Bank of Russian Constructions and Valencies

Olga Lyashevskaya
University of Tromsø, Norway
Institute of Russian Language RAS, Moscow, Russia
E-mail: olesar@mail.ru

Abstract

The Bank of Russian Constructions and Valencies (Russian FrameBank) is an annotation project that takes as input samples from the Russian National Corpus (http://www.ruscorpora.ru). Since Russian verbs and predicates from other POS classes have their particular and not always predictable case pattern, these words and their argument structures are to be described as lexical constructions. The slots of partially filled phrasal constructions (e.g. vzjal i uexal 'he suddenly (lit. took and) went away') are also under analysis. Thus, the notion of construction is understood in the sense of Fillmore's Construction Grammar and is not limited to that of argument structure of verbs.

FrameBank brings together the dictionary of constructions and the annotated collection of examples. Our goal is to mark the set of arguments and adjuncts of a certain construction. The main focus is on realization of the elements in the running text, to facilitate searches through pattern realizations by a certain combination of features. The relevant dataset involves lexical, POS and other morphosyntactic tags, semantic classes, as well as grammatical constructions that introduce or license the use of elements within a given construction.

1. Introduction

The paper describes a new initiative aimed at the annotation of constructions that emerge around verbs, (predicate) nouns, adjectives, adverbs as well as phrasal constructions (in the sense of Fillmore's Constructicon subproject of FrameNet, cf. *let alone, in its own right* etc., Fillmore 2008). The work is based on samples from the Russian National Corpus (RNC) that currently contains 180M words provided with morphological and semantic annotation (Lashevskaja & Shemanaeva 2008, Kustova et al. 2009). A small part of the RNC (0.5M words) bears syntactic annotation.

The project continues the tradition of lexical resources such as FrameNet (Johnson et al. ER), VerbNet (Kipper et al. 2006), PropBank (Palmer et al. 2005), NomBank (Meyers 2007), Vallex (Lopatková et al. 2006) developed for English, Czech and other languages, which are mostly considered to be dictionaries of valencies and constructions. At the same time, it results in deeply annotated corpus data for NLP tasks such as automatic semantic role annotation, machine translation, information retrieval etc. (cf. Gildea & Jurafsky 2002, Gerber et al. 2009, Bick & Pilar Valverde 2009, among others).

The Bank of Russian Constructions and Valencies combines a dictionary and an annotated corpus in a single device. It promises to be the largest on-line pattern lexicon available for Russian that gives an exhaustive number of corpus examples balanced within time periods (18th – 21st centuries) and genres and allows us to estimate the distribution of diverse shallow patterns assigned to a lexical unit.

The paper is organized as follows. Section 2 the outlines the types of information encoded in the annotation while section 3 gives a description of the annotation scheme. Section 4 presents the software and how is used in an annotation mode, section 5 provides a brief overview of its use in a search mode. Section 6 offers conclusions.

2. Data

The Bank is certainly a FrameNet-oriented application. Yet, more attention is paid here to surface morphology and word order and both lexical and semantic restrictions on construction slots are placed under the microscope. The first feature can be explained by the very nature of Russian as inflection-rich language while the second is inspired by theoretical principles of Moscow School of lexical semantics (Ju. Apresjan, I. Boguslavsky, E. Paducheva, G. Kustova and others).

The Bank contains the following types of information:

0) construction: a set of elements, obligatory and optional arguments and adjuncts;

1) shallow morphosyntax: POS, case and other grammatical features that constrain the element position;

2) syntactic rank of the element (Subject, Object, Peripheral, Adjunct, No);

3) semantic role or explication of the element;

4) the phrase that matches the element of the construction;

5) head of the phrase: the lemma and its semantic class;

6) grammatical constructions that license the overt expression of the elements;

7) constructions that introduce additional arguments (new participants to the frame);

8) other pragmatic and information structure parameters that explain omission of the participants;

9) word order.

There are two parts in the project, Dictionary and Realizations. The dictionary provides standard templates that describe arguments of the construction and their expression in neutral context not affected by other grammatical constructions. Points (1)-(3) and (5) mirror the Argument structure section of the "Lexicograph" project (Kustova&Paducheva 1994, www.lexicograph.ru). The preliminary list of constructions was adopted from Apresjan&Pall 1982 and Mel'chuk&Zholkovsky 1984 dictionaries of Russian valencies as well as morphosyntactic data (1).

Realizations describe which elements of the construction are expressed in a sentence and in what particular way. All types of information (1)-(9) are encoded here.

All constructions attested in the dictionary are connected in a graph with labeled relations between them; another sort of relation is established between individual elements.

In the reference section, dictionary templates contain links to other Russian lexical resources: lexical entries in the on-line dictionaries MAS and Ozhegov&Shvedova; "Lexicograph" database of Russian verbs; RussNet, a Russian version of WordNet (Azarova 2008), hopefully accessible online in the near future. Two other links encourage the user to look up the uses of a word in the Russian National Corpus including Main corpus and Syntactic TreeBank.

Link relations are also established with well-known English resources: Unified verb index, the crossmap of VerbNet, FrameNet, PropBank and WordNet (Loper et al. 2007), and NomBank.

3. Annotation scheme

The project presupposes annotation of a large sample of patterns realized in corpus texts. At present, each target lexical unit is illustrated by 100 sentences accompanied by their pre- and post-context.

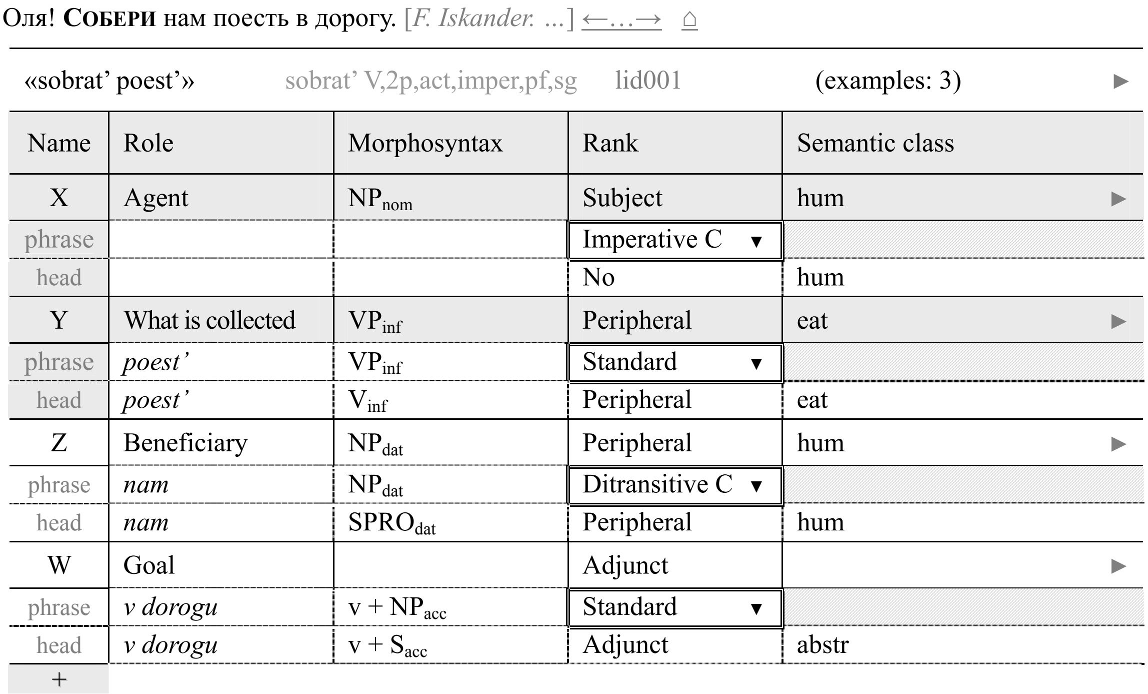
To minimize the amount of mistakes and subjective decisions, each sentence is to be tagged by two annotators.

As a first step, examples are matched to a particular target word entry in the dictionary, i. e. to a certain word sense and an appropriate argument pattern attested for this sense. After that, an annotator marks up the relevant pieces of a sentences linking them with elements of a construction.

The next task is to define the marked arguments in terms of semantic roles, syntactic ranks as well as provide explanation about missing arguments. Figure 1 shows an annotated verb pattern «sobrat' poest'» 'pick up something to eat' as it is realized in example (1).

(1)	Olja!	**SOBERI**	nam	poest'	v	dorogu.
	Olja.S.NOM	collect.V.IMPER	we.SPRO.DAT	eat.V.INF	in.PR	way.S.ACC

'Olja! Pick up something to eat for us on the way.'



Оля! **СОБЕРИ** нам поесть в дорогу. [*F. Iskander.* ...] ←...→ ⌂

«sobrat' poest'» sobrat' V,2p,act,imper,pf,sg lid001 (examples: 3) ▶

Name	Role	Morphosyntax	Rank	Semantic class
X	Agent	NP$_{nom}$	Subject	hum ▶
phrase			Imperative C ▼	
head			No	hum
Y	What is collected	VP$_{inf}$	Peripheral	eat ▶
phrase	*poest'*	VP$_{inf}$	Standard ▼	
head	*poest'*	V$_{inf}$	Peripheral	eat
Z	Beneficiary	NP$_{dat}$	Peripheral	hum ▶
phrase	*nam*	NP$_{dat}$	Ditransitive C ▼	
head	*nam*	SPRO$_{dat}$	Peripheral	hum
W	Goal		Adjunct	▶
phrase	*v dorogu*	v + NP$_{acc}$	Standard ▼	
head	*v dorogu*	v + S$_{acc}$	Adjunct	abstr
+				

Figure 1. Realization of the construction «sobrat' poest'» (target verb *sobrat'* 'pick up') in example (1).

The slots X and Y are described in the valency dictionary as an Agent and What is collected, respectively. The former is a human ('hum') Subject expressed by a Nominal noun (S_{nom}), and the latter is a verb of eating in the Infinitive form (V_{inf}; semantic restriction: 'eat'). This general information attested in the dictionary is colored grey. In a particular example, the Subject is omitted, so there is null instantiation of the argument X licensed by the Imperative construction ("Imperative C"). Y matches its infinitive verb phrase *poest'* in a predictable way ("Standard").

In addition to that, the annotator has marked two additional participants Z and W in example (1), namely Beneficiary and Goal. The Beneficiary argument is introduced by the Ditransitive construction that allows almost every verb to add dative arguments. It is expressed by the Dative pronoun ($SPRO_{dat}$) *nam* ‘for us’. The last slot is treated as Adjunct and corresponds to the prepositional phrase *v dorogu* ‘on the way’. The head noun in the Accusative case (S_{acc}) *dorogu* ‘road, way’ is used in an abstract sense here, so it is marked as ‘abstr’ in the Semantic class column.

4. Software

The data are stored in MySQL tables for online search and in XML files suitable for import into other applications. Mikhail Kudinov (Moscow State University) has developed an online annotation and search tool.

The annotation process is designed to run in semi-automated mode: after the annotator selects an appropriate pattern/construction from the dictionary and matches its slots to the relevant parts of a sentence, the tool extracts information about case marking, POS, lemma and semantic class from tags available in the corpus in its grammatical and semantic layers. In some cases this information needs correction and can be edited manually afterwards. The program checks the completeness and consistency of annotated data and may highlight some possible problems in sentence markup for the annotator (for example, words assigned more than one role) as well as lack of coordination in sentence representation and dictionary data. The dictionary templates can be improved manually, and the software maintains a history of changes made by each annotator.

The tool also measures some parameters of project statistics, time spent by an annotator and inter-annotator agreement.

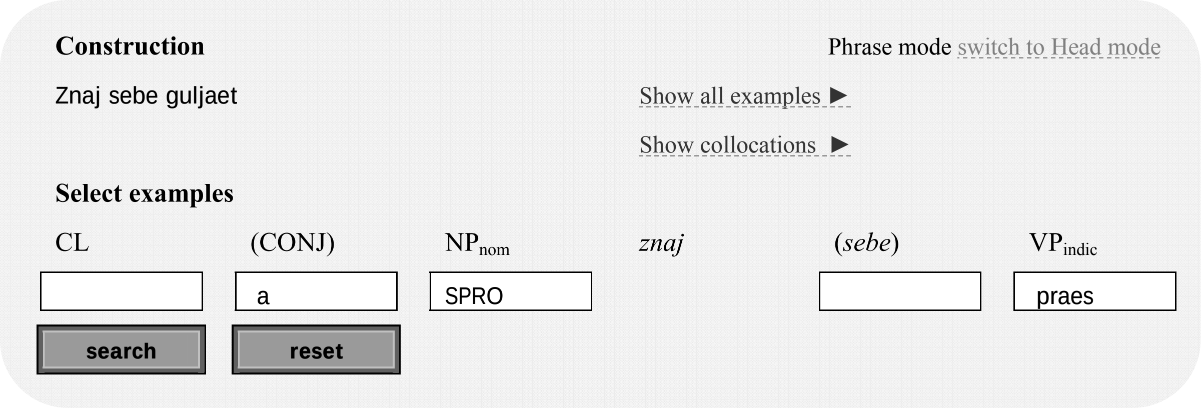
5. Online search engine

In the search mode, the main page presents the user with lists of annotated Constructions and target Lexical units. A click on a lexical unit (lemma) takes the user to the List of constructions attested in the dictionary and the List of examples where it is used. Example Passport tells the user about authors, genres, date of creation and other meta-textual information. Word Passport shows all types of grammatical and semantic information for each word in a sentence.

Construction Passport page contains a pattern template from the dictionary; a set of links redirects the user to the list of examples and the list of collocations attested in the RNC. Another template illustrates the Realization of the construction in a particular sentence, see Fig. 1 above.

Since each construction is documented in terms of grammatical and semantic restrictions on construction slots, on the one hand, and lexical, grammatical and semantic tags of words attested in the corpus, on the other hand, the user may be interested to see how these features interact. The Select examples by pattern form allows the user to specify a subset of examples using the combination of particular features.

Figure 2 illustrates how the user can select those examples of the construction *znaj sebe guljaet* ‘(he) would just go on walking’ where the initial clause (CL) is followed by the conjunction (CONJ) *a* ‘and, but’, the subject of the second clause (NP_{nom}) is a pronoun (SPRO) and the head of the last verb phrase (VP_{indic}) shows up in Present form (praes).



Construction

Phrase mode switch to Head mode

Znaj sebe guljaet

Select examples

CL	(CONJ)	NP_{nom}	*znaj*	(*sebe*)	VP_{indic}
	a	SPRO			praes

Figure 2. Search form based on the features of elements within the construction.

6. Results and future work

The project reported in this paper is a work in progress. So far, a pilot set of ca. 100 verbs, 20 nouns, 20 adjectives, 20 adverbs and 20 phrasal constructions has been annotated, amounting to a total of 18 000 annotations (100 sentences per unit). These data are intended to serve as a model for the improvement of the methodology and guidelines, evaluation of the annotation scheme and development of the software tool. The planned 1.0 release will be focused mostly on verbs, the number of which will amount to 1000 units, and 100 sentences of each perfective and corresponding imperfective verb as well as their reflexive counterparts will be annotated. The release will be stored for distribution in XML format.

The main focus thus is on realization of valencies in the running text, which makes it possible to search through pattern realizations by a certain combination of features.

Future work will involve developing the hierarchy of verb patterns (frames) and their elements (FEs), as well as exploring the domain of nominal predicates and phrasal constructions. The latter task presents a substantial challenge to the project since the argument structure of Russian nouns and adjectives as well as slots of phrasal constructions are much less documented.

7. References

Apresjan, Jurij, and Erna Pall. 1982. *Russkij glagol – vengerskij glagol: upravlenie i sočetaemost'*. Budapest.

Azarova, Irina. 2008. RussNet as a computer lexicon for Russian. *16th International Conference Intelligent Information Systems 2008.* Zakopane, Poland. 341-350. http://iis.ipipan.waw.pl/2008/proceedings/iis08-33.pdf.

Bick, Eckhard, and M. Pilar Valverde Ibáñez. 2009. Automatic semantic role annotation for Spanish. *NODALIDA 2009*. http://dspace.utlib.ee/dspace/bitstream/10062/9762/1/paper47.pdf.

Bohmova, Alena, Silvie Cinkova, Eva Hajicova. 2005. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank. Electronic resource: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

Fillmore, Ch. J. 2008. Border conflicts: FrameNet meets Construction Grammar. *EURALEX 2008*. http://www.hf.uib.no/forskerskole/0415FNMCG.pdf.

Gerber, Matt, Joyce Y. Chai, and Adam Meyers. 2009. The role of implicit argumentation in nominal SRL. *HLT: The 2009 Annual Conference of the North American Chapter of the ACL*, 146–154.

Gildea, Daniel, and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

Johnson, C., Fillmore C., Petruck M, Baker C., Ellsworth M., Ruppenhofer J., and Wood E. *FrameNet: theory and practice*. Electronic resource, available from: http://www.icsi.berkeley.edu/ framenet.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. *LREC 2006*. Genoa, Italy. June, 2006.

Kustova, Galina, and Elena Paducheva. 1994. Semantic dictionary as a lexical database. *EURALEX-94*. Amsterdam, 1994.

Kustova, Galina, Olga Lashevskaja, Elena Paducheva, and Ekaterina Rakhilina. 2009. Verb taxonomy: From theoretical lexical semantics to practice of corpus tagging. In B. Lewandowska-Tomaszczyk, K.Dziwirek (eds.), *Studies in Cognitive Corpus Linguistics*. Frankfurt, 41–56.

Lashevskaja, Olga, and Olga Shemanaeva. Semantic Annotation Layer in Russian National Corpus: Lexical Classes of Nouns and Adjectives. *LREC 2008*. Marrakesh, Morocco, May 2008.

Lopatková, Markéta, Zdeněk Žabokrtský, and Karolína Skwarska. 2006. Valency lexicon of Czech verbs: alternation-based model. *LREC 2006*, vol. 3, pages 1728-1733. ELRA, 2006. http://ufal.mff.cuni.cz/vallex/2.5/publ/06-LREC-alter.pdf.

Loper, Edward, Szu-ting Yi and Martha Palmer. 2007. Combining lexical resources: mapping between PropBank and VerbNet. *7th International Workshop on Computational Semantics, 10-12 January 2007, Tilburg, the Netherlands*. http://www.cis.upenn.edu/~edloper/publications/semlink_iwcs07.pdf. Unified verb index available at http://verbs.colorado.edu/verb-index/index.php.

Mel'chuk, Igor, and Aleksandr Zholkovsky. 1984. *Explanatory combinatorial dictionary of modern Russian*. Wiener Slawistischer Almanach, Vienna.

Meyers, Adam. 2007. *Annotation guidelines for NomBank – noun argument structure for PropBank*. Technical report, New York University.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.