ОЦЕНКА РЕЗУЛЬТАТОВ ПАРСЕРА: РАСПОЗНАВАНИЕ СЕМАНТИЧЕСКИХ РОЛЕЙ УЧАСТНИКОВ ФРЕЙМОВ В ЯЗЫКЕ С ПАДЕЖНЫМ МАРКИРОВАНИЕМ

Ляшевская О. Н. (olesar@gmail.com)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия;
Институт русского языка
им. В. В. Виноградова РАН, Москва, Россия

Кашкин Е. В. (egorkashkin@rambler.ru)

Институт русского языка
им. В. В. Виноградова РАН, Москва, Россия

В статье обсуждаются подходы к оценке парсеров, задачей которых является автоматическое определение семантических ролей (semantic role labeling, SRL). Как было показано ранее, качество распознавания именованных семантических ролей в стиле FrameNet в большой степени зависит от количества выделяемых ролей и может падать, если инвентарь ролей в ресурсе, используемом для обучения, и инвентарь ролей в целевом ресурсе разнится. Наше исследование представляет первый шаг к созданию системы 'умной' оценки SRL-парсеров, которая вводила бы лингвистически мотивированные критерии оценки работы SRL-системы; позволяла бы классифицировать ошибки от незначительных до критически важных; была бы устойчива к возможным расхождениям между инвентарями ролей.

Статья описывает эксперимент, материалом для которого служит база данных FrameBank—общедоступный онлайн-ресурс, идеологически связанный с системой FrameNet и объединяющий словарь лексических конструкций частотных русских глаголов и размеченный корпус их реализаций в примерах из НКРЯ. Одним из параметров разметки аргументов конструкций служат их семантические роли, инвентарь которых в системе FrameBank устроен иерархически и представлен в форме графа. Исследуются статистические критерии дистрибуции ролей в словаре конструкций и расположение ролей на графе для того, чтобы сопоставить ответ системы и ответ золотого стандарта.

Ключевые слова: конструкции, семантические роли, полисемия, автоматическое определение семантических ролей, корпусная лингвистика, лексические ресурсы, эвалюация парсеров

EVALUATION OF FRAME-SEMANTIC ROLE LABELING IN A CASE-MARKING LANGUAGE

Lyashevskaya O. N. (olesar@gmail.com)
National Research University
Higher School of Economics, Moscow, Russia;
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

Kashkin E. V. (egorkashkin@rambler.ru)
Vinogradov Russian Language Institute
of the Russian Academy of Sciences, Moscow, Russia

The paper discusses evaluation techniques for semantic role labeling in Russian. It has been shown that the quality of FrameNet-style semantic role labeling largely depends on the quantity of roles and may decrease if the inventory of roles in the training set differs from that in the output resource. Our study is the first step towards the 'smart' evaluation tool which would introduce linguistically relevant criteria to evaluation; be able to put the mistakes on a scale from minor to critical ones; make evaluation easier in case the grid of roles varies.

We run an experiment based on the data from the Russian FrameBank, a FrameNet-oriented open access database which includes a dictionary of Russian lexical constructions and a corpus of tagged examples. The semantic role is one of the parameters that define the predicate-argument patterns in FrameBank. The inventory of roles is modeled hierarchically and forms a graph. We explore the cases when the role induced by the system and the answer of the gold standard do not match. We analyze the statistical criteria of distribution of roles in the patterns and the distance between the source and the target in the graph of roles as a mean to assess the goodness of fit.

Keywords: constructions, semantic roles, polysemy, semantic role labeling, corpus linguistics, lexical resources, evaluation

1. Background

Syntactic parsing and semantic role labeling (SRL) are two closely related tasks employed in the shallow semantic understanding of natural text. SRL focuses on the automatic identification and labeling of the relations between a predicate and its arguments which involves generalization over surface (morpho)syntactic patterns. It is a further step towards finding projections of semantic arguments in syntactic structures. With the advent of large-scale annotated data resources such as treebanks, PropBank (Palmer et al. 2005) and FrameNet (Fillmore et al. 2003), both domains have recently benefited from an enormous boost in machine learning methods. What

matters even more is the development of standard test data sets and evaluation metrics such as CoNLL 2007, CoNLL 2008 and SemEval 2007.

SRL can be divided into the following steps:

- Step 0. Target predicates (or frame-evoking words) are marked in the sentence.
- Step 1. Each target is disambiguated to a particular sense or semantic frame.
- Step 2. Words in context are classified into arguments and non-arguments; if a dependency tree is available, nodes are classified into actants and circonstants (in Tesnière 1959's tradition), i. e. 'inner arguments' and 'free modifiers'.
- Step 3a. The arguments are labeled as ARG0, ARG1, ARG2, etc. (PropBank-style SRL).
- Step 3a. The arguments are labeled with particular frame-relevant roles such as Agent, Experiencer, Stimulus, Path, etc (FrameNet-style or 'deep' SRL).

SRL is usually constrained to the target's locally expressed semantic arguments, i. e. syntactic dependencies. More advanced tasks, such as finding and resolving null instantiations from the surrounding context (Gorinsky, Ruppenhofer 2013); finding new edges introduced by the semantic structure, are currently out of the scope of industrial standards (see Das et al. forthcoming, Màrquez et al. 2008, Palmer et al. 2013 for a comprehensive overview).

Most SRL parsers stop after Step 3a. They solve the classification task for up to 10 clusters and there is an excess amount of training data on hand to reach good results. CoNLL 2008 shared task training and test data set (Surdeanu et al. 2008) provides a standard benchmark.

FrameNet-style SRL presupposes identifying targets that could evoke frames in a sentence, identifying the correct semantic frame for a target, and finally determining the arguments that fill the semantic roles of a frame. The parsers of this type use much more fine-grained structure of clusters as an input. SemEval 2007 benchmark data set (Baker et al. 2007) provides 665 labels whereas FrameNet 1.5 release has as much as 877 labels, so the optimization of verb classes and semantic roles clusters is considered helpful to overcome the sparse data problem. Other related tasks include discovering new semantic frames and roles, i.e. associating frames to 'unseen' lexical items which cannot be found either in FrameNet or in training data.

Both PropBank-style and FrameNet-style SRL tasks are language dependent. We can assume that such factors as left/right position against the predicate, voice, lexical and semantic cues, case and preposition marking, the general surface obligatoriness of arguments, would have uneven impact, depending on the language. Yet a more important factor seems to be the amount of corresponding annotation in training resources available across languages. Hajič et al. 2009 show that if a SRL tool is applied to different languages, its performance can drop by 10% (e.g. from F1≈85.5 for English to F1≈76.5 for Japanese and Spanish).

The objectives of this paper are to propose a benchmark and evaluation scenario for Russian frame-semantic role labeling. We target two well-known problems in parser evaluation: the comparability of output role labels and insufficiency of traditional performance measures (precision P, recall R, F1) in evaluation against the gold standard. The paper is organized as follows. Section 2 outlines the design and evaluation metrics. We argue that if the standard list of roles is connected into a graph, this

can help assess the SRL results which otherwise may be difficult to compare. In Section 3 we introduce FrameBank as an open resource that can be used in SRL training and/or evaluation based on Russian data. Section 4 summarizes an experiment on SRL for Russian prepositional phrases, including the structure of the data used, the rules and the qualitative analysis of the results. In Section 5 we come back to the metrics proposed earlier and show the evaluation scenario at work.

2. SRL Evaluation Metrics

The standard approach to NLP evaluation assumes that there exists a test corpus provided with a 'Gold Standard' (GS) annotation. Let G = {s^g_1, s^g_2, … s^g_N} be a set of semantic roles in the GS. Given the output from an NLP tool E = {s^e_1, s^e_2, … s^e_N}, we can compare it against the GS set and compute the number of matches M with respect to the number of answers E returned by the parser (i.e. precision P = #M / #E), the number of matches M with respect to the total number of elements G in the GS (i.e. recall R = #M / #G), and their harmonic mean F-score.

But what if a parser is either developed in a different framework or trained on a different data set, or trained on unlabeled data? RU-EVAL evaluation forum[1] has shown that many Russian parser developers rely heavily on the size and quality of their own training resource. If we project this to the domain of SRL, we can expect that the inventories of possible answers (semantic roles) in the SRL resources might vary significantly (cf. Azarowa 2008; Ermakov, Pleshko 2009; Petrova 2013; Smirnov, Shelmanov 2014; Kashkin, Lyashevskaya 2013, among others), what would make the comparison not straightforward.

Lang and Lapata (2011) suggest another set of evaluation metrics that assess an overall goodness of clustering and can work with unsupervised machine learning. Cluster purity (PU) is a measure of the degree to which the induced role clusters meet the goal of containing only instances with the same GS role label:

$$Pu = \frac{1}{n}\sum_{i=1}^{n_C} \max_{j=1,\ldots,n_C} |C_i \cap C_j|$$

where C_i is an induced role cluster (a set of answers with the same semantic role label) and G_j is the best matching GS role cluster. Cluster collocation (CO) measures how well the clustering meets the goal of clustering all gold instances with the same label into a single predicted cluster:

$$Co = \frac{1}{n}\sum_{j=1}^{n_C} \max_{i=1,\ldots,n_C} |C_i \cap C_j|$$

The harmonic mean of PU and CO is reported as F-score (Lang, Lapata 2011; Fürstenau, Rambow 2012; Titov, Klementiev 2012).

---

[1] See Toldova et al. 2012 on morphological parsing and Gareyshina et al. 2012, Lyashevskaya et al. 2010 on dependency parsing.

In this article, we consider two other types of measures, taking into account (1) the distributional properties of semantic roles over the network of frames; (2) the path between two roles in a graph. From a common sense point of view, the pairs of roles like Instrument and Means, Theme and Patient are perceived as similar whereas Addressee and Reason, Experiencer and Direction are not. Meanwhile, it is important to distinguish the roles which can occur in the same frame such as Instrument and Means, Patient and Result. If the roles are distributed complementary over the frames of the same target verb, this allows us to downgrade the matching score between the NLP answer and GS.

Since our evaluation scheme is based on FrameBank (see next section), we will use the dictionary of valencies in order to calculate the co-occurrence statistics of the induced role RoleE and the corresponding gold standard RoleG. We compute the repulsion of roles as:

$$repulsion = \frac{\#Verbs(RoleE\_RoleG\ OR\ RoleE!RoleG)}{\sqrt{\#Verbs(Role\ E) \times \#Verbs(Role\ G)}}$$

Here, the numerator is the number of verbs in the dictionary for which the roles allow us to distinguish participants in the same frame (RoleE_RoleG, i.e. they co-occur in the same pattern: …V…RoleE…RoleG); plus the rest of verbs for which the roles distinguish frames within the same verb (RoleE!RoleG, i.e. the patterns like …V…RoleE… and …V…RoleG… are in complementary distribution). The repulsion is 0 if the roles do not compete with each other and 1 otherwise.

Thus, the roles Patient and Stimulus (of perception) can hardly stand in contrast to each other and their repulsion is expected to be low. However, we can easily anticipate a system in which both roles are labeled identically (e.g. as Patient). Quite the opposite, the pair Patient—Source_location appears to be a good candidate to have high repulsion since we expect to find lots of cases like *ja sorvalsja s dereva* 'I broke from the tree' where both roles co-occur in the same frame and label different kinds of participants (RoleE_RoleG). In addition, we can expect a number of cases such as *vino brodit* 'wine is fermenting'—*kochevniki brodili s mesta na mesto* 'the nomads wondered from place to place' where the roles Patient and Source_location are attested in different frames of the verb (RoleE!RoleG).

Furthermore, we assume that semantic roles are structured data which form a graph. If so, we can calculate the distance between the nodes just as the distances between senses in WordNet are calculated (cf. the review in Budanitsky, Hirst 2006). Presumably, these metrics will help us rank non-matching results as false alerts and major discrepancies (i.e. mistakes).

3. Data: Russian FrameBank

FrameBank is an open access database (www.framebank.ru) which consists of a dictionary of Russian lexical constructions (originally based on the valency dictionary Apresjan, Pall 1982) and a corpus of their uses tagged with a FrameNet-like annotation scheme (Lyashevskaya, Kuznetsova 2009, Lyashevskaya 2010, Kashkin,

Lyashevskaya 2013). FrameBank 1.0 offline release includes constructions for ca. 1500 frequent Russian verbs provided with up to 100 annotated examples per verb. Examples are randomly taken from the Russian National Corpus (RNC, http://ruscorpora.ru).

The theoretical framework adopted in FrameBank includes Construction Grammar (Ch. Fillmore, A. Goldberg, etc.) as well as some approaches developed in the Moscow Semantic School (Ju. D. Apresjan, E. V. Paducheva et al.). Another resource which has obviously influenced the Russian FrameBank is Berkeley FrameNet (http://framenet.icsi.berkeley.edu). However, in contrast to FrameNet, the core of FrameBank is constituted by the constructions of particular lexemes rather than by generalized frames. Each construction is stored in the dictionary as a pattern followed by a mnemonic sentence label. The pattern includes:

(1) the syntactic ranks and
(2) the morphosyntactic features of the arguments (incl. case and preposition marking),
(3) the semantic roles of the arguments,
(4) the lexical-semantic classes of the participants,
(5) the morphosyntactic features of the target lexical unit itself (e.g. impersonal, passive participle, etc),
(6) one or several examples.

Figure 1 shows a sample pattern in the dictionary.

The dictionary of constructions is supplemented by corpus examples tagged manually (see Fig. 2). An example is tied to a suitable pattern, which includes establishing correspondences between their elements, assigning morphosyntactic and semantic features of the arguments in a particular example, and also marking non-standard types of use (e.g., participial or converbal constructions). Adjuncts and focus particles are also tagged but remain beyond the construction pattern. The coordinates of phrases filling the slots and their heads are calculated automatically, so we can track the position of the filler against the predicate. If an expert comes across a corpus example which does not fit any existing pattern, they are expected to add a new pattern into the database.

There are two other components in FrameBank aimed at making generalizations on how the construction network of Russian verbs is organized. These are the graph of semantic roles and the graph of lexical constructions and frames. As regards the inventory of semantic roles, its volume and structure may shrink and expand depending on a particular research task and theoretical framework (see Fillmore 1968, 1977, 1982, Dowty 1991, Apresjan 1974/1995: 125–126, Apresjan et al. 2010: 370–377, Paducheva 2004: 587–588, etc.). The most important principles governing the inventory of semantic roles in FrameBank are as follows (Kashkin, Lyashevskaya 2013):

- the inventory should be hierarchical in order to support flexible search options (it may be reduced to 5–10 basic roles, and at the same time enlarged to several dozen labels);
- the roles should correlate with the semantic classification of verbs (what follows from it is that traditionally "broad" roles such as Agent or Patient should get different labels in different semantic classes, cf. Agent in destruction vs. speech vs. motion).

	Synt_Rank	Morph.	Semantic role	Lexical class
svesti	Predicate	Vimpers		
Y	Object	Sacc	Part of subject of physiological state	body part
X	Periphery	ot + Sgen	Reason	abstract

Fig. 1. The pattern of the construction *Pal'tsy*$_Y$ *svelo*$_V$ *ot xoloda*$_V$ 'The fingers$_Y$ cramped$_V$ from the cold$_X$'

	Synt_Rank	Morph.	Semantic role	Lexical class	Alternation predicted by	Filler
X	Subject	Snom	Reason	abstract		
	Periphery	Sins	=	=	Passive participle	*prostudoj*
Svesti	Predicate	V				
	Predicate. attrib	V.partcp. pass.full. acc				*svedennyje*
Y	Object	Sacc	Part of subject of physiological state	body part		
	Agreement controller	Sacc	=	=	Passive participle shift	*pal'tsy*

Fig. 2. The annotation of the construction *Sudoroga*$_X$ *svela*$_V$ *pal'tsy*$_Y$ 'A cramp$_X$ in (lit. took down$_V$) the fingers$_Y$' in the example ... *ona podsela k pechi, svedennyje*$_V$ *prostudoj*$_X$ *pal'tsy*$_Y$ *zasovyvala v samyj ogon'—grela* '... she sat down next to the stove trying to warm at the fire her fingers$_X$ cramped$_V$ by flu$_X$'. For each element, the first line reports data from the dictionary, the second line reports annotation of the example.

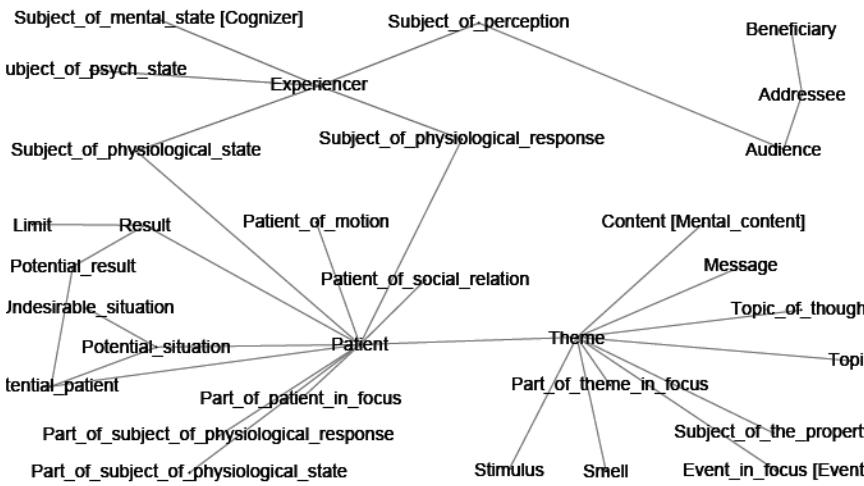

Fig. 3. A fragment of the sematic roles graph illustrating the domains of Experiencer, Addressee and Patient

The detailed list of semantic roles currently contains 96 items classified into six domains (those of Agent, Patient, Experiencer, Instrument, Addressee, Settings), which are further subdivided into smaller units. As an instance, the domain of Experiencer includes Subject of Perception ('see', 'hear'), Subject of Mental State ('think', 'understand'), Subject of Psychological State ('love', 'be afraid'), Subject of Physiological State ('feel pain', 'have a buzzing in one's ears'), and Subject of Physiological Response ('laugh', 'feel sick'). In addition, the last two roles are linked to the node of Patient whereas the subjects of Perception, Mental State and Psychological State are linked to the node of Agent. As a result, all the roles make up a united graph, see Figure 3.

The graph of lexical constructions and frames is an ongoing project; for the present it has covered 55% of the data. The graph of constructions documents the systematic relations between constructions. First, it systematizes semantic shifts in verbal lexemes (metaphor, metonymy and some more complex relations). Second, the graph represents formal changes in argument structure, such as omission of a participant, diathetic alternations, the inheritance of a pattern from another verb etc. The semantic part of the project is inspired by FrameNet grapher as well as by E. Rakhilina et al.'s research on Russian polysemous adjectives and adverbs summarized in a database (see Rakhilina et al. 2010 and references therein). The formal part is guided by E. Paducheva and G. Kustova's theoretical and empirical analysis of polysemy in Russian verbs (Paducheva 2004, see also the Lexicographer database at http://lexicograph.ruslang.ru) The frame grapher shows how lexical constructions map into the frame structure, so the graph of lexical constructions goes in parallel with the graph of frames.

4. Experiment

The next two sections will focus on how the FrameBank data can be used in SRL evaluation. The goal of our experiment was to build up a simplified SRL system adapted to use FrameBank data as a source and to do a basic evaluation in both quantitative and qualitative terms.

4.1. SRL Prototype

In order to get a data set for the evaluation we produced a list of 62 heuristics simulating a rule-based SRL tool. Unlike machine learning tools adapted to corpus data, our system works with data from the FrameBank dictionary of constructions. Our experiment focuses on semantic role labeling of four prepositional phrases: *za* + NPins, *za* + NPacc, *ot* + NPgen, *po* + NPdat. These particular PPs have been chosen since they are very frequent (e.g. ca. 900,000 hits of the PP *ot* + NPgen in the RNC) and highly polysemous. FrameBank annotations show that za + NPacc is mapped into 14 roles, such as Destination Point (*Mal'chik zabezhal za derevo* 'The boy ran behind the tree'), Motivation (*nakazat' syna za vran'je* 'to punish a son for his lies'), Price (*On kupil dom za million dollarov* 'He bought a house for a million dollars'), Period (*Eto mozhno sdelat' za chas* 'It can be done in an hour'), etc.

We produced a list of heuristics that take into account the morphosyntactic pattern of the construction, the lexical class of the PP argument, the lexical class of other arguments and the lexical class of the target predicate (the similar feature types were used for the rule-based verb sense disambiguation on RNC data, see Toldova et al. 2008).

The rules can be illustrated by the following two uses of *za* + NPacc. If the PP is added to a transitive construction, the target verb refers to the change of a possessor (e.g. *kupit'* 'to buy', *prodat'* 'to sell', *otdat'* 'to give', etc.) and the NPacc embedded into *za* + NPacc is a quantitative expression (e. g. *sto rublej* 'one hundred rubles', *bol'shaja summa* 'a large sum [of money]'), then the semantic role of *za* + NPacc is Price. However, if the class of the NPacc (within za + NPacc) is a time period (*dva dnja* 'two days', *nedelja* 'a week', etc.), then the semantic role of the PP in focus is Period.

Some rules suggest two possible outcomes, cf. two constructions: *Militsioner pobezhal za prestupnikom* 'A policeman ran after an offender', where the pragmatically correct choice is Counter-Agent, and *Mal'chik pobezhal v bol'nitsu za vrachom* 'A boy ran to hospital to call the doctor', where *za* + NPacc is more likely to describe Goal, because the doctor does not seem to escape or perform any other action here. This distinction is hard to formalize, as it requires taking into account rather vague pragmatic factors, so the rule assigns two roles with a 50% probability when the NPacc in the PP is animate.

4.2. Training and Test Data

The rules were formulated for the constructions found in the so-called 'old' part of the dictionary and evaluated against the 'new' constructions. As a training set, the constructions attested in Apresjan, Pall 1982 were taken. The constructions recently added by annotators (in order to cover RNC examples) were used as a test data set. Though this was but one of many possible folds, yet the split between the 'old' and 'new' parts was chosen as a matter of convenience since we presumed the productive patterns to prevail in the new part. Table 1 shows the distribution of training and test data set[2].

[2] Patterns with generalized locative and directional PPs were also taken into account if the use of one of the four PPs was attested in FrameBank. E.g. such patterns as [NPnom] V [za + NPacc] include those of [NPnom] V [PRkuda + NPx], where PRkuda stands for any directional preposition corresponding to Russian kuda 'where (direction)', and x denotes the NP case governed by a particular preposition.

Table 1. Type frequency (the number of constructions in the dictionary) and token frequency (the number of annotated sentences in corpus samples) of four Russian PPs: training and test set

	Training set: 'old' data		Test set: 'new' data	
PP	**# constructions**	**# examples**[3]	**# constructions**	**# examples**
za + NPins	95	80	19	22
za + NPacc	228	223	37	51
ot + NPgen	266	435	70	113
po + NPdat	311	245	65	78
Total	900	983	191	264

The results of SRL for four prepositional groups based on our rules are shown in Table 2.

Table 2. Results of the experiment

PP	Total amount of new patterns	'Strong' matching (the role is identified correctly and unambiguously	'Weak' matching (one of the answers is correct)	P_{strong}	$P_{strong+weak}$
za + NPins	19	9	7	0.47	0.84
za + NPacc	37	22	11	0.59	0.89
ot + NPgen	70	41	24	0.59	0.93
po + NPdat	65	32	25	0.49	0.88
Total	191	104	67	0.54	0.90

4.3. SRL Cues

The rules for disambiguation produce the right answers due to taking into account such 'cues' in the data as:

- The semantic class of a verb. Thus, the semantic role of *za* + NPacc has been identified correctly as Reason for an Emotion in the pattern *Beshus' za doch moju* 'I am in a rage because of my daughter', because the verb *besit'sja* 'to be in a rage' used here belongs to the class of emotions, like the verbs *bespokoit'sja* 'to worry about sth.', *bojat'sja* 'to be afraid' etc. which also occur in this syntactic pattern. Similarly, *ot* + NPgen is interpreted as Reason when combined with verbs denoting

[3] The quantity of annotated corpus examples can be less than the quantity of constructions in the dictionary since FrameBank is a project in progress and not all constructions has been tagged so far. For this reason, the rules and evaluation are based on types (i.e. constructions in the dictionary), not tokens.

physiological state, cf. the new construction patterns *V golove gudelo ot udara* 'One was feeling a buzzing in one's head due to the stroke' or *Vo rtu gorelo ot pertsa* 'One's mouth was burning from the pepper' and the old ones *Ushi zalozhilo ot vystrelov* 'One's ears were blocked due to the shots' or *Zhivot podvelo ot goloda* 'One was feeling pinched with hunger (lit.: It brought one's stomach closer due to hunger').

- The semantic class of a participant. For example, in the case of *Po radio igrala muzyka* 'There was music broadcast (lit.: played) by radio' the role of Manner has been assigned to *po* + NPdat, since NPdat in this case is in the semantic class of Communication Facilities, cf. *zvonit' po telefonu* 'to ring sb. up', *vystupat' po televizoru* 'to speak on TV', *poslat' dokumenty po pochte* 'to send documents by post', etc. The same idea works in numerous cases where a participant must be animate (like Agent or Counter-Agent), as well as in the case of the opposition between concrete and abstract entities which is relevant for quite a few examples.
- The pattern in general. Sometimes it is necessary to take into account the interaction between the elements of a construction. A curious example is represented by the verb *begat'* 'to run' and its prefixal derivative *probegát'* when used with *po* + NPdat. Normally they refer to motion events in a syntactic pattern NPnom V *po* + NPdat (*Rebenok begaet po komnate* 'A child is running in the room'). Used metaphorically, these verbs may describe perception, which may be supported by adding a NPins *glazami* 'with one's eyes' or *vzgljadom* 'with one's look' into the pattern, cf. *Ona probezhala glazami po tekstu pis'ma* 'She looked through (lit.: ran with her eyes) the text of the letter' or *Ona probezhala glazami po komnate* 'She looked through (lit.: ran with her eyes) the room'—note that in the latter case *po* + NPdat refers to the same entity (the room) as it does in the situations of motion. What is the main point here is that it is possible to omit a NPins in the contexts of perception, but a NPdat embedded into *po* + NPdat cannot denote then any kind of territory or space. Thus, *Ona probezhala po tekstu pis'ma* 'She looked through (lit.: ran) the text of the letter' is perfect, while *Ona probezhala po komnate* 'She ran along the room' is very odd as a reference to visual perception.

4.4. Challenges for SRL

The main challenges we have faced in our experiment are as follows.

First, it is difficult to deal with such cases in which there are no clear constraints on the classes of a verb and of its arguments. Thus, the use of *po* + NPdat for conveying Reason in *Rasskaz byl prochitan po ego pros'be* 'The story was read at his request' receives an additional interpretation of Information Source (yielded by the semantics of the noun *pros'ba* 'request', which belongs to the class of texts and speech acts, and of the verb *prochitat'* 'to read' dealing with information processing). This case of ambiguity stems from the vagueness of semantic restrictions imposed on *po* + NPdat as Reason, as well as on the verbs possible in this pattern, cf. *zhenit'sja po ljubvi* 'to make a love-match (lit.: to get married due to love)', *uvolit' po sokrashcheniju shtatov* 'to discharge sb. on grounds of staff reduction', *sidet' zdes' po drugomu delu* 'to stay here on some other business'.

Second, a challenge is posed by metonymic shifts of concrete nouns. For example, *ot* + NPgen in a new pattern *Ego nevozmozhno otorvat' ot knigi* 'It is impossible to divert his attention from the book (lit.: to tear him=it from the book)' is wrongly analyzed not as Content of Action, but as Patient & Location (like in *otorvat' listok ot kalendarja* 'to tear a sheet off the calendar'). The role of Content of Action presumes an abstract entity (*otorvat' ot raboty* 'to put sb. off work', *otkazat'sja ot svoih planov* 'to abandon one's plans'), while in the example in question this kind of participant is referred to by a concrete noun *kniga* 'a book' metonymically connected with an abstract entity of reading which is meant here.

There are some more similar examples in the training corpus which pose an obvious difficulty for constructing the rules. For instance, the verb *sidet'* 'to sit' combined with *za* + NPins may refer to sb's posture when the NPins denotes a concrete entity, and *za* + NPins takes the role of Location (*Papa sidit za knizhnym shkafom* 'Father is sitting behind the bookcase'), or it may yield the interpretation of *za* + NPins as Content of Action if the NPins designates an abstract entity (*Papa sidit za rabotoj* 'Father is occupied with his work (lit.: Father is sitting behind his work)'). The latter interpretation may however arise in the case of a concrete NPins making it difficult to automatically distinguish it from locative contexts, cf. the sentence *Papa sidit za knigoj* (lit.: 'Father is sitting behind the book'), which means that father is occupied with reading (as a result an obvious metonymic connection between reading and books) and has nothing to do with the expression of a spatial relationship between father and the book. Even a more challenging example is *Papa sidit za stolom* 'Father is sitting at the table', which evokes a dual interpretation (Location vs. Content of Action—e.g., eating) depending perhaps on a broader context.

It can be seen from the above that what influences the choice of a semantic role is pragmatic factors, which have proved to cause difficulties for the application of our rules. This can be illustrated by the use of *za* + NPins in the roles of Counter-Agent vs. Goal in the frames of motion when the NPacc participant is animate. Thus, a new pattern *Otets pustilsja za gigantskoj akuloj v nebol'shoj motornoj lodke* 'Father rushed after a giant shark in a small motorboat' gets two interpretations, Counter-Agent and Goal.

5. Evaluation at Work. Discussion

The experimental rule-based SRL module for four PPs yields P=0.90 in trade-off evaluation (a rule can induce more than one role as an answer, one of them is correct) and P=0.54 in strong evaluation (exact matching of an answer with the GS); recall is not applicable because we used default settings in our rules. We did not apply the formulae of purity and collocation since there was a very small number of data points to perform cluster modeling.

Table 3 summarizes 15 cases of non-matching answers. The non-matching pairs of roles were tagged manually as Good, Average and Bad match by an assessor (the contexts are provided in the table). For each case, we show the statistics on the co-occurrence of semantic roles in the valency dictionary; the shortest path from RoleE to RoleG in the graph of semantic roles. The borders of domains are marked by square brackets, ↑ marks the path up to the hypernym and ↓ marks the path down to the hyponym. The

last column shows whether the roles belong to the same domain (e.g. Agent, Patient, etc); label (Yes) indicates that the roles are from the same hyper-domain of settings (circumstances) but they belong to different domains such as Place, Time, Reason etc.

Table 3. The goodness of fit for non-matching pairs of roles: manual evaluation and descriptive statistics

Matching Evaluation (human)	Role E	Role G	#Verbs (RoleE)	#Verbs (RoleG)	#Verbs (RoleE! RoleG)	#Verbs (RoleE+ RoleG)	Repul-sion	Same domain
Good	Source	Reason	12	266	3	0	0.05	NO
[Source↑External_cause↑Agent]↑Root↓[Setting↓Reason] *Lovit' kajf* [*ot knig*] 'To be in high from books'.								
Good	Path	Patient	105	712	46	3	0.18	NO
[Path↑Location↑Setting]↑Root↓[Patient] *On bredit i mechetsja golovoj* [*po perekladine*] 'He raves, tossing his head over the crossbar'.								
Good	Property	Reason	175	266	31	5	0.17	(YES)
[Property↑Setting↓Reason] *Ego zabrali* [*po natsional'nomu piznaku*] 'He was arrested on ethnic grounds'.								
Average	Term	Time_ point	52	42	6	2	0.17	YES
[Term↓Time_point] *Vstrecha prodlilas'* [*za polnoch*] 'The meeting lasted past midnight'.								
Average	Term	Target_ location	52	398	26	0	0.18	(YES)
[Term↑Time↑Setting↓Location↓Target_location] *Emu zabralos'* [*za 50 let*] 'He was (lit. It was climbed him) over fifty years old'.								
Average	Source_ of_smell	Source_ location	5	250	2	0	0.06	NO
[Source_of_smell↑Source]↓[Resourse↑Sourse_location] [*Ot tebja*] *za verstu paxnet neprijatiem sotsialisticheskix tsennostej* 'It is evident from a mile away that you reject (lit. it smells from you) socialist values'.								
Average	Source_ location	Poten-tial_coun-ter-agent	250	6	1	0	0.03	NO
[Sourse_location↑Location↑Setting]↑Root↓[Agent↓Counter-agent↓Potential_counter-agent] *Devochki glupo prygali* [*ot nego*] *v trolleybus* 'The girls foolishly jumped from him on a trolleybus'.								
Average	Undesir-able_sit-uation	Event_ in_focus	32	309	7	0	0.07	YES
[Undesirable_situation↑Potential_situation↑Result↑Patient↓Theme↓Event_in_focus] *On uderzhalsja* [*ot sljoz*] 'He hold back the tears'.								
Bad	Patient	Source_ location	712	250	146	127	0.65	NO
[Patient]↑Root↓[Setting↓Location↓Sourse_location] *Ona otorvala glaza* [*ot knigi*] 'She raised her eyes from the book'.								

Matching Evaluation (human)	Role E	Role G	#Verbs (RoleE)	#Verbs (RoleG)	#Verbs (RoleE! RoleG)	#Verbs (RoleE+ RoleG)	Repul-sion	Same domain
Bad	Patient	Manner	712	320	175	91	0.56	NO
[Patient]↑Root↓[Instrument↓Manner] *Probejte* [*po baze dannyh*] *ego prava* 'Check his license status through the database'.								
Bad	Location	Counter-agent	519	285	117	12	0.34	NO
[Location↑Setting]↑Root↓[Agent↓Counter-agent] *Povtorit'* [*za uchitelem*] 'To repeat after the teacher'.								
Bad	Purpose	Location	169	519	84	2	0.29	NO
[Purpose↑Setting↓Location] *Deti prygali by* [*na mogile*] 'His kids would jump on the grave'.								
Bad	Informa-tion_re-source	Message	26	236	12	10	0.28	NO
[Information_resource↑Resourse↑Sourse_location↑ Location↑Setting]↑Root↓[Patient↓Theme↓Message] *Obychaj zvat' doma* [*po familii*] 'The tradition to address by the last name at home'.								
Bad	Informa-tion_re-source	Manner	26	320	9	0	0.10	NO
[Information_resource↑Resourse↑Sourse_location↑Location↑Setting]↑Root↓[Instrument↓Manner] *Izbirateli golosujut* [*po spiskam*] 'The voters take a vote through the lists'.								
Bad	Informa-tion_re-source	Cause	26	144	7	0	0.11	NO
[Information_resource↑Resourse]↑[Source↓External_cause↓Agent↓Cause] *Ja prochitala pis'mo* [*po ego pros'be*] 'I read the letter at his request'.								

The figures of repulsion correlate quite well with the split between Average and Bad matches (repulsion threshold at about .20). However, the formula fails to predict the split between Good and Average matches.

Even though all the Bad matches do not share the same domain, this factor was not sufficient to assess the quality of mismatches. We experimented with a number of approaches to employ a graph-based distance model, but neither of them succeeded to rank the data in accordance with manual assessment.

Nevertheless, the use of graph seems to be promising for the task of 'smart' SRL evaluation. In fact, the nature of mismatches in our experimental set is not necessarily the same as expected with the real SRL data since the chance of deviation from the training model is much higher. Token observations will provide a much larger number of data points and more homogeneous results.

Still, there is a lot of work to be done to explore the answers of external parsers with non-matching grids of semantic roles. *Further goal is to integrate a better representation of* the graph of semantic roles. So far, the graph uses two types of edges (IS-A and association). As a rule, the hyponyms cannot co-occur with their hypernym in the same frame (cf. Agent, Cause, Speaker), so another type of edges would be useful.

References

1. *Apresjan Ju. D.* (1995), Selected papers, Vol. 1, Lexical Semantics [Izbrannye trudy, tom I. Leksicheskaja semantika], Jazyki Russkoj Kul'tury, Vostochnaja Literatura, Moscow.
2. *Apresjan Ju. D., Boguslavskij I. M., Iomdin L. L., Sannikov V. Z.* (2010), Theoretical issues of Russian syntax: the interrelation between grammar and vocabulary [Teoreticheskie problemy russkogo sintaksisa: vzaimodejstvie grammatiki i slovarja], Jazyki slavjanskih kul'tur, Moscow.
3. *Apresjan Ju. D., Pall E.* (1982), Russian verb—Hungarian verb. Government and combinability [Russkij glagol—vengerskij glagol. Upravlenie i sochetaemost'], Tankyonvkiado, Budapest.
4. *Azarowa, I.* (2008). RussNet as a computer lexicon for Russian. Intelligent Information Systems, pp. 341–350.
5. *Baker C., Ellsworth M., Erk K.* (2007) SemEval'07 task 19: Frame semantic structure extraction, Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, June 2007, pp. 99–104.
6. *Budanitsky A., Hirst G.* (2006), Evaluating WordNet-based measures of lexical semantic relatedness, Computational Linguistics, 32 (1), pp. 13–47.
7. *Das D., Chen D., Martins A. F. T., Schneider N., Smith N.* (forthcoming), Frame-semantic parsing. Computational Linguistics.
8. *Dowty, D. R.* (1991), Thematic proto roles and argument selection, Language 67, pp. 547–619.
9. *Ermakov A. E., Pleshko V. V.* (2009), Semantic interpretation in text processing computer systems [Semanticheskaja interpretatsija v sistemah komp'juternogo analiza teksta]. Information technologies [Informatsionnye tehnologii], Vol. 6, pp. 2–7.
10. *Fillmore Ch. J.* (1968), The Case for Case, in Bach E. and Harms (Ed.), Universals in Linguistic Theory. New York, pp. 1–88.
11. *Fillmore Ch. J.* (1977), The case for case reopened, in Cole P., Sadock J. M. (eds.), Grammatical Relations, Acad. Press, New York, pp. 59–81.
12. *Fillmore Ch. J.* (1982), Frame semantics, Linguistics in the morning calm: Selected papers from the SICOL-1981, Hanship, Seoul, pp. 111–137.
13. *Fillmore, Ch. J., Johnson C. R., Petruck M. R. L.* (2003), Background to FrameNet. International Journal of Lexicography, 16(3), pp. 235–250.
14. FrameNet. An online resource, available at: http://framenet.icsi.berkeley.edu.
15. *Fürstenau H., Rambow O.* (2012), Unsupervised induction of a syntax-semantics lexicon using iterative refinement, in Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).
16. *Gareyshina, Anastasia, Maxim Ionov, Olga Lyashevskaya, Dmitry Privoznov, Elena Sokolova, Svetlana Toldova.* (2012). RU-EVAL-2012: Evaluating dependency parsers for Russian. In: Proceedings of COLING 2012, Mumbai, December 2012: Posters. Pp. 349–360.
17. *Gorinski, P., Ruppenhofer, J., Sporleder, C.* (2013), Towards weakly supervised resolution of null instantiations. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013). Long Papers, pp. 119–130. http://aclweb.org/anthology//W/W13/W13-0111.pdf

18. *Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L.* et al. (2009), The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–18. http://aclweb.org/anthology//W/W09/W09-12.pdf
19. *Kashkin E. V., Lyashevskaya O. N.* (2013), Semantic roles and construction net in Russian FrameBank [Semanticheskie roli i set' konstrukcij v sisteme FrameBank], Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 12–1, pp. 297–311.
20. *Kuznetsov I.* (2013), Semantic role labeling system for Russian language, in: Joho H., Ignatov D. (eds.), ECIR 2013 Doctoral Consortium, 24 March 2013, Moscow, pp. 15–18. http://www.hse.ru/pubs/lib/data/access/ticket/1391119334ea59778f895f58081f821387ee54322f/text_DC.pdf
21. *Lang J., Lapata M.* (2011), Unsupervised semantic role induction with graph partitioning, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1320–1331.
22. *Lyashevskaya O.* (2010), Bank of Russian Constructions and Valencies, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, pp. 1802–1805.
23. *Lyashevskaya O., Astaf'eva I., Bonch-Osmolovskaya A.* et al. (2010), NLP evaluation: Russian morphological parsers [Otsenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka], Computational linguistics and intellectual technologies. Proceedings of International Conference "Dialog", Vol. 9 (16), pp. 318–326.
24. *Lyashevskaya O. N., Kuznetsova, Ju. L.* (2009), Russian FrameNet: constructing a corpus-based dictionary of constructions [Russkij FrejmNet: k zadache sozdanija korpusnogo slovarja konstruktsij], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog", Vol. 8 (15), pp. 306–312.
25. *Màrquez L., Carreras X., Litkowski K. C., Stevenson S.* (2008), Semantic role labeling: an introduction to the special issue, Computational Linguistics, Vol. 34 (2), pp. 145–159.
26. *Paducheva E. V.* (2004), Dynamic patterns in lexical semantics [Dinamicheskie modeli v semantike leksiki], Jazyki slavjanskoj kul'tury, Moscow.
27. *Palmer, M. S., Gildea D., Kingsbury P.* (2005), The proposition bank: an annotated corpus of semantic roles. Computational Linguistics, 31(1), pp. 71–106. www.cs.rochester.edu/~gildea/palmer-propbank-cl.pdf.
28. *Palmer M. S., Wu Sh., Titov I.* (2013), Semantic Role Labeling Tutorial. NAACL 2013 tutorials. Electronic access: http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-1-naacl-2013-tutorial.pdf, http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-2-naacl-2013-tutorial.pdf, http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-3-naacl-2013-tutorial.pdf
29. *Petrova M.* (2013). The Compreno Semantic Model: The Universality Problem. International Journal of Lexicography, 26 (4).

30. *Rakhilina E. V., Reznikova T. I., Karpova O. S.* (2010), Semantic shifts in attributive constructions: metaphor, metonymy, and rebranding [Semanticheskie perehody v atribultivnyh konstruktsijah: metafora, metonimija i rebrending], in Linguistics of Constructions [Lingvistika konstrukcij], Azbukovnik, Moscow, pp. 398–455.
31. *Smirnov I. V., Shelmanov A. O., Kuznetsova E. S. and Hramoin I. V.* (2014), Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov II. Metod semantiko-sintaksicheskogo analiza tekstov]. Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], Vol. 1, pp. 95–108.
32. *Surdeanu M., Johansson R., Meyers A., Màrquez L., Nivre J.* (2008), The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies, Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008). Manchester, England, August 2008, pp. 159–177. http://aclweb.org/anthology//W/W08/W08-2121.pdf
33. *Titov I., Klementiev A.* (2012), Semi-supervised semantic role labeling: approaching from an unsupervised perspective, in Proceedings of COLING 2012: Technical Papers, pp. 2635–2652.
34. *Toldova S. Ju., Kustova G. I., Lashevskaja O. N.* (2008), Semanticheskie fil'try dlja razreshenija mnogoznachnosti v Natsional'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 Conference “Dialog”. Vol. 7 (14), pp. 522–529.
35. *Toldova S. Ju., Sokolova E. G., Astaf'eva I., Gareyshina A., Koroleva A., Privoznov D., Sidorova E., Tupikina L., Lyashevskaya O. N.* (2012). NLP evaluation 2011–2012: Russian syntactic parsers [Otsenka metodov avtomaticheskogo analiza teksta 2011–2012: Sintaksicheskie parsery russkogo jazyka]. Computational linguistics and intellectual technologies. Proceedings of International Conference “Dialog”, Vol. 11 (18), pp. 797–809.