Dictionary of Valencies Meets Corpus Annotation: A Case of Russian FrameBank[1]

Olga Lyashevskaya

Keywords: *frame semantics, FrameNet, Construction Grammar, Russian*

Abstract

The Russian FrameBank project aims at the development of a hybrid lexical resource that links a dictionary of valencies and an annotated corpus. Two types of data present generalized lexical constructions (LexCxs) and their realizations in contemporary written texts (1950-present).

The predicate-argument structure for verbs, nominalizations, adjectives, adverbs, and other lexical units in Russian is mostly encoded in case and prepositional marking while word alignment is determined by information structure. This means that an argument can be found in any part of the sentence and the window for argument detection is infinitely wide. Russian predicates reveal more than 1000 typical morphosyntactic patterns; the number of shallow realizations under certain grammatical and discourse constraints is even greater.

Morphosyntactic patterns are not fully predictable by semantics (Apresjan 1967), and, hence, we can speak here about lexical constructions. The patterns with lexical slots evoked by two or more target lexemes (e.g. idiomatic phrases like *vzjal i <uexal>* 'he suddenly (lit. took and) <went away>') are also treated as LexCxs. As experiments on unsupervised LexCx retrieval have shown (Toldova et al. 2008, Lashevskaja and Mitrofanova 2009), there is a great need for an open data pool annotated manually for lexical frames. In a wider perspective, the project on tagging the form and meaning pairings is of great significance for lexical and syntactic research, lexicography, and IR tasks.

The dictionary of lexical constructions matches frames evoked by a particular target word into morphosyntactic patterns. The relevant dataset here is semantic explications (roles), lexico-semantic constrains (e.g. human, emotion, etc.), morphosyntactic constraints on the elements, their syntactic ranks.

FrameBank is an offspring project of the Russian National Corpus (http://www.ruscorpora.ru) and involves a large illustrative sample taken from the corpus. The goal of framenet-like corpus annotation is to reveal the diverse realizations of a certain LexCxs in the running text and to mark the elements that correspond to constructional arguments and adjuncts. The corpus part of FrameBank details morphological and syntactic mismatches, violation of lexical and semantic constraints, and focuses on the grammatical constructions that introduce or license the use of elements within a given construction. This is a report on work in progress, which can be followed at http://framebank.ru.

1. Background

FrameBank (Lyashevskaya 2010, Lyashevskaya and Kuznetsova 2009, http://framebank.ru) is a pilot lexicographic resource that combines a dictionary of valencies and other LexCxs with a vast collection of annotated corpus examples. While the systematic theoretical research on lexical semantics, frame semantics and lexical syntagmatics in Russian started almost as early as Fillmore's "The case for case" (1968), cf. Apresjan 1967, 1974, Mel'čuk 1974, Mel'čuk, Zholkovsky et al. 1984, there has been no appropriate large-scale data sources publicly available yet. While experimental TKS dictionary (Mel'chuk and Zholkovsky1984) and more recent "Lexicographer" database (Padučeva 2004) provide a kind of frame information for the argument structure constructions, the size of these sources cannot be considered enough since they include about one hundred entries each. Apresjan and Pall's dictionary of Russian verb valencies (Apresjan and Pall 1984) is focused on morphological patterns and does not involve frame information. An expanded and semantically informed version of this dictionary is used in ETAP-3 machine translation system and in Russian Treebank annotation (SynTagRus, cf. Apresjan et al. 2006), but this resource is not public. The Russian WordNet (RussNet, Azarova 2008) and commercial Russian Word Sketch Engine (Khokhlova 2009) which partially refer to argument structures and semantic roles are still work in progress.

The first question that arises immediately is why Russian FrameBank is not 'a yet another clone' of English FrameNet-Constructicon (Fillmore 2008, Fillmore et al. forthcoming). Although our project appears to be similar to FrameNet in many aspects, there are some crucial differences in their ideology and design. First, we do not support the hypothesis about universal frames, which should be the same in all languages. Second, due to the grammatical properties of Russian, syntax plays less important role in definition of semantic relations within a frame while it is more influenced by morphosyntactic patterns (e.g. different case and PP structures help to profile the situation differently). Thus, FrameBank is more focused on morphosyntactic patterns. Third, FrameBank follows the traditions of Moscow Semantic School including its interest to lexical constraints and semantic motivation for word co-occurrences. Forth, adjuncts are not considered a part of frames and LexCxs; rather, they are treated as forming another LexCx that joins the basic predicate-argument Cx and interacts with it. Unlike FrameNet, frames in Russian FrameBank are associated with a particular target word and not a list of semantically related words. This presupposes that even synonyms reveal (slightly) different frames, so we deal with separate frames nested under their parent rather than with one frame.

However, Russian FrameBank is close to FrameNet-Constructicon in that it is based on Construction Grammar theory. Both argument structures and idiomatic expression are treated as LexCxs with certain lexically fixed element(s) as a target and one or more variable slots. The entry for verbs, nouns, adjectives and adverbs in the dictionary look like a set of LexCxs where each construction presents a unique combination of meaning and form. Finally, both projects aim to study how constructions evolve and interact with each other.

2. Dictionary of lexical constructions

2.1. *Short list: target words and their constructions*

Each verb or other predicate word is followed by a list of LexCxs where it serves as a target word. LexCxs are grouped in frames where participants have the same semantic role. As a rule, a cluster of constructions correspond to a particular lexical meaning, thus demonstrating variability of case patterns and whether a particular frame element can be omitted. Figure 1 shows two groups of LexCxs of the verb *vystupit'* 'to step forward' which correspond to the frame of motion and the frame of coming into existence, respectively.

Target Lexeme: *vystupit'*
1. 'to step forward'
ID220. <Snom V> *Vystupilo srazu pjat' soldat* 'Five soldiers stepped forward at once'
ID221. <Snom V PR_from+S>. *Iz stroja vystupil čelovek* 'A man stepped forward from the line'
ID222. <Snom V PR_to+S> *On vystupil na seredinu komnaty* 'He stepped forward to the center of the room'
...
5. 'to appear (about blood, tears, stains, etc.)'
ID 230. <Snom V *na*.PR+Sloc> *Pjatno vystupilo na rubaške* 'A stain appeared on the short'
ID 231. <Snom V *na*.PR+Sloc *u*.PR+Sgen> *Sljozy vystupili u nee na glazax* lit. 'Tears appeared on the eyes at her'
ID 232. <Snom V *u*.PR+Sgen *ot*.PR+Sgen> *U nee ot smexa vystupili sljozy* lit. 'Tears appeared at her from laughing'
...

Figure 1. A list of LexCxs of the verb *vystupit'* 'to step forward'.

This part of a dictionary provides only brief information about constructions including ID, morphosyntactic pattern (where the target word is replaced by V) and a short example that serves as Cx name.

Idiomatic phrases are also included in a list of LexCxs; the only difference is that there is more than one target word, and all of them will be listed in the Lexical Index of target words. For example, in the lexical entry of the verb *dut'* 'to blow' a list of argument structure Cx is followed by an idiomatic Cx with the meaning 'He does not care a straw', lit. 'He does not blow through (his) moustache' (see Figure 2). This construction includes seven elements: two variables (Snom and CL, that is a Subject and a clause that express a negative situation) and five constant lexical units (the verb *dut',* the prepositional group *v us* 'through (a) mustache', the conjunction *a* 'but', the particle *i* 'even', and the negative particle *ne*). Three elements are optional and are shown in square brackets.

Target Lexeme: *dut'*
1. 'to blow (about wind etc.)'
...
2. 'to blow (about person)'
...
ID3088 <Snom V {PR_to+S/ADV}> *Dut' v trubku* 'to blow through the tube'
...
6. (idiomatic) 'not to express or reveal concern over something'
ID12934. <[CL,] [*a*] Snom [*i*] *v*.PR+*us*.Sacc *ne dujet*> *A on i v us ne dujet* 'He does not care a straw'

Figure 2. A list of LexCxs of the verb *dut'* 'to blow'.

2.2. *Passport of the construction*

Detailed information about lexical constructions is stored as a Cx template (Padučeva 2004, see Figure 3). The dictionary provides standard templates that describe arguments of the construction and their expression in neutral context not affected by other grammatical constructions. This includes:

1) Item ID and its Place within the Cx. The order of elements is quite conventional: usually Subject is followed by Object and then by syntactically peripheral elements;
2) latin Letters (X, Y, Z, etc.) help to identify elements in other LexCx; dash mark identifies predicated and other target lexemes;
3) shallow morphosyntax: POS (e.g. S(ubstantive), ADV(erb)), case, prepositional phrase and other grammatical features that constrain the element position; this information is displayed in two columns that reflects the tradition of dependency syntax (Head) and phrase structure grammar (Phrase). Target words may also have their own grammatical constraints like preferable tense, mood, etc.;
4) Explication or semantic role of the element;
5) Syntactic Rank of the element (Subject, Object, Peripheral, Clause, Adjunct, No);
7) Lexico-semantic constraints on a slot (semantic group or a list of lexemes); head of the phrase: the lemma and its semantic class;
8) Status (target word or variable; obligatory or optional).

ID230. Cx name: *Pjatno vystupilo na rubaške* [‘a stain appeared on the short’]. Cx Pattern: Snom V na + Sloc.								
Cx Item ID	Pl	Letter	Head	Phrase	Explication	Syntactic Rank	Lexico-semantic constraints	Status [obligatory / optional]
2077	1	**X**	**Snom** [Nominative case]	**NPnom**	**substance**	**Subject**	**natural object**	**Oblig.**
2078	2	**–**	***vystupit’*** [‘to appear; lit. to step forward’]	–	**to appear**	**Predicate**	**–**	**Oblig.**
2079	3	**Y**	***na* + Sloc** [preposition *na* ‘on’ + Locative case]	***na* + NPloc**	**location: surface**	**Peripheral**	**space and place**	**Oblig.**

Lexical Index of target words

Index of Morphosyntactic Items

Figure 3. The passport of the construction *Pjatno*[Noun.Nom] *vystupilo*[Verb] *na rubaške*[PREP + Noun.Loc] ‘a stain appeared on the short’.

2.3. *Frames, graphs and indexes*

A Frame Index entry includes a head word, definition, and a list of participants with explication about their semantic role in the frame. This information is copied in LexCx templates (cf. the field Explication). Sometimes the semantic role of a participant varies a bit from one LexCx to another (cf. Instrument vs. Instrument-Place); this can happen due to the effect of different frame profiling (Padučeva 2004) caused by certain lexical constructions. In this case all possible explications will be listed in the Frame Index entry.

Both frames and LexCxs are arranged in graphs with three types of relations: (i) mother-daughter, (ii) use (cf. FrameNet frame grapher) and (iii) polysemy. The hierarchical design uses the criteria of semantic derivation and comparable frame element structure nesting daughter frames under frames with more general meaning. Since frames are associated with target words, the frames of polysemous lexical units can also be related via a polysemy link.

LexCxs are linked via their target words and frames; at the same time, they can be grouped formally on the base of common morphosyntactic patterns. Of particular interest are formally identical LexCxs which represent two or more frames with a polysemy link and novel LexCxs that borrow (a part of) their morphosyntactic structure from the LexCx evoked by another target word or by another sense of the same word. For example, the LexCx *Nos.NOM sobralsja skladkami.INS* ‘(One’s) nose[Snom] gathered in pleats[Sins]’ uses the instrumental pattern of the LexCx *pojti skladkami.INS* ‘to go in pleats [Sins]’, which in turn is motivated by the manner of motion construction *pojti galopom.INS* ‘to go gallop [Sins]’ and the construction of transformation *stat’ drugom.INS* ‘to become a friend [Sins]’. The morphosyntactic pattern of the idiomatic construction *A on i v us.ACC ne dujet* ‘He does not care a straw’ <[CL,] [a] Snom [i] v.PR+us.Sacc ne dujet> uses the pattern of the lexical construction *Dut’ v trubku* ‘to blow through the tube’. The lexically fixed *v us* ‘through (a) mustache’ mirrors the PP with allative (directional) meaning (e.g. the preposition *v* ‘in’ + Accusative case) and demonstrates specification of grammatical constraints.

The dictionary of lexical constructions also includes

– an index of target words;
– a general index of lexemes (a user can consult their POS, semantic group they belong to, their definition in dictionaries of Russian);
– an index of morphosyntactic items (e.g. Snom, Sins, *v* + Sacc, etc.);
– an index of Explications.

3. Corpus annotation

Unlike standard dictionaries of valencies, FrameBank presupposes manual annotation of a sample of real uses. This is not a full-text corpus framenet annotation but rather a 'cherry-picking' approach: each target lexical unit is illustrated by 100 sentences accompanied by their pre- and post-context. The instantiations of LexCxs patterns in running texts are called 'realizations'.

The examples are drawn randomly from the Russian National Corpus (modern texts from 1950 till present, about 100 MW). The source corpus is tagged for POS, lemmas, morphosyntactic features and lexico-semantic information, and this information is stored in FrameBank database as well. It is used when we match realizations into Cx templates stored in the dictionary.

Since examples are picked up at random, the distribution of realizations over frames and LexCxs is uneven. This is done in order to get the picture of LexCxs use in real texts. The disadvantage of this approach is that some LexCxs are left without illustration. Hunting for realizations of such rare LexCxs (if any) and their annotation is a future task of the project.

Annotation of sentences is done manually by one annotator and checked by a supervisor. An annotator validates an instance against an archetype, i.e. compares an example with the model entry from the dictionary of constructions. First, examples are matched to one of the constructions associated with the target word, i. e. to a certain word sense and an appropriate argument pattern attested for this sense. Second, an annotator marks up the relevant pieces of a sentence linking them to the elements of a construction. Then she defines the marked arguments in terms of syntactic ranks, identify non-standard case marking and provide explanation about missing arguments.

Figure 4 shows the LexCx *sobrat' použinat'* 'pick up something for supper' as it is realized in example (1).

(1) *Olja!* *soberi* *nam* *poest'* *v* *dorogu.*
Olja.Noun.NOM collect.V.IMPER we.SPRO.DAT eat.V.INF in.PR way.Noun.ACC
'Olja! Pick up something to eat for us on the way.'

ID5914. Cx name: *Soberi použinat'* [‘pick up something for supper’]. Cx Pattern: Snom V Vinf.								
ID	Pl	Letter	Head	Phrase	Explication	Syntactic Rank	Lexico-semantic constraints	Status / Realization
18589	1	**X**	**Snom**	**NPnom**	**Agent**	**Subject**	**human**	**Oblig.**
20089	-	X	–	–	–	No	–	Omitted. Licensed by Imperative Cx
18590	2	**–**	***sobrat'***	–	**to collect**	**Predicate**	**–**	**Oblig.**
20090	1	–	*soberi.Vimper*	–	to collect	Predicate	–	Standard
18591	3	**Y**	**Vinf**	**VPinf**	**what is collected**	**Peripheral**	**eat**	**Oblig.**
20091	3	Y	*poest'*.Vinf	*poest' v dorogu*.VPinf	what is collected	Peripheral	eat	Standard
20089	2	Z	*nam*.SPROdat	*nam*.NPdat	Beneficiary	Peripheral	human	Added. Licensed by Ditransitive Dative Cx
20089	4	W	*v dorogu*	*v dorogu*.NPdat	Goal	Adjunct	abstract	Added
+								

Figure 4. Realization of the construction «sobrat’ poest’» (target verb sobrat’ ‘pick up’) in example (1).

The first two slots are described in the dictionary as an ‘Agent’ and ‘What is collected’, respectively. The former is a human Subject expressed by a Nominal noun (Snom), and the latter is a verb of eating in the Infinitive form (Vinf; 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. Y matches its infinitive verb phrase poest’ in a predictable way (Standard realization). It is unlikely that semantic roles will change in realizations, so Explications are just copied in the annotation templates.

In addition to that, the annotator 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 (SPROdat) *nam* ‘for us’. The last slot is treated as Adjunct and corresponds to the prepositional phrase *v dorogu* ‘on the way’. The noun in the Accusative case (Sacc) *dorogu* ‘road, way’ is used in an abstract sense here, so it is marked as ‘abstract’ in the Semantic class field.

Corpus annotations allow us to get the following new types of information:

1) the phrase matching an element of the construction (if any);
2) the head of the phrase: its lemma and semantic class;
3) elements expressed in a wider context;
4) differences in morphosyntax;
5) differences in syntactic rank;
6) differences in word order (defined against the so called ‘neutral word order’);
7) grammatical constructions that license omission or overt expression of the elements;
8) constructions that introduce additional arguments (external participants to the frame);

9) other pragmatic and information structure parameters that explain omission of the participants.

A Mismatch Index lists about twenty Cx licensing types that trigger realizations of Cx elements. The most frequent are embedding a construction into other syntactic constructions ('control') and omission of Subject if the predicate itself takes Passive Participle, Infinitive or Gerund. To put it simple, in most cases these grammatical constructions work as transformational rules that bring us from a generalized LexCx stored in the dictionary to the large variety of its surface realizations in the corpus.

4. Present state of the source and its future

At the moment, the dictionary contains 12000 LexCxs anchored by 2055 target words. For the most part, the target words are verbs and the LexCxs are of argument structure type. At the same type, the number of idiomatic Cxs rises rapidly as the body of annotated examples becomes larger.

The index of morphosyntactic items involves, among others, 59 prepositional phrase types (given they are used in more than one LexCx) and ten combinations of conjunctions and clauses. 88% LexCxs were automatically converted from other lexicographic sources, namely, morphosyntactic patterns, affiliation with target words and their senses, an example of use were extracted. After that, these constructions were tagged with frame information manually. The other 12% LexCxs are a by-product of corpus annotation. They were created as novel Cxs and annotated by hand.

The manually annotated corpus part is 28 363 sentences now (plus untagged sentences that display pre- and post-context). This number corresponds to 3250 LexCxs evoked by 500 target words. The planned size of this corpus is 100 000 example (2 MW of tagged sentences). We are planning also to include more LexCxs evoked by nouns, adjectives and adverbs both into the dictionary and corpus parts. The further step will be full-text annotation of LexCxs in this corpus.

Note

[1] The FrameBank project is supported by the Corpus Linguistics Program of the Russian Academy of Science Presidium and RFBR Foundation grant 10-06-00586a.

References

A. Dictionaries

Apresjan, J. and E. Pall 1982. *Russkij glagol – vengerskij glagol: upravlenie i sočetaemost'* [*The Russian Verb – the Hungarian Verb: Government and Combinability*]. Budapest.

Mel'chuk, I., and A. Zholkovsky 1984. *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach, Vienna.

B. Other literature

Apresjan J. 1967. *Ėksperimental'noe issledovanie semantiki russkogo glagola* [Experimental study of semantics of Russian verb]. Moscow: Nauka.

Apresjan J. 1974. *Leksičeskaja semantika* [Lexical semantics]. Moscow: Nauka.

Apresjan, J., I. Boguslavsky, B. Iomdin, L. Iomdin, A. Sannikov and V. Sizov 2006. 'A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects.' In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, 1378–1381.

Azarova, I. 2008. 'RussNet as a Computer Lexicon for Russian.' In *16th International Conference Intelligent Information Systems 2008*. Zakopane, Poland, 341–350. http://iis.ipipan.waw.pl/2008/ proceedings/iis08-33.pdf.

Fillmore, C. J. 1968. 'The Case for Case.' In E. Bach and R. T. Harms (eds.), *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1–88.

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

Fillmore, C. J., R. R. Lee-Goldman, R. Rhodes. (forthcoming). 'The FrameNet Constructicon.' In H. C Boas,. and I. A. Sag (eds.), *Sign-based Construction Grammar*. Stanford: CSLI Publications.

Goldberg, A. E. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago, IL: University of Chicago Press.

Khokhlova, M. 2009. Applying Word Sketches to Russian. In: *Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 91–99.

Kustova, G., O. Lashevskaja, E. Paducheva, and E. Rakhilina 2009. 'Verb Taxonomy: From Theoretical Lexical Semantics to Practice of Corpus Tagging.' In B. Lewandowska-Tomaszczyk and K. Dziwirek (eds.), *Studies in Cognitive Corpus Linguistics*. Frankfurt: Peter Lang, 41–56.

Lyashevskaya, O. 2010. 'Bank of Russian Constructions and Valencies.' *LREC 2010. Malta, Valletta, May 19-21, 2010.* http://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/77_Paper.pdf.

Lyashevskaya, O. and J. Kuznetsova 2009. 'Russkij FrameNet: k zadache sozdanija korpusnogo slovarja konstrukcij [Russian FrameNet: towards a corpus-based dictionary of constructions].' *Computational Linguistics and Intellectual Technologies. Proceedings of International Workshop Dialogue'2009*. Vol. 8 (15). Moscow: RGGU, 306–312.

Mel'čuk, I. 1974. *Opyt teorii lingvističeskikh modelej 'Smysl⇔Tekst': Semantika, sintaksis*. [Essay on a theory of linguistic simulations "Meaning⇔Text": Semantics, syntax]. Moscow: Nauka.

Mel'čuk, I., A. Zholkovsky, J. Apresjan et al. 1984. *Explanatory Combinatorial Dictionary of Modern Russian: Semantico-Syntactic Studies of Russian Vocabulary*. Wien: Wiener Slawistischer Almanach.

Lashevskaja, O. and O. Mitrofanova 2009. 'Disambiguation of taxonomy markers in context: Russian nouns'. *17th Nordic Conference on Computational Linguistics (NODALIDA 2009). Odense, Denmark, May 14-16, 2009*. [NEALT Proceedings Series, Vol. 4], 111–117.

Padučeva, E. 2004. *Dinamičeskie modeli v semantike leksiki* [Dynamic models in lexical semantics]. Moscow.

Toldova, S., G. Kustova and O. Lyashevskaya 2008. 'Semantičeskie fil'try dlja razrešenija mnogoznačnosti v Nacional'nom korpuse russkogo jazyka: glagoly [Semantic filters for word sense disambiguation in the Russian National Corpus: Verbs]'. *Computational Linguistics and Intellectual Technologies. Proceedings of International Workshop Dialogue'2008*. Vol. 7 (14). Moscow: RGGU, 522–529. http://www.dialog-21.ru/dialog2008/materials/pdf/81.pdf.