SynTagRus – a deeply annotated corpus of Russian[1]

Igor Boguslavsky
Universidad Politécnica de Madrid /
Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow

Abstract.

The Russian dependency treebank, SynTagRus, is a subcorpus of the National Corpus of the Russian Language and at the time of writing (Spring 2013) contains over 52,000 sentences (roughly 770,000 words). It is supplied with several types of annotation. First, it contains comprehensive morphological and syntactic annotation. The latter is presented in the form of a full dependency tree that uses about 75 distinct dependency labels. Second, SynTagRus partly contains lexical semantic annotation, which means that, for all cases of word sense ambiguity in the corpus, the concrete lexical meaning should be identified and explicitly marked. So far, the number of SYNTAGRUS sentences fully tagged for word senses is over 10,000, and it is constantly growing. Third, a part of SynTagRus is annotated for collocate Lexical Functions (in Igor Melčuk's sense). SynTagRus is freely available for research and educational purposes.

1 Introductory Remarks

Tagged corpora are primarily intended for providing the basis for linguistic research in all fields of the vocabulary and the grammar (including changes occurring in the language throughout its history). There are two significantly different areas of such research. On the one hand, there are traditional linguistic studies for which mass material of texts is needed: such demand is much easier met if good and deeply tagged corpora are available. On the other hand, modern computational linguistics itself becomes an eager and interested user of such corpora as these are used on an increasing scale as training sets in machine learning. As a result of such learning, computer programs enhance their capability for extracting sophisticated types of data, which are contained in training text sets, from new texts. Generally speaking, the deeper the level of corpus annotation, the more advanced types of information could be learned from the corpus.

We describe the SynTagRus corpus developed by the Laboratory of Computational Linguistics (LCL) of the Institute for Information Transmission Problems of the Russian Academy of Sciences supplied with several types of annotation – morphological, syntactic, lexical semantic and lexical functional. This corpus serves both areas of linguistic research: traditional and computational.

2 Text Corpora of Russian

The largest corpus resource available for Russian is the so called National Corpus of the Russian Language ("Национальный корпус русского языка"), abbreviated as НКРЯ or NCRL, and available through the portal **www.ruscorpora.ru**. It combines several independently created subcorpora that contain different texts and are supplied with different types of annotation.

1) the **main corpus** provided with morphological annotation, which comprises over 300 million words belonging to written texts of a variety of genres starting from the 18th century. Since Russian is a morphologically rich language, with large paradigms for nouns, adjectives and verbs, in most cases, the annotation is morphologically ambiguous: a word may have more than one set of morphological tags corresponding to different parts of speech and/or morphological features. The main corpus contains a subcorpus of texts with resolved morphological ambiguity, which counts over 7 million words. Lexical ambiguity is not resolved independently; words can be considered to be lexically disambiguated only if the word senses belong to different parts of speech, as in the adjective *slepoj* 'blind' vs. the noun *slepoj* 'blind person'. The corpus now contains partial semantic annotation presented in the form of simple semantic features of words.

2) the **syntactic corpus** SYNTAGRUS provided with morphological and syntactic annotation. The corpus comprises about 770,000 words (over 52,000 sentences) and provides a syntactic dependency structure for all sentences. The corpus is fully disambiguated, both morphologically and syntactically, so that every word is supplied with one part-of-speech tag and a unique set of morphological features, while every sentence is has one, and only one dependency tree structure. Lexical semantic and lexical functional annotation exists in the stand-alone version of the corpus and is not displayed at the site of NCRL.

3) the **newspaper corpus**, built on the same principles as the main corpus and comprising the articles of seven mass media since the year 2000 (four newspapers published in Moscow and three electronic mass media), which now counts over 170 million words.

4) several aligned **parallel corpora** (English-Russian, Russian-English, German-Russian, Ukrainian-Russian, Russian-Ukrainian, Belarussian-Russian, Russian-Belarussian, and multilingual).

5) a **dialectal corpus**, composed of samples of dialect speech from various regions of Russia, presented in quasi-standard orthography. The corpus disregards phonetic variation but demonstrates morphological, lexical and syntactic peculiarities of regional and dialectal usage.

6) a **poetry corpus**, primarily comprising Russian poetic work of the 18th and 19th centuries supplemented by work of a number of 20th century poets. In addition to non-disambiguated morphological tagging built on the same principles as that of the main corpus, the texts are provided with information on the poetic meters used in them.

7) an **educational corpus**, intended for learners of Russian, which offers disambiguated morphological tagging for simple prosaic texts.

8) a **corpus of Spoken Russian**, which consists of transcripts of samples of public and private oral speech of the 20th and 21st centuries as well as film transcripts. The corpus is supplied with morphological and partial semantic annotation.

9) an **accentological corpus** (corpus of history of Russian word stresses).

10) a **Church Slavonic corpus**, which comprises modern lithurgical texts of the 19th and 20th centuries, as well as older religious and biblical texts. The corpus enables the search of words in three orthographic systems.

11) a **multimedia corpus** is composed of fragments of films released since 1930 and presented as video files, audio files and textual transcripts, as well as lists of gestures present in these fragments.

3 SYNTAGRUS Treebank

The Russian dependency treebank, SynTagRus, developed and maintained by the LCL (Boguslavsky et al. 2000, Apresjan et al. 2005), currently contains over 52,000 sentences (roughly 770,000 words) belonging to texts from a variety of genres (contemporary fiction, popular science, newspaper, magazine and journal articles dated between 1960 and 2013, texts of online news, 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 can be freely consulted on the Web (http://www.ruscorpora.ru/syntax-search.html).

Since Russian, as other Slavic languages, has a relatively free word order, SynTagRus adopted a dependency-based annotation scheme, in many respects parallel to the Prague Dependency Treebank (Hajič et al., 2001).

So far, SYNTAGRUS is the only corpus of Russian supplied with comprehensive morphological and syntactic annotation. The latter is presented in the form of a full dependency tree provided for every sentence. In the dependency tree, nodes represent words annotated with parts of speech and morphological features, while arcs are labeled with syntactic dependency types. There are about 75 distinct dependency labels in the treebank, half of which are taken from Igor Mel'čuk's Meaning ⇔ Text Theory (see e.g. Mel'čuk, 1988).

Fig.1 below is a sample dependency structure for the sentence:

Наибольшее возмущение   участников   митинга   вызвал
Most$_{NEUT,SG,ACC}$ indignation$_{SG,ACC}$ participant$_{PL,GEN}$ meeting$_{SG,GEN}$ cause$_{PAST,PERF,,SG,MASC,GEN}$

продолжающийся   рост   цен   на   бензин,
continue $_{PART,PRES,,IMPERF,,SG,MASC,NOM}$ growth $_{SG,NOM}$ price $_{PL,GEN}$ on $_{PREP}$ petrol$_{SG,ACC}$

устанавливаемых   нефтяными   компаниями
set$_{PART,PRES,,IMPERF,,PASS,PL,GEN}$ oil-Adj $_{PL,INSTR}$ company $_{PL,INSTR}$

'It was the continuing growth of petrol prices set by oil companies that caused the greatest indignation of the participants of the meeting'.
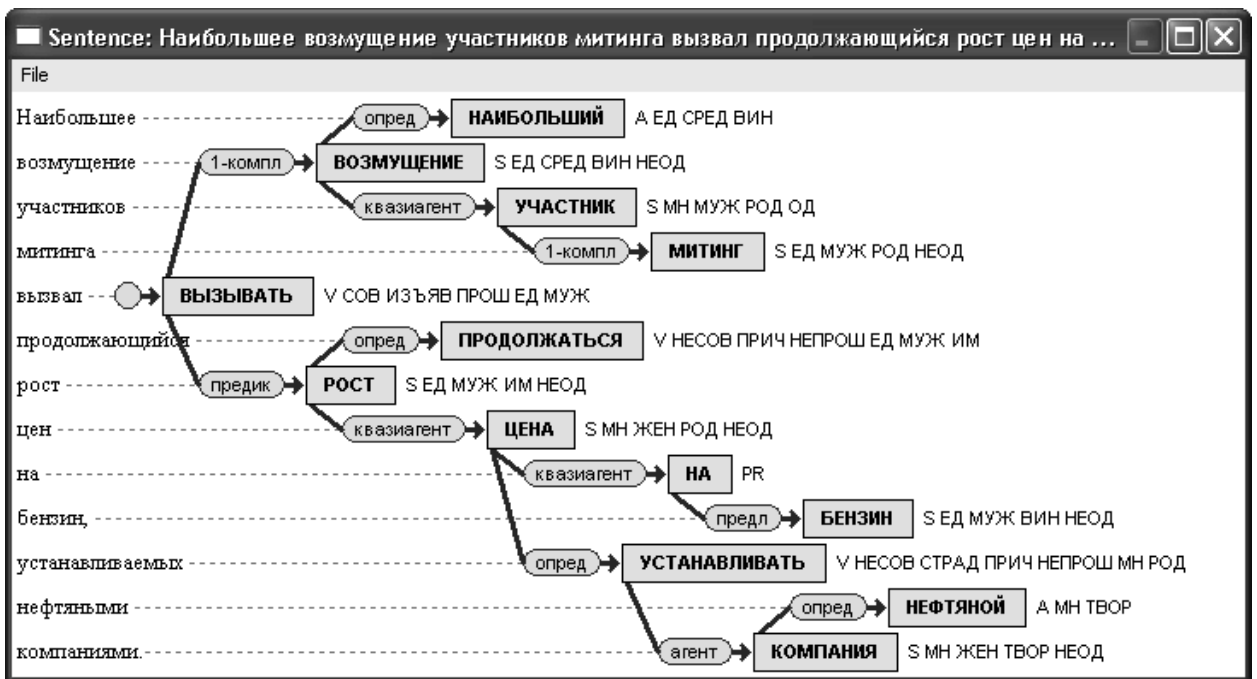
Fig.1. A syntactically tagged sentence.

Dependency types used in Fig. 1 include:

1. предик (predicative), which, prototypically, represents the relation between the verbal predicate as head and its subject as dependent;

2. 1-компл (first complement), which denotes the relation between a predicate word as head and its direct complement as dependent;

3. агент (agentive), which introduces the relation between a predicate word (verbal noun or verb in the passive voice) as head and its agent in the instrumental case as dependent;

4. квазиагент (quasi-agentive), which relates any predicate noun as head with the word implementing its first syntactic valency as dependent, if such a word is not eligible for being qualified as the noun's agent;

5. опред (modifier), which connects a noun head with an adjective/participle dependent if the latter serves as an adjectival modifier to the noun;

6. предл (prepositional), which accounts for the relation between a preposition as head and a noun as dependent.

Dependency trees in SYNTAGRUS may contain non-projective dependencies.

Normally, one token of the sentence (roughly, a word taken from space to space) corresponds to one node in the dependency tree. There are however a noticeable number of exceptions, the most important of which are the following:

1. compound words like *пятидесятиэтажный* 'fifty-storied', *стопятидесятипятимилиметровый* 'one hundred fifty five millimeter wide', where one token corresponds to two or more nodes;

2. so-called phantom nodes for the representation of hard cases of ellipsis, which do not correspond to any particular token in the sentence; for example, *я купил рубашку, а он галстук* 'I bought a shirt and he a tie', which is expanded into *я купил рубашку, а он купил*PHANTOM *галстук* 'I bought a shirt and he bought PHANTOM a tie';

3. multiword expressions like во что бы то ни стало 'whatever happened', where several tokens correspond to one node.

Morphological and syntactic annotation for SYNTAGRUS is performed semi-automatically: each sentence of the corpus is first processed by the rule-based Russian parser of an advanced multipurpose NLP system, ETAP-3 (Apresjan *et al*., 2003) and then edited manually by linguists, who correct errors made by the parser and handle cases of ambiguity that cannot be reliably resolved without extralinguistic knowledge.

Morphological annotation in SYNTAGRUS is based on a comprehensive morphological dictionary of Russian that counts about 130,000 entries (over 4 million word forms). The ETAP-3 morphological analyzer uses the dictionary to produce morphological annotation of words belonging to the corpus, including the lemma, the part-of-speech tag and additional morphological features dependent on the part of speech: e.g. values of such features as 1) animacy, gender, number, case, degree of comparison, short form – for adjectives and participles, 2) representation (with values of finiteness, infinitive, participle, or gerund), aspect, tense, mood, person, voice – for verbs, etc. The morphological analyzer operates in a context-free manner, offering almost no morphological disambiguation for the sentence.

The syntactic parser processes morphologically analyzed sentences using a set of syntactic rules, or syntagms, that produce one binary syntactic link each. Like many similar parsers, ETAP-3 uses a statistics-based prior part-of-speech tagging module.

When editing SYNTAGRUS annotation, the developers use a powerful software tool, Structure Editor, which enables them to easily access all sorts of data necessary for efficient work (ETAP-3 dictionaries and rules) and handle even the hardest cases in a smooth and consistent way.

4 Lexical Semantic Annotation

Lexical semantic annotation means that, for all cases of word sense ambiguity of the corpus, the concrete lexical meaning should be identified and explicitly marked. In its present state, SYNTAGRUS does not provide exact lexical meanings, showing only the lemmas of the words occurring in texts. This means that lexical ambiguity is only resolved in the corpus if ambiguous words happen to have different lemmas and/or different part of speech tags. Accordingly, SYNTAGRUS distinguishes between *печь* as a verb ('bake') and *печь* as a noun ('oven') or the pronominal adjectives *сам* 'oneself" and *самый* 'very', so that ambiguous sentences containing ambiguous word forms like *Она любит печь* 'She likes to bake' vs. 'She likes the oven', or *Я знаю самого главного инженера* 'I know the chief engineer himself' vs. 'I know the most important engineer' can be distinguished in the corpus. Conversely, if ambiguous lexemes (no matter whether they belong to one polysemic vocable or are lexical homonyms) have the same lemmas, they are not distinguished. For this reason, even very different words are underrepresented if they happen to have the same lemmas.

In the new corpus, all ambiguous lemmas will be supplied with concrete word senses as they are specified in the combinatorial dictionary of Russian. This dictionary is a vital component of the ETAP-3 linguistic processor that counts almost 100,000 words. Thanks to this annotation, corpus users will be able to search for lexical meanings of words and study lexical ambiguity in broad linear and syntactic context. Among other things, we expect that such data will contribute to the development of a statistically driven module of automatic word sense disambiguation for Russian.

Benefits that accrue from lexical semantic annotation of the corpus can be illustrated by the ambiguous Russian verb *толковать*. This verb has (at least) three manifestly different lexical meanings: *толковать 1* 'interpret', "define' (in a dictionary, law etc.), as in *Русские словари толкуют честолюбие как негативную черту характера* 'Russian dictionaries interpret ambition as a negative character trait', *толковать 2* 'explain insistently', 'try to convince' *Он толковал мне, почему я ошибаюсь* 'He was explaining to me why I am wrong, and *толковать 3* 'converse', 'discuss', 'reason', as in *Они долго толковали о чем-то* 'They conversed long about something'. Importantly, these lexical units have very different linguistic properties. These properties, fully documented in the dictionaries of ETAP-3, include:

(1) valency structures (*толковать 1* has a subcategorization frame close to that of *интерпретировать* 'interpret': *толковать что-л. Как что-л.* 'define smth. As smth' or *толковать что-л. Через что-л.* 'define smth. Through <with, by way of> smth.; the subcategorization frame of *толковать 2* resembles (but is not identical to!) that of *объяснять* 'explain': *толковать о чем-л. Кому-л.* 'explain smth. To smb', whilst *толковать 3* approaches the behavior of the symmetrical verb *беседовать* 'talk': *толковать о чем-л. С кем-л.* 'speak about smth. With smb';

(2) derivation (*толковать 1* has a deverbal noun *толкование* 'act of interpretation' or 'lexicographic definition', while *толковать 2* and *толковать 3* have no derivatives), and even

(3) morphological peculiarities '*толковать 1* is a transitive verb which has passive forms but it has no perfective aspect; *толковать 2* is, formally, a transitive verb (even though its direct object can only be realized by certain pronouns in the accusative case, like *толковали что-нибудь, <такое, свое>* 'explained something <something of this kind, their own thing>') but has neither passive forms or perfective aspect, whereas *толковать 3* is an intransitive verb that has no passive forms but it has the perfective aspect *потолковать*).

In a lexically underspecified corpus, it is impossible to sort out sentences that contain *толковать* in one particular sense, so it would be hard to establish, validate or rectify the information on specific lexical units, which could otherwise be used in many actual tasks (including those requiring machine learning).

It should be emphasized that, since SYNTAGRUS is compiled semi-automatically, in many cases the linguist expert that edits the results of automatic parsing corrects the resulting structure containing particular words even it is not corroborated by the existing dictionary or grammatical data (which may be incomplete or not very accurate), without actually updating such data – the natural reason being that the expert may lack expertise, authority, or simply time. As a result, the deeply tagged corpus – not only SYNTAGRUS but any corpus built on similar principles – acts, in many respects, as a source of invaluable data for linguists.

To continue with the example of *толковать*, a corpus that distinguishes word senses will enable us to see that e.g.

(a) the sentence *Ресторанные словари толкуют о каком-то соусе и каштанах* 'restaurant dictionaries [whatever these are!] talk about some sort of sauce and chestnuts' contains the word *толковать 2,* rather than *толковать 1,* which could be anticipated in the context of the word *словари* 'dictionaries';

(b) for subtle semantic reasons (*writing techniques* hardly needs interpretation, and interpretation hardly requires an addressee) the sentence *Помню, как он улыбался, толкуя мне китайскую письменность* 'I remember how he was smiling, explaining to me Chinese writing' also contains the word *толковать 2* despite the fact that the verb in this sense hardly accepts a non-pronominal direct object, unlike *толковать 1* – and the respective piece of information on this verb sense should be added to the dictionary;

(c) the sentence *Боюсь, что она это превратно понимает и толкует, как будто я забыл ее и не хочу ее видеть* 'I am afraid that she misapprehends it and interprets (it) as though I have forgotten her and do not wish to see her' contains *толковать 1* even though its third valency (of content) is presented in a highly non-canonical way – by a subordinate clause introduced with the conjunction *как будто.*

As follows from these examples, it is not at all easy to provide quality lexical semantic annotation of the corpus: this endeavour requires much time – and intellectual labour – of experienced annotators. The amount of work to be done can be properly assessed if we take into account the number of ambiguous words in 100,000-strong ETAP-3 dictionary (ca. 3,300 vocables whose lexemes share the same lemma names, representing about 6,700 word senses). We strongly believe, however, that the resulting corpus will be well worth this effort.

Fig. 2 below presents the structure of a corpus sentence from (c) with ambiguous words marked for concrete senses (here, words *что 1, толковать 1, как будто 1, и 1* and *не 1* specify such senses), while Fig. 3 summarizes the information on one of the respective lexemes – *толковать 1.*
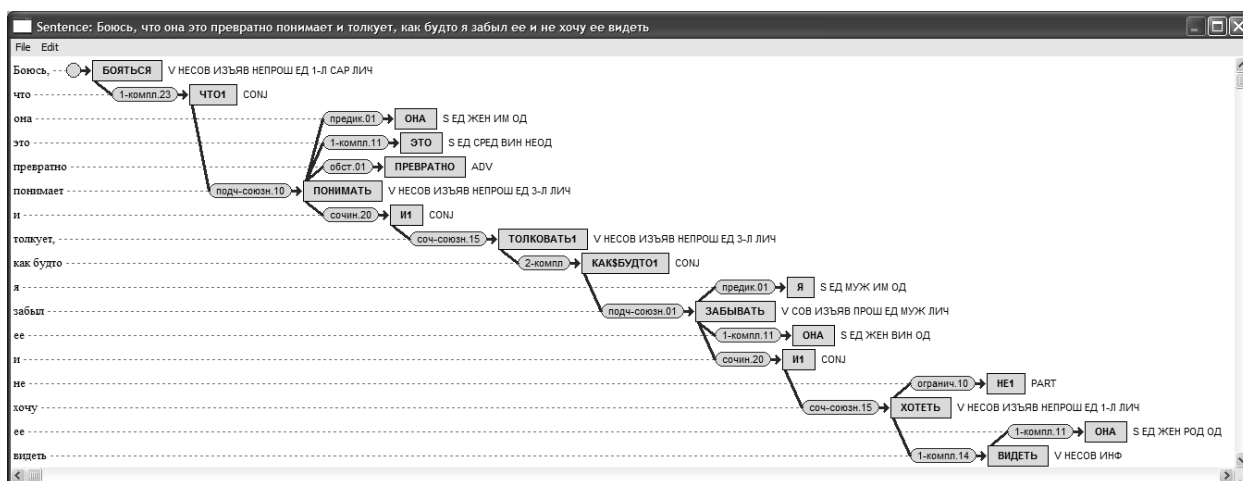

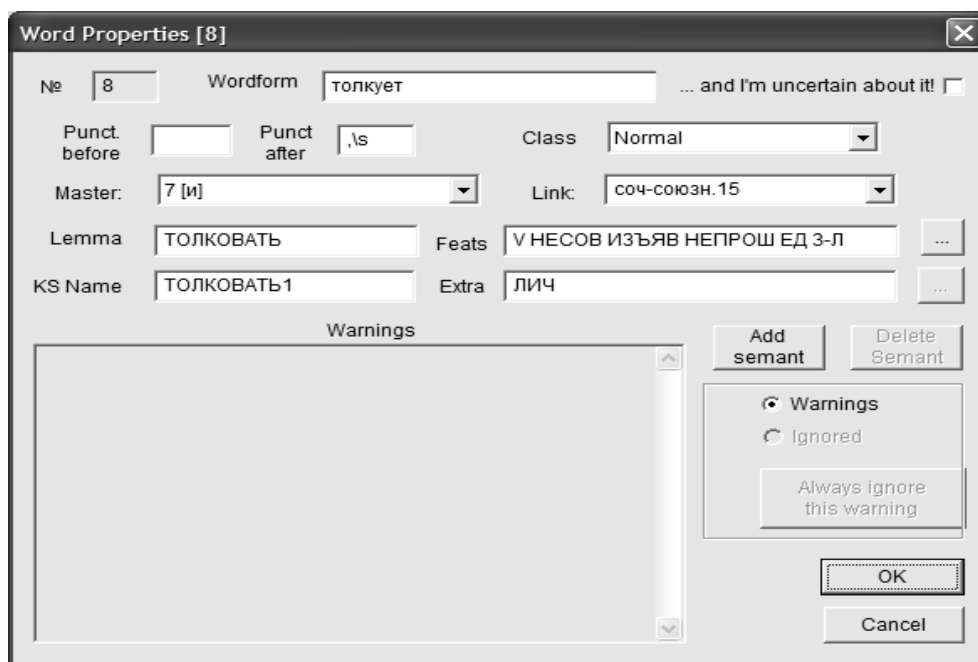Fig.2. A sentence tagged for syntax and lexical semantics

Fig.3. Properties of a specific word from the corpus sentence.

Most of the boxes of the view presented by Fig. 3 are self-evident. KS name is that of the the word's entry in the combinatorial dictionary of ETAP-3: it is clear that the corpus essentially relies on this particular dictionary, so that future researchers working with this corpus may require access to it.

So far, the number of SYNTAGRUS sentences fully tagged for word senses is over 10,000, and it is constantly growing.

Apparently, lexical semantic annotation adds predictive power to the corpus and makes it a much more valuable linguistic resource.

5 Lexical Functional Annotation

Lexical functional annotation consists in detecting lexical functions (LF) and their values in texts and tagging them for this type of data. Specifically, we plan to reveal occurrences of collocate type LFs in the sentences of SynTagRus and record them as part of sentence annotation. As a result, ample LF material will be available to researchers. So far, very few dictionary resources have provided such data. If collocates are marked in the text, direct observation and research of contexts in which lexical functions are realized will be possible. These data are of immense value for natural language processing systems.

The notion of lexical function was first proposed by the author of the Meaning ⇔ Text linguistic theory, Igor Mel'čuk, in 1970s and has been extensively studied and developed by the Moscow Linguistic School, in particular, by the Laboratory of Institute of Information Transmission Problems with active participation of Juri Apresjan. We have developed a number of NLP applications using LFs, including machine translation, where LFs are used to resolve lexical and syntactic ambiguity and achieve idiomatic translation of collocations, and an experimental system of synonymous paraphrasing for Russian.

A prototypical LF is a triple of elements {R, X, Y}, where R is a certain general semantic relation obtaining between the argument lexeme X (the keyword) and some other lexeme Y which is the value of R with regard to X (by a lexeme in this context we mean either a word in one of its lexical meanings or some other lexical unit, such as a set expression). Y is often represented by a set of synonymous lexemes Y1, Y2, …., Yn, all of them being the values of the given LF R with regard to X. To give a simple example, MAGN is a LF for which the semantic relation is 'high degree'. Respectively for English,

MAGN (*desire*) = *strong / keen / intense / fervent / ardent / overwhelming*,

and for Russian,

MAGN (*желание*) = *сильный /упорный/ настойчивый / горячий / страстный / неудержимый / неутолимый / большой.*

Two types of LFs are distinguished: – paradigmatic LFs (substitutes) and syntagmatic LFs (collocates, or parameters).

Substitute LFs replace the keyword in the given utterance without substantially changing its meaning or changing it in a strictly predictable way. Examples are synonyms, antonyms, converse terms. A special subclass of substitute LFs is represented by various types of derivatives of X (*nomina actionis*, as in *to encourage – encouragement*, typical agents, as in *to build – builder* or *to judge – judge*, typical patients, as in *to nominate – nominee, to teach – student* and the like). All of them play an important role in paraphrasing sentences. Cf., for example: *She bought a computer for 500 dollars from a retail dealer – A retail dealer sold her a computer for 500 dollars – She paid 500 dollars to the retail dealer for a computer – The retail dealer got 500 dollars from her for a computer*

Collocate LFs appear in an utterance together with the keyword. Typically, such LFs either dominate the keyword syntactically or are dominated by it, although more elaborate syntactic configurations between the keyword and an LF value are not infrequent. Typical examples of collocate LFs are adjectival LFs, such as MAGN, or support verbs of the OPER / FUNC family.

This family of LFs can be exemplified by OPER 1 – a semantically empty verb such that the **first** actant of a certain situation functions as the subject of this verb and the name of the situation itself is the verb's first object: In Russian, OPER 1 (*контроль*) = *осуществлять* (cf. in English OPER1 (control) = *to exercise (control)*).

In much the same way, OPER 2 is a semantically empty verb such that the **second** actant of a certain situation functions as the subject of this verb and the name of the situation itself is the verb's first object: OPER2 (*контроль*) = *подвергаться (контролю), находиться под (контролем), быть под (контролем)* (cf. in English OPER2 (control) = *be under (control)*).

Collocate LFs play a leading role in the paraphrasing system of ETAP-3, providing paraphrases like *He respects his teachers – He has respect for his teachers – He treats his teachers with respect – His teachers enjoy his respect*", or *The United Nations ordered Iraq a report on chemical weapons – the United Nations gave Iraq an order to write a report on chemical weapons – Iraq was ordered by the United Nations to write a report on chemical weapons – Iraq received an order from the United Nations to write a report on chemical*

weapons.

We are planning to mark a substantial part of our corpus with lexical functional annotation, too. As with syntactic and lexical semantic annotation, this work will be done semi-automatically. Since the ETAP-3 parser has a set of special post-syntactic rules that identify argument and values of most collocate LFs (primarily if they appear in prototypical syntactic positions), the results will be used as raw material for manual correction and tagging by the annotator.

To give an example, for the sentence *Лил проливной дождь* 'A heavy rain was pouring' the parser will provide the following information on lexical functions:

```
MAGN(ДОЖДЬ) = ПРОЛИВНОЙ
FUNC0(ДОЖДЬ)  ЛИТЬ1
```

These data will supplement the syntactic and lexical semantic tagging of SYNTAGRUS.

6 How can SynTagRus be used?

SYNTAGRUS has already been used for a number of linguistic research and application tasks. In particular, it has been used as benchmark in regression tests designed to ensure stable performance of the ETAP-3 Russian parser in the course of its development (see e.g. Boguslavsky *et al.* 2008) and as a source for the creation, by machine learning methods, of a successful statistical parser for Russian (Nivre *et al.*, 2008).

International academic community showed considerable interest in using SYNTAGRUS for research. Over 30 free licenses were provided to universities and academic institutions, and several commercial licenses were provided to big IT companies.

References

Apresjan *et al.*, 2003: Apresjan, Ju., I. Boguslavsky, I., Iomdin, L., Lazursky, A., Sannikov, V., Sizov V, and Tsinman, L. ETAP-3 linguistic processor: A full-fledged NLP implementation of the MTT. // Proceedings of the First International Conference on Meaning-Text Theory, 279–288.

Apresjan *et al.*, 2005: Apresjan, Ju.D., Boguslavsky, I.M., Iomdin, L.L., Iomdin, B.L., Sannikov, A.V., Sannikov, V.Z., Sizov, V.G., Tsinman, L.L. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы. (Syntactically and Semantically Annotated Corpus of Russian: State-of-the-Art and Prospects) // National Corpus of Russian 2003–2005 (Results and Prospects). M: Indrik, 2005. P.193-214. (In Russian.)

Boguslavsky *et al.*, 2000: Boguslavsky, I., Grigoriev N., Grigorieva S., Kreidlin L., Frid N. Dependency Treebank for Russian: Concept, Tools, Types of Information // Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), 2000, p. 987-991.

Boguslavsky *et al.*, 2008: Boguslavsky, I.M., Iomdin, L.L. Valeev, D.R., and Sizov, V.G. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов (ETAP Parser and its Evaluation Using a Deeply

Annotated Corpus of Russian Texts) // Corpus Linguistics 2008. St. Petersburg, St. Petersburg State University, 2008. ISBN 978-5-288-04769-5. P. 56-74.

Hajič et al., 2001: Hajič, J., Vidova Hladka, B., Panevová, J. E. Hajičová, P. Sgall, and P. Pajas. 2001. Prague Dependency Treebank 1.0. LDC, 2001T10.

Mel'čuk, 1988: Mel'čuk, Igor. Dependency Syntax: Theory and Practice. State University of New York Press, 1988.

Nivre *et al.*, 2008. Nivre, Ioakim; Boguslavsky, Igor; Iomdin, Leonid. Parsing the SYNTAGRUS Treebank of Russian // Coling 2008. 22nd International Conference on Computational Linguistics. Proceedings of the Conference. ISBN: 978-1-905593-47-7. Vol. 2. pp. 641-648.