Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5995–6002
Marseille, 11–16 May 2020

Russian PropBank

Sarah Moeller, Irina Wagner, Martha Palmer, Kathryn Conger, Skatje Myers
University of Colorado Boulder
{first.last}@colorado.edu

Abstract

This paper presents a proposition bank for Russian (RuPB), a resource for semantic role labeling (SRL). The motivating goal for this resource is to automatically project semantic role labels from English to Russian. This paper describes frame creation strategies, coverage, and the process of sense disambiguation. It discusses language-specific issues that complicated the process of building the PropBank and how these challenges were exploited as language-internal guidance for consistency and coherence.

Keywords: lexicon, lexical database, corpus (creation, annotation, etc.), semantics

1. Introduction

This paper presents a proposition bank for Russian (RuPB) that balances parallelism with the English PropBank against guidance from linguistic properties specific to Russian. A proposition bank, or PropBank, is a lexical resource that follows the PropBank scheme (Palmer et al., 2005) to provide consistent labeling of semantic roles for large corpora. Semantic Role Labeling (SRL) provides consistent semantic information for natural language processing at a level appropriate for statistical machine learning of semantic relations. Data annotated with proposition bank labels supports the training of automatic SRL which improves question answering (Zapirain et al., 2013), information extraction (Moreda et al., 2005), and textual entailments (Sammons et al., 2010), and statistical machine translation (Bazrafshan and Gildea, 2013). Semantic roles communicate "Who does What to Whom and How and When and Where?" They may appear in various positions in the sentence, as in the examples below where "John" is always the semantic agent, but appears in (1) as the subject noun phrase, in (2) as the direct object in a passive transitive clause, and in (3) as the object of the final prepositional phrase of a passive ditransitive clause. Despite these syntactic alternations, in all three sentences the semantic role of "John" as the Agent of the hitting event never changes.

(1) John hit the nail with the hammer.

(2) The nail was hit by John.

(3) The nail was hit with a hammer by John.

It is important that natural language processing models works with multiple languages and language structures. Systems that have been evaluated on multiple languages are more likely to generalize well to new languages (Cohen, 2020). The initial goal for developing a proposition bank for Russian was to support alignment of English and Russian predicates and automatic projection of English semantic roles to Russian texts. This is a complicated task because Russian is more complex morphologically and exhibits more flexible word order than English or other Indo-European languages currently modeled by proposition banks. This project tests the portability of the PropBank scheme. During development of the the Russian PropBank (RuPB), language-specific issues presented unique challenges. Yet, an in-depth linguistic understanding of these issues also provided guidance towards a more consistent and appropriate representation of the semantic structure of Russian verbs.

Russian is spoken by 150 million people across twelve time zones as a first language, in most of the former Soviet Union as a second language, and is one of the official languages of the United Nations. It is written with the Cyrillic alphabet which is used by other languages such as Bulgarian, Ukrainian, and Kazakh. With respect to available related resources, there is a rich lexicon of verb structure in Czech, another Slavic language with many similarities in morphosyntax, but the only Russian semantic-syntactic resource available today relies as much on syntactic as semantic structure.

The following sections describe the process of developing RuPB. Related resources that supported the process are introduced in Section 2. The creation and subdivision of a predicate's rolesets, as well as the rationale for inclusion of derivations as predicates or aliases of another predicate, are explained in Section 3. Language-specific issues that both complicated and enriched the process and the language internal guidance that they provided when creating the PropBank are described in Section 4. RuPB's coverage is described in Section 5. We conclude with some notes on Inter-Annotator Agreement and plans for future work in Sections 6. and 7.

2. Related Work

The Russian PropBank provides a unique representation of the semantic information of Russian predicates. This section surveys resources that guided the representation choices made in RuPB: a similar resource for Czech and another for Russian. It also looks at the English PropBank, with which RuPB maintains parallels, and proposition banks in other languages.

2.1. English PropBank

The English PropBank provides annotation coverage for verbal, nominal, and adjectival predicates. The original English PropBank established semantic annotation on top of the Penn Treebank dependency parses of the Wall Street

Journal. It established the PropBank scheme that creates a semantic frame for the various senses of a given predicate, defining the semantic roles on a predicate-by-predicate basis (Palmer et al., 2005). Each sense of a polysemous verb has a different frame, called a roleset, based on its specific semantics. Rolesets can be considered coarse-grained sense distinctions. All arguments of a given word sense are assigned an argument structure. Arguments are numbered from 0-6 so as to be theory neutral but take the diathesis alternations in Levin (1993) and Dowty's (Dowty, 1991) proto-roles into account, with Arg0 (argument 0) corresponding to the most agentive argument and Arg1 corresponding to the most patient-like argument (e.g., patient, theme). Arguments 2-6 are more diverse but include benefactive, recipient, instrument, attribute, end state, start point, direction, and verb specific roles such as "cost" in a purchase scenario. An example frame for the sense of "hit" meaning 'to strike' appears below. Arg0 is the proto-agent or the most agentive argument. Arg1 corresponds to the most patient-like argument, called the proto-patient. Finally, Arg2 corresponds to the modifier role of instrument.

(4) hit.01
Arg0: Hitter
Arg1: Thing Hit
Arg2: Hit with

RuPB maps compatible verb senses and semantic roles between the English and Russian PropBanks. This mapping will support automatic role projection from English to the Russian rolesets, allow for easier comparison between the two resources, and facilitate frame creation. RuPB follows the pattern of the English PropBank as closely as possible; at the same time, it does not slavishly follow the English resource. This allowed RuPB to better address the semantic structure of Russian verbs. For example, if we were to follow English PropBank strictly, the verb видеть 'to see' would have lost an additional meaning, which is 'to perceive, or to hear.' Moreover, Russian verbs provide intrinsic semantic information that English verbs do not. This information ensures data-driven decisions about divisions of rolesets which would have been lost otherwise.

2.2. Non-English PropBanks

RuPB joins the collection of PropBanks for languages other than English. A collection which covers at least six languages from six language families. PropBanks have been built for Chinese (Xue, 2006), Korean (Palmer et al., 2006), Arabic (Zaghouani et al., 2010), Hindi (Vaidya et al., 2013 06), Portuguese (Duran and Aluísio, 2012), Finnish (Haverinen et al., 2014 09), and Turkish (Şahin and Adalı, 2018). Language-specific issues differ among the various languages. For example, like Russian, the Korean PropBank handles free word order and complex morphology. In Arabic, in contrast with Russian, non-verbal predicates play a prominent role in expressing eventive predicates, requiring an early inclusion of light verb constructions. However, Arabic has verbal morphological alternations that is unexpectedly similar to Russian in its effect. The morphological process produces a verb form which can act as a passive or middle voice counterpart to the base form, but can also express other, unpredictable meanings (Bonial et al., 2017). Like the RuPB, the Arabic PropBank deals with the morphological variation on a individual basis and gives alternate forms a separate frame whenever they express something other than passive or middle voice.

At least one method has been developed to automatically generate PropBanks (Akbik et al., 2015). This method first uses a parallel corpus where the source language is annotated with PropBank role labels and the target language is syntactically annotated. High-confidence semantic roles are filtered and used to project source language roles onto the target language, resulting in high precision. In a second stage, a subset of target language labeled sentences train a classifier that adds new labels in order to increase recall. This approach resulted in about a 72% F1 score of correct labels for Russian.

2.3. FrameBank

The Russian FrameBank (Lyashevskaya and Kashkin, 2015) is a digital lexical resource oriented towards FrameNet (Baker et al., 1998) that also includes detailed syntactic information similarly to VerbNet (Kipper et al., 2008 03). It divides a given verb into fine-grained senses. FrameBank then subdivides the verb senses into its syntactic realizations, supporting each one with example sentences attested in the Russian National Corpus (Apresyan et al., 2003).

RuPB relates each of its Roleset to the coarse division of verb senses in FrameBank for easy reference. However, unlike FrameBank, RuPB does not subdivide frames by syntactic structure. It also does not make fine-grained sense divisions where the senses are not significantly different in meaning and the same semantic roles are preserved. This means there is often a one-to-many mapping between RuPB and FrameBank. The Russian PropBank uses the same general argument labels as other PropBanks to maintain consistency between PropBanks.

At least one attempts to perform SRL with neural networks used FrameBank parsing. The results was a 82% F1 score. This provides an openly available benchmark for evaluating automatic SRL in Russian (Shelmanov and Devyatkin, 2017).

2.4. Czech Vallex

The Czech Vallex (Kettnerová et al., 2012) is a well-developed example of an enriched lexical resource for Slavic languages. The Vallex contains a collection of annotated data that maps the argument structure of Czech verbs. Its main goal is to render a consistent dictionary of the Czech verb structure for NLP applications. It allows for a verb to be mapped to other resources, such as FrameNet or VerbNet, on the basis of the selected core arguments. Should the other resource not include an argument as either core or non-core, the Vallex may be required to either exclude arguments that otherwise fit its frame or include arguments that do not fit.

The Vallex provides the syntactic valency structures of the most frequent Czech verbs and their senses. It includes information such as number and optionality of arguments.

Unlike PropBank, only those arguments that are deemed core syntactic arguments exhibit such information. Some morphological characteristics, such as aspectual marking are included, but all the aspectual counterparts of a verb are not necessarily included in the resource. Each sense of a verb is listed with definitions and examples of usage. The Vallex covers idiomatic constructions, characteristics of control, reflexivity, reciprocity, and the verbs' syntactico-semantic class. The latest version added light verb constructions.

3. The Russian PropBank

In designing RuPB, several crucial considerations were made. Since the main goal of the project was to create a resource based on the semantic distinctions within the language and, at the same time allow for a cross-resource integration, the identification of core arguments of a predicate was an initial step. In this theoretical model, frequent and core arguments create a possible roleset for a singe meaning of the predicate. For example, the verb быть (byt') 'to be' has seven rolesets due to the seven possible meanings of the verb, including two idiomatic senses, listed below. Although the rolesets were based on English rolesets where possible, the creation of new Russian rolesets was determined by the sense of the predicate rather than the previously developed framework for English. In the example below, senses 06 and 07 have no equivalents in English and thus they required new rolesets.

(5)
- быть.01 'copula' (be.01[1])
- быть.02 'have, possess' (have.03)
- быть.03 'want to consume' (eat.01)
- быть.04 'be, exist' (exist.01)
- быть.05 'to do, to be or not to be' (do.01)
- стало_быть.06 (lit. 'became_be') 'therefore, consequently'
- была_не_была.07 (lit. was_not_was) 'let's risk it!'

The criteria for dividing rolesets are primarily semantic; however, some syntactic considerations were also used. Often a new roleset is created when an otherwise dubiously distinct sense has a distinct argument structure. More often, if two senses have one argument structure but require different semantic roles, a new roleset is created in the Frame File, as illustrated below with the verb стать (stat'). Unlike FrameBank, metaphoric or figurative senses of a verb are collapsed in RuPB if they do not require different semantic roles.

(6) стать.01
'become' (e.g., стать грустно "become sad")
Arg1: entity changing state
Arg2: new state

(7) стать.04
'stand' (e.g., стать к лесу задом "stand with one's back to the forest")
Arg1: thing standing
Arg2: location, position

The process for developing new rolesets takes the following steps, working through the data one predicate at a time. First, annotators examine supporting resources, particularly FrameBank, to determine the predicate's semantic range and syntactic structure.[2] Second, they draft rolesets based on the verb's semantic-syntactic interaction. Semantic roles include either very frequent arguments or those arguments that are necessary to complete the semantics of the verb. Frequency is determined by examining several sentences in the data or from the Russian National Corpus. Third, to exemplify the use of the rolesets, annotators chose sentences from the data when possible. In such cases when data had no examples for some verb senses and their rolesets, annotators suggested their own grammatical sentences or found appropriate examples in the Russian National Corpus. Fourth, example sentences are annotated with the roleset that was created for the sense they illustrate.

Two native or near native speakers consulted with each other as they constructed the rolesets. A third native speaker linguist made a final check of the drafts. Annotators strove to maintain consistent mapping with the English PropBank by choosing semantic roles from English predicates similar in meaning to the Roleset under construction. The definition and the descriptions of the semantic roles are given in English, similarly to the Arabic PropBank. However, the annotators did not directly follow or translate from the English PropBank. The number and definition of the roles were adjusted wherever necessary to fit the semantic and syntactic structure of the Russian verb.

Each roleset includes the chosen predicate with an identifying number, a definition in English, its semantic roles, and annotated example sentences, as well as notes that point to the parallel English roleset and notes regarding any deviation from the English template. A verbs' derived nouns, some reflexive forms and canonical aspectual derived forms are listed as a semantic aliases. An alias in PropBank is an alternative word or form of a word (e.g. adjectival participials) treated as different realizations of the same semantic concept (e.g, for the verb закончить (zakonchit') 'complete,' aliases include окончить, кончить, кончать, кончавший, оконченный, (okonchit', konchit', konchat', konchavshiy, okonchenniy)). The reflexive form (the "-ся" (s'a) form) of a verb is also treated as an alias if, and only if, the two forms do not differ in the qualities described in section 4.

Frame Files include idiomatic phrases but not light verb constructions. We define idioms as any phrase that includes the predicate in question but the meaning of which does not easily decompose into its individual words. New rolesets were constructed for idioms when the meaning of the predicate inside the phrase does not match any existing roleset, or when its meaning does fit another roleset but requires

[1] The rolesets given in parentheses are the parallel roleset in the English PropBank. Where no roleset is given, no cross-linguistic roleset mapping between English and Russian was found.

[2] Many thanks to our annotators: Oksana Melnyk and Alexandra Romanova.

different semantic roles. For example, the phrase сыграть в ящик (sygrat' v yashchik) literally means 'to play a box' but it has the same meaning as an English idiom "kick the bucket." This idiom was given its own roleset since the meaning of сыграть (sygrat'), literally 'play', differs significantly in this particular usage from any of the verb's other rolesets and requires different semantic roles.

(8) сыграть_в_ящик.06
'die, kick the bucket'
Arg1: the deceased

4. Language-specific Issues

When developing a proposition bank, constraints that are unique to the language shape the approach and provide an understanding of the theoretical shortcomings. While these constraints usually challenge the completion of the project, for RuPB, language-internal complexities guided the construction of a more coherent PropBank. This section describes three significant issues specific to Russian and how the RuPB developers exploited them to their advantage.

As mentioned earlier, the rich verbal morphology of Russian often presents unique challenges. One issue that arose during the development of RuPB concerned the organization of the predicates into Frames. In English PropBank, when verbs use affixation to derive new words (e.g., appropriate - misappropriate), each word is organized into its own Frame Files.[3] This organization follows standard English dictionaries, in which each word is a headword in its own entry. However, in a Russian dictionary, verbs may be cross-referenced to other verbs if they have close morphological and semantic relations to each other. These relations are created by affixation which changes grammatical aspect, nominalizes the verb, and switches its reflexivity. By considering the complex derivational morphology of Russian during the development of RuPB uncovered a mismatch between verbs and the deverbal noun forms (nominalized verbs). Moreover, it elucidated a wide variation of grammatical aspect that sometimes combined with changes in lexical meaning, as well as variations in semantic-syntactic functions that occurred when verbs switched their reflexivity. Such variation is not found in English and, therefore, could not be leveraged in the development of the English PropBank. However, in the development of future proposition banks, these strategies could be useful for languages that exhibit similar derivational processes on verbs, such as other Slavic languages.

Like other Slavic languages, including Czech, Russian exhibits rich grammatical aspectual range. All but a few Russian verbs are either perfective or imperfective, and the grammatical aspectual distinctions are manifested at a lexical level. No single method exists for deriving perfective and imperfective verbs from each other. The imperfective form is usually the closest to a morphological "base" form. Perfective verbs are often formed by adding a prefix to the imperfective form, but they can also be created by other affixation processes or by altering the root morpheme. Verbs typically form a part of a canonical pair, consisting of one imperfective and one perfective form, each with the same lexical meaning, differing only in perfective and imperfective aspects. As illustrated below with the base form кусать (kusat'), many predicates have more than one perfective form, expressing different aspectual meanings. The range of meanings in perfective forms includes inchoative, temporary duration, non-/iterative, distributive, and more. Some perfective forms may act as derivative forms that alter the lexical meaning as well.

kusat'	"bite"	imperfective
perekusat'	"bite all over"	perfective
perekusyvat'	"bite in two, snack"	imperfective & habitual
perekusit'	"snack"	perfective
perekusnut'	"quickly snack"	perfective & punctual
zaperekusat'	"begin biting all over"	perfective & inchoative

Many Russian verbs have both a reflexive and non-reflexive form, although several have only one of them. Russian reflexive verbs are formed by adding the suffix -ся to the non-reflexive verb form. Historically, the suffix derives from the reflexive pronoun. Similarly to Czech, the difference in meaning between the forms is not limited to reflexivity. Reflexive verbs act not only as reflexives, but also express reciprocal actions, allow unspecified object deletion, and function as passive voice. They nearly always result in a different syntactic structure, usually reducing valency, but some also alter the lexical meaning of the predicate. For example, the reflexive form of the verb кусать (kusat') 'to bite' becomes кусаться (kusat's'a) 'to bite (everyone) around'.

Different or additional lexical senses of the aspectual pairs or (non) reflexive forms were given their own rolesets. FrameBank and dictionaries assisted in the determination of what the "default" imperfective or perfective pair of each verb was and whether the possible aspectual variations should be unified under one roleset or not. They also helped decide when a reflexive form acts as passive or middle voice or possesses distinct senses that demand a separate roleset.

A final challenge in RuPB development is the handling of deverbal, or derivative, nouns. These nouns are formed from a verb via morphophonological processes. English employs processes such as umlauting, stress changes, or affixation (including zero affixation) to derive a noun from a verb (e.g. pr[a]céss (v.) - pr[ó]cess (n.), walk (v.) - walk (n.), educate (v.) - education (n.)). Russian employs affixation to turn verbs into nouns. For example, взять (vz'at'; v. 'take') - взятие (vz'atiye; n. 'taking, conquering'), читать (chitat'; v. 'read') - читание (chitaniye; n. 'reading'). In RuPB, the decision on their inclusion was most challenging.

At the initial stages of the project rolesets were created for the verbs only and it was not clear how the inclusion of such forms would affect the organization and the theoretical framework. This project takes a conservative approach in the development of the rolesets, that is, creating the smallest number of the rolesets necessary to capture the essence

[3] Phrasal verbs are added to the Frame File of its verbal element.

of the predicate's semantic distribution of arguments. Ultimately, this conservative approach, along with the careful adherence to the language-specific challenges, the RuPB enhances the accuracy of the semantic model.

4.1. Grammatical and Morphological Aspect

Where they exist, the canonical im-/perfective pair are found in the roleset as aliases or alternative forms of the predicate. In most cases, the contrast between the two forms is the canonical difference between the progressive or continuous sense of imperfective aspect versus the completive sense of the perfective. However, Russian grammatical aspect interacts with a predicate's lexical aspect (Aktionsart) in unpredictable ways. Occasionally, the "aliasing" of a canonical aspectual pair is only possible with certain senses of the predicate. Sometimes one sense requires a different for as its canonical match or simply does not occur in the other aspect. This was taken as an indication that a new roleset should be created.

For example, the pairing of утверждать/утвердить (utverzhdat'/utverdit') is not acceptable for every sense, as shown below. The perfective and imperfective forms each have one sense that does not pair with another aspectual form. Only the roleset утверждать.02 allows both perfective and imperfective forms.

(9) утверждать.01
'insist, claim, assert'
NO PERFECTIVE

(10) утверждать.02
'confirm, appoint, establish'
ALIASES:

(11) утвердить.03
'reinforce, strengthen, maintain'
NO IMPERFECTIVE

Many Russian verbs have multiple possible perfective forms. These forms add a range of aspectual information, and sometimes lexical meaning as well, to the verb. For some predicates new imperfective forms can be formed from the perfective forms. Chains of imperfective-perfective-imperfective morphological alternations, where each altered form adds some additional meaning, are frequent in the language. However, due to the time and resources limitations of the project, these full chains are generally not included in the rolesets. Although some additional perfective forms are added as aliases when the changes in meaning are limited to inchoative as in (12), iterative as in (13), or durative as in (14).

(12) иметь (imet') 'have' *vs.*
заиметь (zaimet') 'obtain.'

(13) терять (ter'at') 'lose' *vs.*
перетерять (pereter'at') 'lose one after the other.'

(14) играть (igrat') 'play' *vs.*
поиграть (poigrat') 'play for a bit.'

Other morphological alternations that strongly affect the lexical meaning are generally not included as rolesets unless that specific form was attested in the data. However, forms that do not add additional arguments make an exception of that general rule. For example, verbs with derived forms that essentially express negation, as in (15) were added to the Frame Files, requiring their own rolesets.

(15) хотеть (khotet') 'want' *vs.*
расхотеть (raskhotet') 'stop wanting'

Overall, aspectual variations required certain flexibility on the part of the RuPB. The decisions to collapse some aspectual forms under one roleset but to separate others were made based on the semantic and syntactic characteristics of the lemma, which were analyzed by the annotators in multiple contexts.

4.2. Reflexive Verbs

The semantic structure and usage of reflexive forms determines how they are handled in RuPB. The range in meaning of the reflexive suffix "-ся/-s'a" reaches beyond reflexivity. The reflexive form may function as a passive voice, promoting the Proto-Patient role to subject, as a middle voice, and, of course, as the canonical reflexive. In some cases, the reflexive form may so distinct in the meaning that it the commonalities with the non-reflexive form may not be immediately clear, as in 16.

(16) смываться (smyvat's'a) 'sneak away' *vs.*
смывать (smyvat') 'wash off'

In the latter case, the RuPB treats the reflexive form as a separate predicate with its own roleset. In other cases, the reflexive form is treated as an alias of the non-reflexive form, as long as its semantic roles are the same. For example, the reflexive form смываться (smyvat's'a) can function as a passive voice for two senses of the non-reflexive form смывать (smyvat'): 'wash off' and 'move by liquid'. However, it also has a distinct sense of its own ('sneak out/away') that does not allow a non-reflexive form. As a result, an additional roleset was required to capture that sense. In contrast, the Czech Vallex uses morphological form and syntactic structure to decide whether to handle reflexive forms as variants of the non-reflexive form or as separate lemmas.

4.3. Deverbal Nouns

Unlike the English PropBank, which covers some eventive nouns and predicate adjectives, the Russian PropBank currently focuses on verbs. Deverbal nouns are included in a roleset as aliases when the sense of a deverbal noun does not equate with the verb's fine-grained sense range. Including these nouns demonstrates how RuPB used language-internal guidance for separating rolesets. It also provides a template for a unified representation of Russian nominals and verbs.

Some Russian deverbal noun forms have a different range of meaning than the verb form it derives from. For example, the deverbal noun взятие (vz'atiye) matches in sense and usage to sense 'gain control of, achieve, conquer, possess, arrest, rape' but is not as acceptable for the sense 'take,

hold.' When a deverbal noun pertains only to certain senses of the verb, the noun is included as an alias to help disambiguation the rolesets. This occasional mismatch between the sense distinctions of the verb and its derived noun also informed decisions about when to create a new roleset. For example, the fact that the sense of the noun only corresponds to one subset of the verb's range of meaning supported the decision to divide the verb взять (vz'at') into its first two rolesets, as shown below.

(17) взять.01
'take, hold'
ALIASES: взяться, брать, браться

(18) взять.02
'gain control of, achieve, conquer, possess, arrest, rape'
ALIASES: брать, взяться, взятие

Ultimately, the RuPB will include eventive nouns. The approach shown here demonstrates how attention to the language constraints offers an elegant solution for standardizing the otherwise difficult determination of how to subdivide coarse-grained senses. The English PropBank includes verbs and their nominalization but it does not employ a difference in the range of senses between deverbal nouns and verbs as a guide to determining the creation of rolesets.[4]

5. Data

All the verbs in RuPB were automatically extracted from the Russian Language Pack of the LORELEI (Low Resource Languages for Emergent Incidents) data (Strassel and Tracey, 2016). The subset of the data used included two genres: newswire (91 sentences / 2,228 tokens) and phrasebook (496 sentences / 2471 tokens). Rolesets for the most frequent verbs in the data were developed first. Additional verbs are being added which widens the coverage to include verbs that are common in genres other than news; all these verbs occur at least once in the data. The RuPB includes rolesets for approximately 60% of the 500 most frequently used verbs in the Russian language (Sharoff, 2002) across genres. Example sentences in each roleset are primarily drawn from the LORELEI data or the Russian National Corpus.

A primary goal for building PropBanks is to develop an annotated corpus to train machine learning systems, meaning that typically only the rolesets needed for immediate annotation are created. The RuPB also began with corpus data annotated with syntactic information, but, in part thanks to Framebank, its coverage of a given verb's senses reaches beyond the the needs of its initial annotation goal. RuPB provides rolesets for each predicate that nearly complete its full range of senses. At this moment, RuPB does not include light verb nor the archaic forms of verbs. In some cases, rolesets for more well-known archaic senses are covered; so are any non-standard usages (slang, conversational, colloquialisms) that were familiar to the annotators. Common non-standard orthographic forms are included as aliases. Derived forms were not extracted from the data, nor are they typically included in the RuPB rolesets with the exception of those discussed in Section 4.

[4]The English FrameNet, on the other hand, does make distinctions between some nominalizations and their verb forms based on differences in semantics. For example, "observe" and "observance" have separate frames.

6. Inter-Annotator Agreement

To ensure consistency and standardization of the RuPB annotation framework, the previously described subset of the LORELEI data, consisting of newswire and phrasebook text, was double annotated with the rolesets included in this iteration of RuPB. A pilot development created rolesets for the most frequent verbs in the data. Then the data was annotated with RuPB role labels. Annotators tagged occurrences of predicates and their arguments independently of each other. If an annotator felt that a roleset was unsuitable or a sense or role was unaccounted for, they discussed the issues and modified the rolesets as necessary.

A second round of annotation was completed after the first. Overall agreement was calculated in two ways. The more conservative IAA computes "exact" match. This considers two annotations to be in agreement only if both the chosen word and the chosen argument label are the same. The second, "partial" score is computed if Arg0 and Arg1 labels are the same but higher numbered arguments differ, essentially dealing only with proto-agents and proto-patients and treating all other arguments the same.

In the more conservative exact calculation, the annotation agreement was 81.1 and 74.5 F1-score for Phrasebook and Newswire respectively. The more generous partial agreement was 88.9 and 82.5 respectively.

Annotators had very high agreement for identification of which word in the sentence was the predicate on both the Phrasebook and Newswire datasets (99.5 and 98.8 F1-score respectively). Within identified predicates, overall agreement about rolesets (i.e. sense of a verb) was high, though agreement on Phrasebook (89.0) was lower than Newswire (95.8). If we take into account the entirety of predicates, including rolesets, agreement was 88.6 and 94.6 respectively.

In the Phrasebook dataset, a significant source of disagreement was consistently different choice of rolesets for the lemmas быть ('to be/to have/to exist') and мочь ('to be able/can'). The second verb appears in the data where the Russian verb was translated from the English wherever the English text apparently used "can" to mean "may/might". It appears one annotator was more attuned to this "translationese". In both datasets, other disagreements were inconsistent. Some were clearly due to annotator error; some are due to ambiguous context or complicated sentence structure. However, an handful of disagreements pointed to framesets that need to be reexamined and perhaps changed. The framesets that need to be considered with care are остаться 'to stay/ to leave/ to live', отправить 'to send/to set off', идти 'to go, move, walk', and утверждать 'to confirm/to strengthen'.

Agreement on numbered arguments (i.e. Arg0-Arg5) was higher on the Phrasebook dataset, for instance ARG0 (92.9 vs. 74.1), with decreasing accuracy on higher-numbered arguments. This is likely related to the more complicated

sentence structure of newswire data compared to simple phrases. Also, higher numbered arguments tend to occur less often and are less core to the verb, so they may have been confused with modifier arguments (e.g. Goal, Location).

	Phrasebook	Newswire
Predicates (no roleset)	99.5	98.8
Predicates + roleset id	88.6	94.6
Roleset id agreement	89.0	95.8
ARG0	92.9	74.1
ARG1	86.6	73.6
ARG2	72.7	51.6
ARG3	47.1	50.0
ARG4	40.0	50.0
ARG5	0.0	-
ARGM-ADV	0.0	0.0
ARGM-COM	50.0	-
ARGM-EXT	0.0	50.0
ARGM-GOL	46.2	33.3
ARGM-LOC	76.2	65.2
ARGM-MNR	25.0	0.0
ARGM-MOD	40.0	-
ARGM-NEG	63.0	0.0
ARGM-TMP	79.3	74.1
ARGM-ADJ	0.0	-
ARGM-PRP	0.0	0.0
ARGM-DIS	0.0	-
ARGM-PRD	0.0	50.0
ARGM-DIR	0.0	-
ARGM-CAU	0.0	21.1
Overall (Exact)	81.1	74.5
Overall (Partial)	88.9	82.5

Table 1: Inter-annotator agreements, reported as F-score. Dash means that particular argument did not occur in that dataset

.

7. Conclusion

This paper presents a semantic role labeling resource for Russian called the Russian PropBank. The PropBank preserves consistent mapping with the English PropBank frames and semantic role labels as much as possible. The RuPB, freely available online, contains annotated sentences to illustrate the usage of the Russian predicates. The creation of the Russian PropBank rolesets followed language-specific issues such as grammatical aspect, reflexive forms, and nominalization. The main contribution of this paper is the Russian PropBank itself, as well as the linguistic criteria for distinguishing the rolesets. Although the paper describes criteria that is specific to Russian, the principle of following intrinsic semantics of a language is applicable to the development of new proposition banks in any language. Developing proposition banks in multiple languages allows for consistent cross-linguistic mapping of semantic roles and verb senses. The Russian PropBank is facilitating a project to align semantic roles across languages. A small corpus has been annotated with RuPB labels and is serving as evaluation data for a project to automatically project of semantic role labels from English to Russian texts.

PropBanks are designed for interoperability with Abstract Meaning Representations (AMR) (Banarescu et al., 2013). The goals of the two resources are complementary. AMRs make heavy use of the English PropBank frames as semantic concepts in order to abstract away from English syntax. Uniform Meaning Representations (UMRs) (Myers and Palmer, 2019; Pustejovsky et al., 2019; Vigus et al., 2019; Xue et al., 2019) is a current development to extend AMRs to multi-lingual settings. RuPB is designed to support the development of UMRs. For example, its parallel to English PropBank frames supports cross-linguistic mapping and testing.

8. Acknowledgements

We gratefully acknowledge the support of DARPA HR0011516904-Lorelei, Semantic Annotation and Technology Transfer, a subaward from LDC, DARPA FA8750-18-2-0016-AIDA – RAMFIS: Representations of vectors and Abstract Meanings for Information Synthesis, and NSF 1764048 RI: Medium: Collaborative Research: Developing a Uniform Meaning Representation for Natural Language Processing. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, NSF or the U.S. government.

Thanks to Adam Pollins for his contribution towards calculating the inter-annotator agreement.

9. Language Resource References

The Russian PropBank Roleset files are available at https://github.com/cu-clear/RussianPropbank.

10. Bibliographical References

Akbik, A., chiticariu, l., Danilevsky, M., Li, Y., Vaithyanathan, S., and Zhu, H. (2015). Generating high quality proposition banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407. Association for Computational Linguistics.

Apresyan, Y. D., Boguslavsky, I. M., Yomdin, B. L., Yomdin, L. L., Sannikov, A. V., Sannikov, V. Z., Sizov, V. G., and Tsinman, L. L. (2003). *: // Syntactic and semantic annotated corpus of the Russian language: Current situation and perspectives*. National Russian Corpus.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M.,

and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bazrafshan, M. and Gildea, D. (2013). Semantic Roles for String to Tree Machine Translation. In *Association for Computational Linguistics (ACL-13) short paper*.

Bonial, C., Conger, K., Hwang, J., Mansouri, A., Aseri, Y., Bonn, J., O'Gorman, T., and Palmer, M., (2017). *Current Directions in English and Arabic PropBank*, pages 737–769. 06.

Cohen, K. B., (2020). *Biomedical computational linguistics and natural language processing*. Oxford University Press.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Duran, M. S. and Aluísio, S. M. (2012). Propbank-br: a brazilian treebank annotated with semantic role labels. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2014-09). Building the essential resources for finnish: the turku dependency treebank. 48(3):493–531.

Kettnerová, V., Lopatková, M., and Bejček, E. (2012). The syntax-semantics interface of czech verbs in the valency lexicon. In Ruth Fjeld et al., editors, *Proceedings of the 15th EURALEX International Congress*, pages 434–443, Oslo. Department of Linguistics and Scandinavian Studies, University of Oslo.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008-03). A large-scale classification of english verbs. 42(1):21–40.

Levin, B. (1993). *English Verb Classes and Alternations*. University Of Chicago Press.

Lyashevskaya, O. and Kashkin, E. (2015). FrameBank: a database of russian lexical constructions. pages 350–360. Springer International Publishing.

Moreda, P., Navarro, B., and Palomar, M. (2005). Using semantic roles in information retrieval systems. In *International Conference on Application of Natural Language to Information Systems*, pages 192–202. Springer.

Myers, S. and Palmer, M. (2019). ClearTAC: Verb Tense, Aspect, and Form Classification Using Neural Nets. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 136–140, Florence, Italy, August. Association for Computational Linguistics.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. 31(1):71–106.

Palmer, M., Ryu, S., Choi, J., Yoon, S., and Jeon, Y. (2006). Korean propbank. *Linguistic Data Consortium Catalogue*, (LDC2006T03).

Pustejovsky, J., Lai, K., and Xue, N. (2019). Modeling Quantification and Scope in Abstract Meaning Representations. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 28–33, Florence, Italy, August. Association for Computational Linguistics.

Sammons, M., Vydiswaran, V., and Roth, D. (2010). Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208. Association for Computational Linguistics.

Sharoff, S. (2002). Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In *In Proc. of Language Resources and Evaluation Conference (LREC02*.

Shelmanov, A. and Devyatkin, D. (2017). Semantic role labeling with neural networks for texts in russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"*, page 12.

Strassel, S. and Tracey, J. (2016). LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Vaidya, A., Palmer, M., and Narasimhan, B. (2013-06). Semantic roles for nominal predicates: Building a lexical resource. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 126–131. Association for Computational Linguistics.

Vigus, M., Van Gysel, J. E. L., and Croft, W. (2019). A Dependency Structure Annotation for Modality. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 182–198, Florence, Italy, August. Association for Computational Linguistics.

Nianwen Xue, et al., editors. (2019). *Proceedings of the First International Workshop on Designing Meaning Representations*. Association for Computational Linguistics, Florence, Italy, August.

Xue, N. (2006). A chinese semantic lexicon of senses and roles. *Language Resources and Evaluation*, 40(3-4):395–403.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The revised arabic PropBank. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10. Association for Computational Linguistics.

Zapirain, B., Agirre, E., Màrquez, L., and Surdeanu, M. (2013). Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.

Şahin, G. G. and Adalı, E. (2018). Annotation of semantic roles for the turkish proposition bank. 52(3):673–706.