Adjectives in RussNet

Irina Azarova[1] and Anna Sinopalnikova[1,2]

[1] Saint-Petersburg State University, Russia
Email: azic@bsr.spb.ru
[2] Masaryk University, Brno, Czech Republic
Email: anna@fi.muni.cz

Abstract. This paper deals with the problem of structuring adjectives in a wordnet. We will present several methods of dealing with this problem based on the usage of different language resources: frequency lists, text corpora, word association norms, and explanatory dictionaries. The work has been developed within the framework of the RussNet project aiming at building a wordnet for Russian. Three types of relations between descriptive adjectives are to be discussed in detail, and a technique for combining data from various resources to be introduced.

1 Introduction

Up to date presenting adjectives within a wordnet remains one of the most difficult and disputable matters of the lexical semantics.

Although there is no common solution for structuring adjectives in wordnets, some general considerations are adopted by most of the researchers. **Firstly**, it is generally accepted that being a 'satellite' words, adjectives posses very specific meaning (vague, highly dependent on the meaning of accompanying nouns). It is usually stressed that adjectives, descriptive ones, in particular, have no denotation scope of their own. **Secondly**, due to their specific semantic and syntactic properties, semantic organization of adjectives is entirely different from that of other open classes of words. Thus, **thirdly**, methods of revealing the semantic organization for nouns and verbs do not hold for the adjectives [1,2,3].

Adopting these statements as a base of our research, we are to describe the ways semantic organisation of Russian descriptive adjectives is examined. Although the facts discovered could not be expanded on all other languages, the methodology applied is of a scientific value and may contribute significantly to the standards of wordnet building.

2 Frequency List Study

Usually a wordnet building process starts with the analysis of most frequent words (extracted either from corpora [4], or explanatory dictionaries [4,5]) in order to obtain the list of general concepts representing the core structure of a language, so-called Base Concepts.

In addition to its main task performing, the frequency analysis yields many subsidiary results that are useful for the next stages of wordnet constructing. As far as frequency lists of Russian [6,7] concern, it appears that among more than 6500 adjectives given descriptive ones occupy most positions, including the 76% of the 50 top positions.

The following conclusions could be made:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 251–258.

Table 1. Top frequent Russian adjectives in a large corpus (according to [7]).

Rank	Word	Eng	Ipm	Rank	Word	Eng	Ipm
62	большой	big, large	1630.96	150	последний	last	630.17
114	хороший	good	853.71	180	старый	old	528.25
116	новый	new	840.18	194	белый	white	493.36
128	конечный	final, last	732.33	203	главный	main	467.77
137	нужный	necessary	690.34	224	маленький	small	411.52

1. The fact discovered confirms the general view of descriptive adjectives as the 'most typical' representatives of this PoS.
2. High frequency of a certain adjective doesn't indicate whether it is caused by its numerous senses or by its preferential status, or by both simultaneously.
3. The adjective's frequency reveals which member of an antonym pair is marked, being more common. The detailed corpora analysis, e. g. usual position of some adjective after the negative particle *не* ('not'), allows us to define precisely which antonym is semantically marked. The positive value of some parameter is usually supposed to be prone to a markedness, e. g. an opposition between 'big' (*большой*) and 'small' (*маленький, малый*). The information of an antonym's 'markedness' is to be used while generating appropriate definitions for adjectives (see the last section).
4. Frequency data helps us to set order into the synsets, to establish the priority of synonyms from the viewpoint of their usage. Being a neutral term, dominant synonym is expected to occur in texts more often then other members of the corresponding synset.
5. Frequency data allow us to verify the hypothesis of the correlation between two modes of synset organization: from the most frequent synonym to less frequent ones, and from a neutral dominant synonym to expressive and terminological ones.

3 Distinguishing word senses

According to the data shown in Table 1, adjective *большой* ('big/large') is the most frequently used Russian adjective. The fact calls for an explanation, regarding that *большой* usually considered to denote so-called visual assessment of size, which is narrower than that of the adjective *хороший* ('good'), ordinarily said to indicate a general assessment of an object, event, or quality. This situation may be accounted for either by high ambiguity of the adjective *большой*, or by the more abstract nature of this adjective.

To specify and to distinguish between word senses of *большой*, we apply 2 language resources: text corpus[3], and association tests[4]. Extracting from both resources data on syntagmatic properties of the adjective, e. g. selectional restrictions, we base our case study on the general consideration: "Every distinction in a meaning is reflected by distinctions

[3] A balanced corpus of Russian texts for the study includes about 16 mln words. Texts belonging to different functional styles were taken in the following proportions: fiction –20%, newspapers and magazines – 40%, popular science texts – 30%, laws – 10%. The time boundaries are defined as 1985–2003.

[4] RWAT – The Russian Word Association Thesaurus by Karaulov et al. [8] and RWAN – Russian Word Association Norms by Leontiev et al. [9] were used.

in form" separately made by many of the linguists working in the area of corpus-based lexicography [10,11].

In our research we focus mainly on the lexical and semantic context markers, and partly domain ones. The analysis of noun collocations with the adjective *большой* is to assist to reach a decision regarding the number of word senses, which should be distinguished in the RussNet.

From RWAT we extract noun-responses of *большой* combining freely with the adjective in question (ignoring idioms like *Большой театр, большой палец*). Noun-responses may be organized into several groups:

(1) spatial artefacts (*house, town, shop,* etc.);
(2) three-dimensional natural objects (*forest, ball, mushroom*, etc.);
(3) animals (*bear, elephant*, etc.);
(4) two-dimensional objects (*sheet, circle*);
(5) persons (*man, boy, son*);
(6) personal characteristics (*friend, fool, coward*, etc.);
(7) parts of human body (*nose, mouth*);
(8) abstract nouns (*brain, experience, talent*, etc.).

By summing up associations in groups (including unique ones) we distinguish those three, which are the most numerically strong: 1, 6, 8. Checking these data across the corpus, we receive the same leading groups of nouns, the top frequent collocants of *большой* being: *money* (127), *man* (39), *eyes* (36), *problem* (22), *opportunity* (21), *hope* (20), *group* (18), *town* (13), *loss* (13), *difficulty* (12), *distance* (11), etc.

Thus, on the base of facts discovered we may draw a conclusion that the most frequent sense of the adjective *большой* (according to the corpus and RWAT data) is the 'indication to the above-average spatial characteristics of an object'. That holds for both natural objects (including animals) and artefacts, the last including objects with absolute above-average size, e.g. *дворец* 'palace', *город* 'city', *слон* 'elephant', *самолет* 'aeroplane', as well as with relative one, e.g. *капля крови* 'blood driblet', *прыщ* 'smirch', *гриб* 'mushroom', etc. It is in this particular sense *{большой*$_1$*}* is related to its augmentative hyponym *{огромный*$_1$, *громадный*$_1$*}* 'very big' and antonym *{маленький*$_1$, *малый*$_1$*}* 'of a minor, less than average size'.

First sense covers its usage with noun-groups (1), (2), (3), (4), (7). Other senses manifested are (ordered by frequency):

- With nouns from group (8) *большой*$_2$ signalizes 'above-average level of quantifying features [intensity, duration, importance] of some event or state', e. g. *большая проблема, большие сложности.*
- With nouns from group (6) *большой*$_3$ is used for indicating to 'high intensity of some human's trait' mentioned by a noun, e.g. *большой друг.*
- With several nouns from group (5) pointing to children *большой*$_4$ refers to 'grown up from infancy', e.g. *большой мальчик.*

4 Establishing Relations

As we have shown in the previous section, both the RWAT and our corpus supply us with the evidences on the syntagmatic relations of the adjectives. But they also allow us to observe their paradigmatic relations as well.

Regarding the frequency of words from the same PoS (probably, paradigmatically related to adjectives under consideration), we may conclude that paradigmatic relations are highly relevant for adjectives: *большой –> маленький 47, огромный 15, малый 12, толстый 6, высокий, длинный, крупный 3,* etc. (the total amount of associations in RWAT counting 536); and *большой – маленький* 98 (MI = 6.072), *малый* 69 (MI = 7.728), *крупный* 15 (MI = 4.095), *мелкий* 15 (MI = 4.817) etc. out of total amount of 9762 lines in the corpus.

1. These lists of co-occurring words give us a hint on what adjectives could belong to the same semantic field, or to the same hyponymy tree. Thus, for example, we may conclude that *маленький, огромный, малый, толстый, высокий, длинный,* etc. probably belong to the same semantic field as *большой.*
2. Comparing the context patterns (see Section 3) for these adjectives, we are able to establish links between them and to organize them into tree structures.

The general approach to this task performance suppose the fulfilment of following conditions:

- To establish a **Hyponymy** link we need the evidences in favour of context inclusion, see Section 4.1.
- **Antonymy** relations are often characterised by the identical contexts. Antonymous adjectives also may co-occur in contrastive sentences (‘and/or/but’), e.g. *большие и малые программы, нажимать большие или маленькие кнопки* or *план большой, а зарплата маленькая.* See Section 4.2.
- For **synonymous adjectives** identity of contexts is believed to be quite a rare phenomenon, rather we observe incompatible contexts (complementary distribution), e.g. *незамужняя женщина* and *неженатый мужчина.* As an additional criterion we may rely upon co-occurrence of synonyms in enumerating phrases (e.g. *большой, крупный нос*). See Section 4.3.

4.1 Adjectives and Hyponymy

Following the GermaNet proposal to “make use of hyponymy relations wherever it’s possible” [12], in RussNet we adopt formal approach based on the adjective collocations with nouns. Empirical data proves that in Russian it’s the adjective that predicts the noun (class of nouns) to collocate with, not vice versa, e. g. *долговязый (ланкы, страппинг)* involves the pointer to a human being, i. e. it can collocate with such nouns as *мальчик* (*a boy*), *человек* (*a man*).

Thus, the main idea underlying our work is that **hyponymy tree** for descriptive adjectives may be built according to that of nouns: i. e. if 2 adjectives from the same semantic field collocate with 2 nouns linked by the hyponymy, we are to build the hyponymy link for these adjectives [13].

We consider the procedure for retrieving the information about hyponyms using the above mentioned adjective *большой*. There are several multiple adjective responses in the RWAT: *огромный* 'huge', *толстый* 'thick', *круглый* 'round', *высокий* 'high', *длинный* 'long', *крупный* 'large-scale', *сильный* 'strong', *красивый* 'nice', *необъятный* 'immense'. The next step is to specify weather these responses are syntagmatic or paradigmatic. For that purpose we apply to the corpus-driven data on adjective co-occurrences. It appears, that some adjectives do collocate with *большой* in our corpus, e.g. *толстый* 'thick' and *круглый* 'round', however, *красивый* 'nice' occurs 4 times with rather high MI-score (8.063). Also syntagmatic relations are manifested by associations with a copulative conjunction *и* 'and' in RWAT, e.g. *и красивый, и круглый*. Thus, we could exclude adjectives *красивый* and *круглый* from paradigmatic associations, consider *огромный*, *высокий*, *длинный*, *крупный*, *сильный*, *необъятный* to be paradigmatic, and *толстый* – ambivalent.

Lists of word associations for *высокий*, *длинный*, *сильный* look nearly-identical: their leading responses are nouns (*путь* 55; *человек* 54), and antonymous adjectives (*низкий* 48; *короткий* 54; *слабый* 42), while for *огромный*, *крупный* and *необъятный* the leading responses compose *большой* and nouns. The former fact may evidence in favour of a hyponymy link, the latter one may count for synonymy or hyponymy. An ambivalent adjective *толстый* has a structure of the first type.

4.2 Adjectives and Antonymy

Although in Princeton WN antonymy is regarded as a relation between words rather than synsets, in RussNet antonymy is considered to be one of the semantic relations between **synsets**.

Yet we by no means are to reject the differentiation of **direct and indirect antonymy**. We suppose that setting order into a synset helps us to manage this problem adequately. As RWAT shows, in Russian it is usually synset representatives ('dominant literals') that are related by antonymy directly, all other members of synsets are opposed through this pair, i.e. indirectly. E. g. *большой* is strongly associated with *маленький, маленький* is associated with *большой*, while *малый* is associated first of all with *маленький*, its association with *большой* is rather weak. But there still is a possibility that several pairs of direct antonyms may appear in the frame of two synsets, like in English *ларге ↔ смалл, биг ↔ литтле.* However, our study of 533 most frequently used descriptive adjectives (on the basis of RWAT) proves this phenomenon is not that characteristic for Russian.

4.3 Adjectives and Synonymy

In its first and second senses *большой* is a dominant of synsets. As syntagmatic data driven from RWAT and the corpus show, these synsets may include an adjective *крупный* as well. **Firstly**, this adjective occurs regularly as a response to *большой* in the RWAT, it belongs to the 10 most frequent ones. Also regarding backward associations, we discover that *большой* is the first and hence, the most strong, response to *крупный.* The same observation holds for *огромный* and *громадный,* but as opposed to *крупный* both this adjectives fail the implicative synonymy test. E.g. *Большая сумма денег* ⇔ *Крупная сумма денег,* but

Огромная сумма денег ⇒ *Большая сумма денег*, and not vice versa. **Secondly**, comparing syntagmatic associations of *большой* and *крупный*, we observe a significant overlap of the lists. Some responses ($\sim$21%) literally coincide, e. g. *человек, город, нос, выигрыш, успех, специалист*, many others are semantically similar (i. e. belong to the same semantic field) e.g. *разговор, план*, etc. So do the micro-contexts patterns for these adjectives. **Thirdly**, more detailed study of the corpus proves that *крупный* is used mainly in specific domains: commerce and finance texts, e.g. *крупный бизнес, крупный московский автоторговец, крупный производственный филиал, крупный "рынок"* и т. д. Thus, it is clear, that in the corpus the adjective *крупный* occurs far less frequent than *большой* (3882 lines against 19566). **Fourthly**, in most of the observed contexts *крупный* may be easily substituted by *большой*. **Fifthly**, analysis of definitions from Russian explanatory dictionaries [14,15] shows the significant overlap in structure of several definitions given to *крупный* and *большой*.

As a side result of the analysis we also observe that the first sense given in the dictionaries for *крупный* 'consisting of large particles or objects of above-average size' (*крупный песок, жемчуг*) includes an indication to an aggregate or collection of identical or similar units, that could not belong to the same semantic field as *большой*$_1$. This is confirmed by the substitution test: *крупный песок*, but **большой песок*. The priority of that sense is not supported by the actual data: in RWAT nouns illustrating this sense of *крупный* (*дождь, снег, град, виноград, корм, порошок, шрифт, слезы*) are obviously peripheral – their absolute frequency never exceeds 5, and their number gives only 2.7% of total amount of responses. Frequency data counts against the actual priority of the historically original 'aggregate' sense: *крупный* is used less frequent in this sense, so it should be treated within a wordnet as a secondary (*крупный*$_3$).

All the facts discovered – similar meanings, substitutability, similarity of responses in RWAT and contexts in the corpus, domain markedness of *крупный* and neutrality of *большой* – enable us to conclude that the adjective *крупный* belongs to the same synsets as *большой*$_1$ and *большой*$_2$. According to the data on usage, the synsets should be ordered as follows: {*большой*$_1$, *крупный*$_1$}; {*большой*$_2$, *крупный*$_2$}.

5 Generating Appropriate Definitions

As for the adequate representation of systemic relations of adjectives, definitions given in conventional dictionaries are considered to be inconsistent and insufficient. The possible explanation for that lies in the difficulty of performing this task within the framework of traditional lexicography. Specific semantic features of adjectives, such as their mainly significative meaning and absence of clear denotation, dependence on the modified nouns etc. make the traditional methods quite an unreliable base for definition generation. In order to construct appropriate definitions for adjectives we rely upon their **relations** to each other and to nouns they co-occur with.

The relevance of relations may be rated from the viewpoint of the definition generation:

1. For descriptive adjectives **antonymy** is by no means one of the most important and rich in content relations [16,17,18]. Semantic markedness of opposition members determines the direction of the definition generation. Unmarked member is to be defined through

the marked one (e.g. *истинный* through *ложный*). Their definitions in Princeton WN are reversed: *true* – ‘consistent with fact or reality; not false’, *false* – ‘not in accordance with the fact or reality or actuality’. In case of definition based on the antonymy relation special attention should be paid to cycles, when antonyms are defined through each other.

2. **Hyponymy** seem to be useful for definition construction in cases of augmentative/diminutive hyponyms. For most descriptive adjectives denote various assessments of gradable properties, intensity or mildness is among the most frequent components of their meanings. E.g. *невысокий* – ‘not very low’.

The semantic structure of adjectives is considered to be dependent on and specified by the nouns they modify [1]. Thus another necessary contribution to definition generation concerns the coding of meanings of nouns, adjectives co-occur with. The relations within noun–adjective collocations may be divided into several types: goal-instrument e.g. *athletic equipment*, result-cause e.g. *healthy air*, feature-whole *big house*, etc. [3]. Each type of relations requires a specific model of definition (specification of how and to what extent meaning of a co-occurring noun modify an adjective’s meaning): *healthy*$_3$ *– promoting health* e.g. *healthy air.*

6 Conclusions and Future Work

Diverse language resources – frequency lists, association norms, corpus analysis – affords us to establish a clear-cut adjective structure in the RussNet (a wordnet for Russian) [19]. The described technique aims at listing different senses of an adjective, differentiating synonymy and hyponymy links, defining antonym pairs, generating proper sense definition explaining the difference between co-hyponyms.

It is important now to apply it consistently to the whole stock of the descriptive adjectives in RussNet, verifying and correcting the method. Using it on the large scale may find difficulties due to the absence of association data, or an insufficient number of occurrences in the corpus for less frequent adjectives.

References

1. Gross D., Fellbaum C., Miller K.: Adjectives in WordNet. International Journal of Lexicography 3 (4) (1990) `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`.
2. Apresjan Ju. D.: Lexical semantics. Vol. 1–2. Moscow (1995).
3. Willners C.: Antonyms in context: A corpus-based semantic analysis of Swedish descriptive adjectives. PhD thesis. Lund University Press (2001) `http://www.ling.lu.se/education/gradstud/disputation/Omslag.doc`.
4. Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht, Kluwer (1998).
5. Pala K., Ševeček P.: The Czech WordNet, EuroWordNet (LE-8928). Deliverable 2D014 (1999) `http://www.hum.uva.nl./~ewn/docs.html`.
6. Zasorina L. N. (ed.): Frequency Dictionary of Russian. Moscow (1977) (40.000 entries).
7. Sharoff S. A.: Frequency List of Russian (2000) (35.000 entries) URL: `www.artint.ru/projects/freqlist`.
8. Karaulov Ju. N. et al.: Russian Associative Thesaurus. Moscow (1994, 1996, 1998).

9. Leontiev A. A. (ed.) Norms of Russian Word Associations Moscow (1977) (about 100 entries).
10. Sinclair, J.: Corpus, concordance, collocation. Oxford: Oxford University Press (1991)
11. Apresjan Ju. D.: Systematic lexicography / translated by K. Windle. Oxford University Press. (2000).
12. Naumann K. Adjectives in GermaNet. (2000) `http://www.sfs.uni-tuebingen.de/lsd/`.
13. Azarova I. et al.: RussNet: Building a Lexical Database for the Russian Language. In: Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas (2002) 60–64.
14. Evgenjeva A. P. (ed.): Dictionary of Russian (vol. 1–4). Moscow (1985–88).
15. Ozhegov S. I., Shvedova N. I.: Explanatory Dictionary of Russian. Moscow (1992).
16. Charles W. G., Miller G. A.: Contexts of Antonymous Adjectives. Applied Psycholinguistics 10 (1989) 355–375.
17. Fellbaum C.: Co-occurrence and antonymy. International Journal of Lexicography 8(4) 281–303.
18. Justeson J. S., Katz S. M.: Co-occurrence of Antonymous Adjectives and Their Contexts. Computational Linguistics 17 (1991) 1–19.
19. RussNet: Wordnet for Russian: URL: `http://www.phil.pu.ru/depts/12/RN/`.