.2 Adjectives in RussNet

As there is no common solution for treatment of adjectives in EWN, we offer the following one.

We comply with the idea of GermaNet to make use of hyponymy relations wherever it is possible, but our German colleges determine hierarchical structure of adjectives according to semantic fields, while we regard adjectival hyperonymy in terms of their collocations with nouns. We received preliminary results which prove that on the level of adjectives grouping and nouns tree hyperlexeme, it is the **adjective** in Russian that **predicts** certain type of **nouns to collocate with** it, and not vise versa. For example, meaning of *долговязый (lanky)* involves the pointer to a human being, i.e. it can collocate with such nouns as *мальчик (a boy), человек (a man), папа (a father).*

We are prone to the opinion that **adjectival hyponymy trees** can be built according to their collocation with nouns from different levels of hyponymy tree. For example, let's take two adjectives, which express the similar semantic quality – denotation of *height*. In case when one adjective – *высокий (tall)*– may collocate with all nouns denoting “entity”: objects, animals, humans and so on, while the other – *рослый (well-grown, srapping)* – collocates only with a certain part of the tree – human beings, the first one may be thought as hyponym for the second one. So checking the cooccurrence of adjectives with nouns, we are to produce hyponymy structure for groups of adjectives denoting the similar quality.

3.3. Verb Valencies

It is generally accepted that syntactic features of words, especially verbs, are determined by their semantic properties, that the meaning of a verb outlines the form and semantic features of words accompanying it.

The semantic and syntactic structure of verb arguments is called the **valency frame**. Valencies may be thought in terms of morphological noun forms, which are obligatory or optional. This characteristic is vital for Russian syntax, as well as for that of other Slavonic languages (Pala, Sevecek, 1999).

Verbs may have different valency frames associated with dfferent meanings, cf.

- ➢ *Бить (посуду) {crash, break up,break apart}*
- ➢ *Бить (в барабан) {drum, beat, trum}*
- ➢ *Бить (врага) {repel, repulse, fight of, rebuff}*

The minimal form of valency description implies the noun case specification; often it is accompanied with the indication of number, gender, and preposition (a number of prepositions).

We fix the **semantic** features of nouns as well, which a verb can take as arguments in a sentence. It means we use top-level concepts, deal with **classes** of words, rather than with separate words, including verb-to-class relations in the synsets. In the example above, the argument of a verb in the first frame is a fragile object, in the second – musical instrument, more precisely – percussion instrument, in the third – human being, military unit and so on. These references to the hyponymic tree structure of nouns would be very helpful for syntactic description as well, though sometimes this relation may be very complicated.

The situation with valency frames is not clear due to versatility of syntactic preferences of verbs included into a synset, while sometimes they behave uniformly. We use **a list of valency frames** for a synset, specifying which frame fits the member of a synset. The set of frames is better than separate verb description, because in this case the paradigm influencing the native speaker is presented.

Moreover, it would be very useful to represent the inheritance of syntactic frames of a hyperonym by its hyponyms, e. g. *двигаться (to move) ==> идти (to walk):* hyperonym *двигаться* has valency frames: (a) “starting point – location”, (b) “destination point – location”, which are inherited by its hyponym *идти.*

4. Definition Generation

4.1. Subset Sense Definition

We still don't speak about definition generation procedure, but it's vital to have in mind guidelines for definition formulation because dictionary ones for a long time have been a target for an extensive criticism. In this respect we propose several key notes.

4.1.1. “Genus proximum + Differentia specificae” Definition

The definition of a synset incorporated into the hyponymic (or troponymic) tree should be constructed on the following pattern "the dominant **lexeme** of the **hyper** level **plus** a **distinguishing part** showing difference between co-hyponyms", e.g. *плыть (to swim)* has hyperlexeme: «to move in certain direction» + differentiation: «on the surface or in depth of water using special organs», *лететь (to fly)* has hyperlexeme: «to move in certain direction» + differentiation: «in the air using wings». In this case there is no Russian hyperlexeme denoting *moving in some direction*, though it's important to oppose this way of moving to the other one in various direction, with repetitions, to and fro.

It's clear that in case of a large number of co-hyponyms the problem may become practically insolvable because of a great number of necessary differential features, then it would be better to use other types of defining or artificial names (used in GermaNet) uniting several lexemes into a cluster.

4.1.2. Meronymic Definition

The definition of a synset incorporated into the meronymic relations may be based on either holonym, or meronym.

In the first case, a holonym is the referential part of the definition (similar to hyperlexeme), but a simple indication that something is a part of the holonym is not sufficient, so it is usually supplied with a special function (for artefacts) or construction peculiarties. For example, structure «part + construction characteristic + holonym + function» may be used: *крыша (roof)* = «the upper part of the building, covering it from precipitation».

In the second case, a limited number of meronyms may be used for generation of list-type definition, e.g. *фигура (chessman)*: «king, queen, castle, knight, bishop in chess opposed to pawns».

4.1.3. Derivational Definition

In those cases when a synset is associated with a purely derivational link we use a definition describing the additional sense of the derivational affixes, e.g. *столик* «a small table», *генеральша* «general's wife».

4.1.4. Semantic Pointer Definition

The simplest way of defining the quality is to show the synonyms expressing it, which are united in the synset, so in this case we have a rudimentary definition equal to an ordinary synset. This type of definition is frequent for adjectives and adverbs.

Antonymic definition is adequate in those cases when one member of the antonymic pair is marked showing the positive content while the other shows its absence, e.g. *глупый (foolish)* «not clever» <=> *умный (clever)* «having the intellect».

Causative definition is alike the derivational one so as it makes implicit the causative copula and the final state of transition, in Russian there is a specific affix with anti-causative meaning, e.g. *поднять (raise)*: *каузировать подняться (cause to rise)*. Usually in such a definition the artificial causative is used, which is the transliteration of English *cause*, because a Russian equivalent *заставить* means ‘to enforce’, that is not neutral at all.

Moreover, using semantic attributes, such as X_HAS_IMPERFECT, X_HAS_PERFECT, X_IS_SMALL, X_IS_BIG etc., incorporated into the WordNet structure, we may later elaborate a procedure for automatic definition generation.

4. Conclusions

To sum up we may say that RussNet presently covers the core of the Russian lexis (the resulting number of synsets is more than 2500). So it can be regarded as a reliable starting point for further extending and elaboration of the system, which will be carried out by addition of peripheral groups of words, emotionally coloured lexis and derivatives, in particular. This should enrich the content of the database. The introduction of new relations allows us to perform more adequate semantic analysis of the Russian language.

5. References

Apresjan, U. (ed.) (1997). Новый объяснительный словарь синонимов русского языка (=New Explanatory Dictionary of Russian Synonyms). Moscow.

Babenko, L. (ed.) (1999).Толковый словарь русских глаголов (=Explanatory Dictionary of Russian Verbs). Moscow.

Evgenjeva, A. (ed.) (1970). Словарь синонимов русского языка. (=Dictionary of Russian Synonyms) (vol 1-2). Leningrad.

Miller, G. et al (1993). Five Papers on WordNet. Technical Report, Cognitive Science Laboratory, Princeton University. ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps

Mitrofanova, O. (1999). Структурный анализ сигнификативного значения: на материале глаголов процесса мышления английского и русского языков (Structural Analysis of Sense: Verbs of Knowing in English and Russian). PhD thesis. St-Petersburg State University, Philological Faculty, Department of Applied, Structural and Mathematical Linguistics.

Naumann, K. (2000). Adjectives in GermaNet. http://www.sfs.nphil.uni-tuebingen.de/Adj.html

Ozhegov, S., Shvedova, N.(1992). Толковый словарь русского языка (=Explanatory Dictionary of Russian). Moscow.

Pala, K., Sevecek, P. (1999). The Czech WordNet, EuroWordNet (LE-8928). Deliverable 2D014. http://www.hum.uva.nl./~ewn/docs.html

Vossen, P. (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht: Kluwer.

Словарь современного литературного русского языка.(1991). (=Dictionary of Modern Literary Russian) (vol. 1-17). Moscow-Leningrad.