A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects[1]

Juri Apresjan[1,2], Igor Boguslavsky[1,3], Boris Iomdin[2], Leonid Iomdin[1], Andrei Sannikov[2], Victor Sizov[1]

[1]Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow
[2]Vinogradov Institute of the Russian Language, Russian Academy of Sciences, Moscow
[3]Universidad Politécnica de Madrid

[1]Bolshoj Karetnyj per. 19, Moscow, Russia; [2]Volkhonka 18/2, Moscow, Russia; [3]Boadilla del Monte, 28660 Madrid, Spain
apr@iitp.ru, igor@opera.dia.fi.upm.es, iomdin@ruslang.ru, iomdin@iitp.ru, sannikov@ruslang.ru, sizov@iitp.ru

Abstract

We describe a project aimed at creating a deeply annotated corpus of Russian texts. The annotation consists of comprehensive morphological marking, syntactic tagging in the form of a complete dependency tree, and semantic tagging within a restricted semantic dictionary. Syntactic tagging is using about 80 dependency relations. The syntactically annotated corpus counts more than 28,000 sentences and makes an autonomous part of the Russian National Corpus (www.ruscorpora.ru). Semantic tagging is based on an inventory of semantic features (descriptors) and a dictionary comprising about 3,000 entries, with a set of tags assigned to each lexeme and its argument slots. The set of descriptors assigned to words has been designed in such a way as to construct a linguistically relevant classification for the whole Russian vocabulary. This classification serves for discovering laws according to which the elements of various lexical and semantic classes interact in the texts. The inventory of semantic descriptors consists of two parts, object descriptors (about 90 items in total) and predicate descriptors (about a hundred). A set of semantic roles is thoroughly elaborated and contains about 50 roles.

1. Syntactic Tagging

The paper is a progress report on a project aimed at creating a deeply annotated corpus of Russian texts. This corpus, jointly developed by two Moscow teams, is largely based on the ideology of an advanced MT system, ETAP-3 (Apresjan et al. 2003), and is so far the only corpus of Russian supplied with comprehensive morphological annotation and syntactic tagging in the form of a complete dependency tree provided for every sentence.

Fig. 1 is a screenshot of the dependency tree for the sentence

(1) *Наибольшее возмущение участников митинга вызвал продолжающийся рост цен на бензин, устанавливаемых нефтяными компаниями* 'It was the continuing growth of petrol prices set by oil companies that caused the greatest indignation of the participants of the meeting'.

Fig.1. A syntactically tagged sentence

Here, nodes represent words assigned morphological and part-of-speech tags, whilst branches are labeled with names of syntactic links. The tagging uses about 80 surface-syntactic links; half of these were proposed in Mel'čuk's Meaning ⇔ Text Theory (see e.g. Mel'čuk 1988) and the rest were adopted from the ETAP-3 system or specifically designed for the project. Annotation is produced semi-automatically: sentences are first processed by the rule-based Russian parser of ETAP-3 and then edited manually by linguists who handle all hard cases, including the cases of ambiguity that cannot be reliably resolved without extralinguistic knowledge, as well as versatile elliptical constructions, syntactic idiomaticity, and the like.

Currently, the syntactically tagged corpus exceeds 28,000 sentences belonging to modern Russian texts of a variety of genres (fiction, popular science, newspaper and journal articles etc.) and is steadily growing. It is an integral but fully autonomous part of the Russian National Corpus developed in a nationwide research project and available on the Web (www.ruscorpora.ru).

2. Semantic Tagging

Recently (Apresjan et al. 2004a), the annotators proposed to enhance the depth of the tagged corpus by adding innovative semantic tags to sentence representations. For this purpose, we developed an inventory of semantic features (descriptors) and a dictionary comprising about 3,000 entries, with a set of tags assigned to each lexeme, and are elaborating tools for

[1] The paper was partially supported by a grant No. 04-07-90179 from the Russian Foundation of Basic Research, which is gratefully acknowledged. In addition to the authors' of the paper, Valentina Apresjan, Olga Boguslavskaya, Tatyana Krylova, Irina Levontina and Elena Uryson have contributed to the creation of the semantic dictionary and the system of descriptors.

handling semantic data in diverse types of linguistic research.

For semantic descriptors, words of natural language (in our case, Russian) are used whenever possible; e.g. *действие* ('action') or *деятельность* ('activity'). In certain cases, linguistic terms like *каузация существования* ('causation of existence') are used.

The set of descriptors assigned to words has been designed in such a way as to construct a linguistically relevant classification for the whole Russian vocabulary and to provide the researchers with comprehensive information about the laws according to which the elements of various lexical and semantic classes interact in the texts.

The inventory of semantic descriptors consists of two parts, object descriptors and predicate descriptors, in accordance with the idea that all words of Russian (and probably any other language) can be of two types: objects (names of animals, birds, fish, fruits, vegetables, stones, mountains, stars, planets, etc.) or predicates (lexical units that have at least one semantic valency). Both parts of the inventory are further subdivided into two subgroups, the generic and the specific semantic features. For generic descriptors (*genus proximum*), nouns are used ('animal', 'vegetable', 'state', 'action', etc), whereas specific descriptors (*differentia specifica*) are adjectives (e.g. 'domestic', 'wild', 'natural', 'physical', 'mental').

Two different classifications are used for the object and the predicate parts of the vocabulary: a taxonomic classification and a fundamental classification, respectively.

Object descriptors reflect the "naïve" perception of the world, rather than a scientific account thereof. This is why the noun *паук* 'spider' is assigned the feature 'insect' and not 'arachnoid'. Currently, ca. 90 object descriptors are used. New descriptors may be added as the semantic dictionary is expanded and additional words are tagged.

The set of predicate descriptors consists of two subsets – predicate descriptors proper (about a hundred) and roles for tagging the semantic valencies of predicate lexemes (over 50 roles). Both have grown out of independent research in the domain of systemic lexicography based on the idea of integrated linguistic descriptions (see Apresjan 2000, 2003). This approach has been partly implemented in dictionaries, above all in the *New Explanatory Dictionary of Russian Synonyms* whose second, updated and enlarged edition came out of print in 2004 (Apresjan et al. 2004b). The system of predicate descriptors may thus be claimed to have received substantial linguistic validation.

This system is based on the version of fundamental predicate classification developed by Juri Apresjan and differs from comparable systems[2] of Juri Maslov, Zeno Vendler, Tatiana Bulygina, Elena Paducheva, Charles Fillmore and other researchers in the following respects:

1) Apart from such commonly used classes (and corresponding tags) as 'action', 'activity', 'process', 'state', 'property' and the like a number of new classes have been added, e. g. 'occupation', 'behaviour', 'impact', 'spatial position', 'interpretation' and so on. Further breakdown of the classes is based on such semantic oppositions as 'beginning' vs. 'cessation', 'causation' vs. 'elimination' etc. and such specific semantic features as 'volitional', 'emotional', 'quantitative', 'qualitative', or 'multiple'.

The set of semantic roles has also been revised. Apart from such familiar roles as 'agent', 'result', 'patient' and 'instrument', assigned for example to the verb *вязать* 'to knit' in such sentences as *Маша* ['agent'] *вяжет шарфы* ['result'] *из шерсти* ['patient'] *маминым крючком* ['instrument'] 'Masha knits scarves from wool with her mother's crochet hook', a number of new roles have been introduced, e.g.. such "temporal" roles as 'duration' (for certain Aktionsarten of Russian, cf. *проработать три часа* 'work for three hours'), 'date' (*Заседание было отложено до понедельника* 'The session was postponed till Monday'), 'term' (*аренда на пять лет* 'lease for five years').

Unlike that of object descriptors, this list of predicate descriptors forms a closed set.

2) The emphasis in selecting and assigning tags was on the continuity of natural-language semantic spaces and the kind of notation capable to reflect it. The semantic system of a natural language is not a hierarchy but a net with multiple "horizontal" and "vertical" intersections of classes. *Дышать* 'to breathe' is usually a process with 'patient' as its first actant, but in the situation of a medical examination it becomes an action, the role of its first actant changing to that of 'agent'. Prototypically, stativity manifests itself in mental states, like *to know that*, *to think that*, *to believe that*, while volitional (*to wish*) and especially emotional states (*to envy*, *to pride oneself on something*) are a step closer to processes.

3) The sum of tags assigned to a certain predicate lexeme is required to have certain explanatory and predictive power with respect to the non-semantic properties of lexemes – their patterns of government, combinatorial potential, or profile, derivational potential and even grammatical paradigms.

No formal restrictions are imposed on descriptor assignment. As is clear from the above, a lexeme in the semantic dictionary may have several descriptors of the same type; e.g., the verb *дышать* 'breathe' is assigned two descriptors, 'process' and 'action', to cover its unintentional and intentional uses. Moreover, the same lexeme may be simultaneously assigned both object and predicate descriptors. Thus, the entry for *отец* 'father' lists object descriptors 'human', and 'male', and predicate descriptors 'relation', 'kindred'. Additionally, it quotes semantic roles assigned to its two actants: 'object' and 'object2' (these are instantiated, respectively, by *Исаак* 'Isaac' and *Иаков* 'Jacob' in the sentence *Исаак – отец Иакова* 'Isaac is the father of Jacob'). The entry for the noun *взятка* 'bribe' lists an object descriptor 'money', three predicate descriptors 'action', 'social' and 'bad' and cites three semantic roles: 'agent', 'patient', and 'recipient', which are exemplified in the sentence *Контрабандисты предложили таможеннику взятку в 1000 долларов* 'The smugglers ['agent'] offered the customs official ['recipient'] a bribe of $1000 ['patient']'.

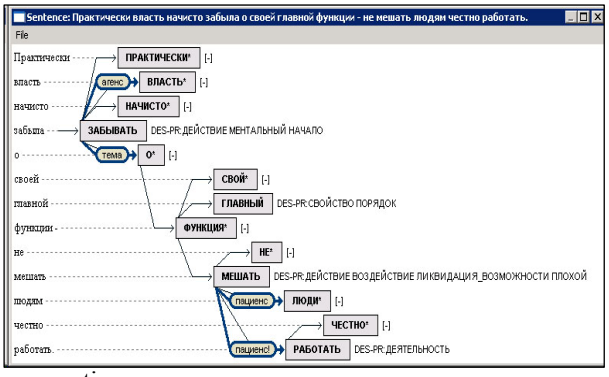
So far, semantic tagging has been produced on a tentative basis for a limited set of sentences. Fig. 2 shows partial tagging, made semi-automatically for sentence

(2) *Практически власть начисто забыла о своей главной функции – не мешать людям честно*

[2] See, in particular, Bulygina 1982, Fillmore et al. 2003, and Gildea and Jurafsky 2002 with further ample references.

работать ‘Practically, the authorities have completely forgotten about their main mission – not to interfere with the people’s fair work’.

This tagging was obtained using the 3,000-strong semantic entries; four of these entries – *забывать* ‘to forget’, *главный* ‘main’, *мешать* ‘to hinder, interfere’ and *работать* ‘to work’ occurred in (2). On the right of these four words, predicate descriptors are listed: ‘action’, ‘mental’ and ‘beginning’ for (the Russian equivalent of) *forget;* ‘property’ and ‘order’ for *main*, ‘action’, ‘effect’, ‘liquidation_of_possibility’ for *hinder*, and ‘activity’ for *work.* Besides, semantic roles are defined for *forget* (‘agent’ fulfilled by *authority* and ‘theme’ fulfilled by *mission*[3]) and *hinder* (‘patient’ fulfilled by *people* and ‘patient!’ fulfilled by *work*). The latter role, ‘patient!’ refers to such aspect of the patient that is directly affected



by an action.

Fig.2. Partial semantic tagging for sentence (2)

It must be added that semantic tagging is performed on sentences already annotated syntactically. On Fig. 2, which is a screenshot of the output produced by the Structure Editor, a software package specifically designed to facilitate corpus compilation, syntactic links between words of the sentence can be seen as thin unlabeled lines. To view the full syntactic annotation of sentence (2), one has to toggle the Editor, which will yield another image:

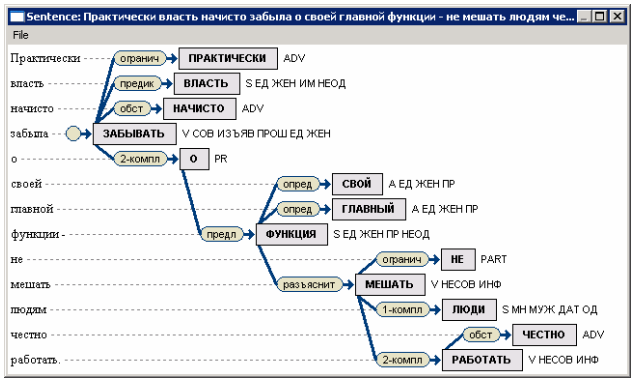


Fig.3. Full syntactic tagging for sentence (2)

Since the internal structures of the corpus (presented in normal XML format) contain both the syntactic and semantic tags for each sentence, the corpus allows for complex queries that may involve all kinds of language properties and as such can be considered, potentially, as a powerful instrument for linguistic research and NLP tasks.

In contrast to partial semantic tagging available now, Fig. 3 represents the syntactically annotated sentence (1) manually supplemented with full semantic tagging as aspired for in the present project. Such results will become possible when the semantic dictionary is expanded.

In Fig. 4, labels in square brackets represent object and

predicate descriptors whilst branches are marked with semantic roles.

Fig.4. Full syntactic and semantic tagging of sentence (1)

Here, the first word, *наибольший* ‘greatest’, is assigned the descriptors ‘characteristic’, ‘size’ and ‘big’; the second one, *возмущение* ‘indignation’ has the descriptors ‘state’ and ‘emotional’; *участник* ‘participant’ is labeled ‘person’, ‘action’, and ‘social’; *митинг* ‘meeting’ is labeled ‘event’ and ‘social’; *вызывать* ‘to cause’ and *устанавливать* ‘to set’ have one descriptor each, ‘causation of existence’; *продолжаться* ‘to continue’ is labeled ‘process’; *рост* ‘growth’ has descriptors ‘process’ and ‘quantitative’; *цена* ‘price’ has descriptors ‘parameter’ and ‘quantitative’; *бензин* ‘petrol’ and *нефтяной* ‘oil’ are labeled ‘substance’ and ‘liquid’, and *компания* ‘company’ is assigned the descriptors ‘aggregate’ and ‘human’. The semantic roles assigned to argument slots of some of the words are as follows: *вызывать* has two slots – ‘cause’ instantiated by *рост* and ‘result’ instantiated by *возмущение*; the latter has its own slot ‘experiencer’ instantiated by *участник*, which in its turn has a slot for ‘situation’ instantiated by *митинг*; *рост* has the slot for ‘patient’ instantiated by *цена*, whose ‘possessor’ slot is realized by *бензин*; finally, the ‘agent’ slot of *устанавливать* is represented by *компания.*

As can be easily seen, the idea underlying the enhancement of the Russian syntactically tagged corpus by semantic annotation is close to the endeavor of building a Proposition Bank from the Penn English Tree Bank (Kingsbury and Palmer 2002). The notable difference is that our semantic annotation envisages descriptors of words in addition to argument structures and roles of argument slots.

References

Apresjan, Ju. D. (2000). Systematic Lexicography. Oxford.

Apresjan, Ju. D. (2003). Фундаментальная классификация предикатов и системная лексикография. (A Fundamental Classification of Predicates and Systematic Lexicography. // Grammar Categories: Hierarchies, Links, Interaction. Proceedings

[3] Technically, this semantic role is instantiated by the preposition *о* ‘about’ which starts the prepositional group *о своей миссии* ‘about their mission’.

of an International Conference. Saint Petersburg, Nauka, pp. 7–21 (In Russian.)

Apresian, Ju., Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov, V., Tsinman, L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. // MTT 2003, First International Conference on Meaning – Text Theory. Paris, Ecole Normale Superieure, Paris, June 16-18 2003, pp. 279–288.

Apresjan, Ju.D. Iomdin, L.L., Sannikov, A.V. and Sizov, V.G. (2004a). Семантическая разметка в глубоко аннотированном корпусе русского языка (Semantic Tagging in the Deeply Annotated Corpus of Russian) // Proceedings of the International Conference "Corpus Linguistics–2004», pp. 41–54. (In Russian.)

Apresjan, Ju.D. *et al*. (2004b). Новый объяснительный словарь синонимов русского языка. (The New Explanatory Dictionary of Synonyms of Russian.) Second Edition. Moscow-Vienna: Yazyki slavjanskoj kultury. Wiener Slawistischer Almanach. Sonderband 60. (In Russian.)

Bulygina, T.V. (1982). К построению типологии предикатов в русском языке (On the Construction of Semantic Types of Predicates) // Semantic Types of Predicates. Moscow, pp. 7–85. (In Russian.)

Fillmore, Ch. J. (2003). Double-Decker Definitions: The Role of Frames in Meaning Explanations. // Sign Language Studies. Volume 3, Number 3, Spring 2003, pp. 263-295.

Fillmore, Ch. J., Johnson, C. R., Petruck, M. R.. (2003). Background to FrameNet. International Journal of Lexicography, 16, pp. 235–250.

Gildea, D., Jurafsky, D. (2002). Automatic labeling of semantic roles. // Computational Linguistics, vol. 28, Issue 3 (September 2002), pp. 245–288.

Kingsbury, P., Marcus, M. and Palmer, M. Adding (2002). Predicate Argument Structure to the Penn TreeBank. In: *Proceedings of the 2002 Conference on Human Language Technology*, San Diego, CA.

Mel'čuk, Igor A. (1988). Dependency syntax: Theory and practice. Albany, NY: SUNY.