АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ РУССКИХ ГЛАГОЛОВ С ИСПОЛЬЗОВАНИЕМ ИНФОРМАЦИИ О МОРФО-СИНТАКСИЧЕСКОМ ОФОРМЛЕНИИ И СЕМАНТИЧЕСКИХ РОЛЯХ УЧАСТНИКОВ ФРЕЙМОВ

Ляшевская О. Н. (olesar@yandex.ru)[1,2],
Кашкин Е. В. (egorkashkin@rambler.ru)[2]

[1]Национальный исследовательский университет
Высшая школа экономики, Москва

[2]Институт русского языка им. В. В. Виноградова РАН, Москва

В статье описываются эксперименты по классификации русских глаголов на основе статистических данных, представленных в системе FrameBank (framebank.ru). Хотя лексикологи в основном отказались от мысли, что группы глаголов должны объединяться на основе способности к синтаксическим трансформациям (Apresjan 1967, Levin 1993), оценка близости контекстов по схожей дистрибуции лексики и синтаксических связей по-прежнему остается ведущим критерием для определения лексических типов. Компьютерная лингвистика заимствовала последний подход для получения глагольных классов для английского, немецкого и многих других языков (Dorr and Jones 1996; Lapata 1999; Schulte im Walde; Lenci 2014 и др.), строя векторы лексических и синтаксических признаков на основе корпусов текстов.

Наши эксперименты по семантической классификации русских глаголов базируются на статистике двух типов тегов, используемых в аннотации системы ФреймБанк, теге семантической роли и теге морфосинтаксического оформления участника. Поле глаголов речи было структурировано с помощью нескольких вариантов автоматической кластеризации на векторах; затем автоматические результаты мы сравнили с классификацией глаголов в словаре Л. Г. Бабенко (2007) и некоторыми другими построенными вручную классификациями. Классификация глаголов смены посессора была построена с помощью правил и затем была верифицирована относительно сети глагольных фреймов в англоязычной системе FrameNet. Проводится лингвистический анализ классификаций, получающихся только на морфосинтаксических признаках, только на признаках семантических ролей и классификаций на объединении этих признаков.

Ключевые слова: лексические классификации, глагол, FrameNet, FrameBank, семантические роли, морфосинтаксис, фреймовая семантика, лексикология, русский язык

INDUCING VERB CLASSES FROM FRAMES IN RUSSIAN: MORPHO-SYNTAX AND SEMANTIC ROLES[1]

Olga Lyashevskaya (olesar@yandex.ru)[1,2],
Egor Kashkin (egorkashkin@rambler.ru)[2]

[1]National Research University Higher School of Economics
[2]V. V. Vinogradov Russian Language Institute of RAS, Moscow

The paper presents clustering experiments on Russian verbs based on the statistical data drawn from the Russian FrameBank (framebank.ru). While lexicology has essentially abandoned the idea of syntactic transformations as the primary basis for grouping verbs into semantic classes (Apresjan 1967, Levin 1993), the hypothesis of the same lexical and syntactic distributional profiles underlying lexical clusters is still attractive. In computational linguistics, some attempts have been made to obtain verb classes for English, German and other languages using observable morpho-syntactic and lexical properties of context (Dorr and Jones 1996; Lapata 1999; Schulte im Walde 2006; Lenci 2014, among others).

Our experiments on semantic classification of Russian verbs are based on two types of tags embedded in the annotation of argument constructions: a) semantic roles and b) morpho-syntactic patterns. The domain of speech verbs is classified automatically on vectors, and the resulting clusters are contrasted against Babenko (2007)'s semantic classes and three other manual classifications. The classes within the domain of possessive verbs are constructed using rule-based solutions and evaluated against Berkeley FrameNet verb clusters. We conclude that clustering on morpho-syntactic (pure formal) patterns loses the race to more intelligent approaches which take into account semantic roles.

Key words: lexical classifications, verb, FrameNet, FrameBank, semantic roles, morpho-syntax, frame semantics, lexicology, Russian language

1. Introduction

The systematic lexicography approach (Apresjan 2000, 2002) generalizes over words according to their properties to share patterns of conceptualization and the regular paths of meaning development and interaction (polysemy, antonymy, etc). Word classes are also expected to manifest similar trends in grammatical, syntactic and lexical co-occurrence behavior. These ideas establish grounds for the unified

[1] This work was partly supported by the Russian Foundation for the Humanities, project #13-04-12020 "New Open Electronic Thesaurus for Russian" and Russian Basic Research Foundation, project # 15-07-09306 "Evaluation benchmark for information retrieval".

representation of word classes or 'types' (in terms of their semantics and 'lexical grammar') in lexicography, functional and cognitive linguistics, and computational linguistics.

In computational linguistics, word classes and nets help to reduce the sparseness of lexical vectors which measure how often each word from the corpus (one coordinate axis) occurs in the context of the target word. If two words belong to the same class, the corresponding dimensions can be collapsed into one; this can also help to associate context elements not available in the training set. As a result, the use of lexical classes (along with other categories available in corpus annotation such as parts of speech, lemmas, ie. the sets of word forms, semantic roles, syntactic relation types etc.) affects data pruning, feature weighting and feature selection and can be considered potentially good way to improve machine learning. The only limitation is the availability of large-scale lexical classifications as open-access resources.

There is a number of manual builds of verb classes for several languages, including English (Levin 1993 and its implementation in VerbNet, Palmer 2009; Baker, Fillmore, and Lowe 1998), Spanish (Vázquez et al. 2000), Russian (Babenko 2007, Shvedova 1998–2007), etc. More attempts have been made to obtain verb classes automatically (Dorr and Jones 1996; Lapata 1999; Korhonen 2002; Schulte im Walde 2006; Lenci 2014, etc). Lenci (2014) distinguishes between the ontology-based and distribution-based classifications. As an instance, the verbs *eat* and *devour* belong to the same group in the ontololology-based classification since they evoke the same frame Ingestion 'an Ingestor consumes Ingestibles'; in contrast, *eat* and *devour* do not share certain syntactic properties such as object drop and conative construction and therefore can be placed in different groups in at least some versions of distribution-based classifications. Lenci's example is misleading since there are two context vector models underlying the distribution-based classification. If the idea of syntactic transformations is taken into account, then the target words are seen as being in two states (cf. two isotopes of a chemical element) in which they behave differently and their context image consists basically of two classes of vectors. In the more straightforward reading of the distributional hypothesis, the context vectors of the target word form a homogeneous image. The transformational hypothesis has been put under question by Construction Grammar and quantitative corpus-based approaches. As corpus data show, the alternations are rather peripheral than central phenomenon (see discussion in Kuznetsova, Lyashevskaya 2009; Kuznetsova 2013), and verbs from the same lexical class demonstrate strong statistical preferences for either one or another alternating construction (Gries, Stefanowitch 2004). Therefore, we leave transformations out of the model in order to make it less computationally complex.

In our approach, we take both latent frame-based cues and observable morpho-syntactic cues in order to evaluate their classification strength in the task of Russian verb clustering. The paper is structured as follows. Section 2 outlines Russian FrameBank as a data source for our case studies. In Section 3, we introduce the case study on speech verbs clusters which were classified by machine learning and contrasted against four gold standards. Section 4 summarizes an experiment where possessive verbs were classified using rule-based solutions and then evaluated against Berkeley FrameNet verb clusters. Section 5 concludes.

2. Data

The Russian FrameBank (framebank.ru, Lyashevskaya, Kuznetsova 2009; Kashkin, Lyashevskaya 2013; Lyashevskaya, Kashkin 2014) includes a dictionary of lexical constructions and a corpus of manually annotated sentences (up to 100 examples from the Russian National Corpus for each target word). In our experiments we use two types of tags embedded in the annotation of argument constructions both in the dictionary and the corpus: a) semantic roles and b) morpho-syntactic properties which form a formal pattern of constructions. For example, the verb *govorit'* is associated with a number of frames, cf. the frame of CONVERSATION:

Semantic roles of frame elements: Speaker, Counter-agent, Topic

Furthermore, each frame is associated with a set of lexical constructions, cf. The two-argument construction in (1) with a particilar pairing of meaning (formalized as a combination of semantic roles) and form (formalized as a set of morpho-syntactic constraints):

(1) *Dmitriev govoril s dochkoy*. 'Dmitriev talked to his daugher'.
Semantic role pattern: <Speaker, Counter-agent>
Morpho-syntactic pattern: <NPnom[2], *s* 'with' + NPgen>

Example (2) presents the frame of INFORMATION TRANSFER (e. g. saying smth. to smb.) and the three-argument construction of the verb *govorit'* 'say':

(2) — *Vsego etogo nedostatochno, — govoril mne Dviniatin*. 'Dviniatin said to me, 'All this is not enough'.
Semantic role pattern: <Speaker, Addressee, Message-as-content>
Morpho-syntactic pattern: <NPnom, NPdat, CL>

In our first case study we explored the contexts of speech verbs which were assigned to frames where at least one participant plays a role of Speaker, Addressee, Topic, or Message-as-content. The data set included vectors for 80 speech verbs having speech frames being associated with their primary meaning.

For the second case study we used only data from the dictionary database, namely, information on semantic roles, morpho-syntactic tags, and their matches in lexical constructions. We included into our experiment 128 verbs having the arguments with the roles of Initial Possessor or Eventual Possessor. If a verb represented such role patterns for more than one frame in the database (e. g., the verb *vz'at'* which may refer either to TAKING or to BYING), these cases were counted as different verbs (e. g., we analyzed a verb *vz'at' 1* 'to take' and the verb *vz'at' 9* 'to buy').

---

[2] Here and throughout, *nom* stands for Nominative case, *gen* for Genitive, *dat* for Dative, *acc* for Accusative, *ins* for Instrumental, *loc* for Locative, *CL* for clause; {ADV / PRfrom_where + Npx} refers to any prepositional phrase or adverb with the meaning 'from a certain source'.

3. Case study 1: speech verbs

In the first case study, we explore subclasses of speech verbs. These verbs differ in terms of associated morpho-syntactic constructions and combinations of participants in the frames they evoke; the most common set of roles usually includes 1) Speaker; 2) Addresse or Counter-Agent; 3) Message-as-Topic and / or Message-as-content. Rare cases include such roles as Motivation (cf. *khvalit' za pirogi* 'to praise (smb.) for cakes'), Quantity (cf. *povtorit' dvazhdy* 'to repeat twice'), Point of Destination (cf. *zvat' v park* 'to call (smb.) to the park'), etc.

Figures 1 and 2 report how often morpho-syntactic and role tags occur in the context of verbs (in each case the verbs are sorted separately according to the ratio of contexts that include a given element).

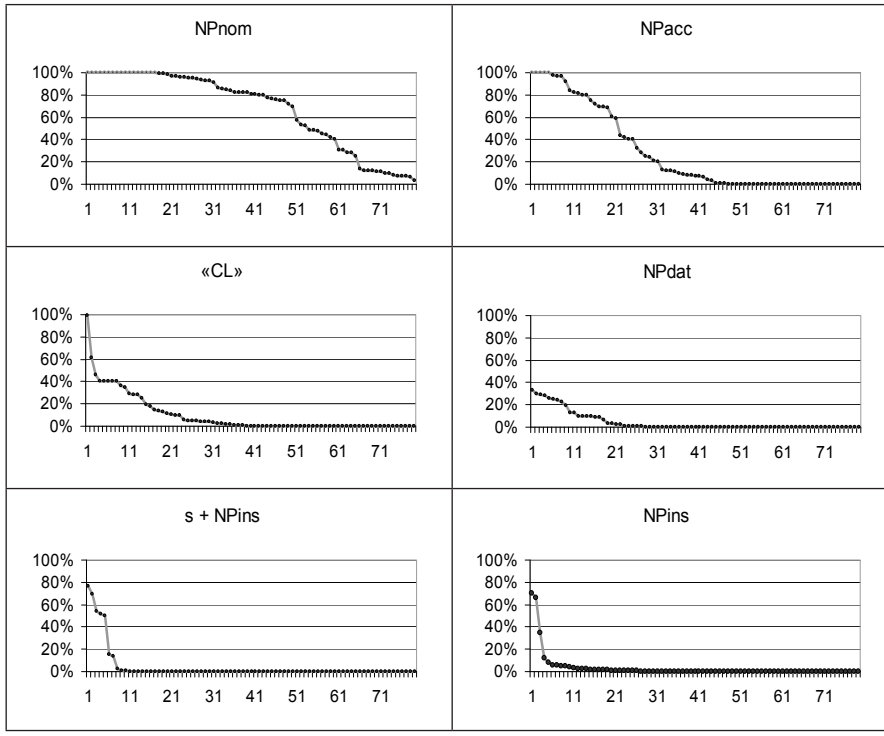

Fig. 1. Verbs sorted by the ratio of top-6 morphological tags[3] in their context, in % of tagged examples for each verb

[3] *s* + S ins is a prepositional group which means "with + S ins', «CL» stands for the direct speech clause; see also footnote 2.

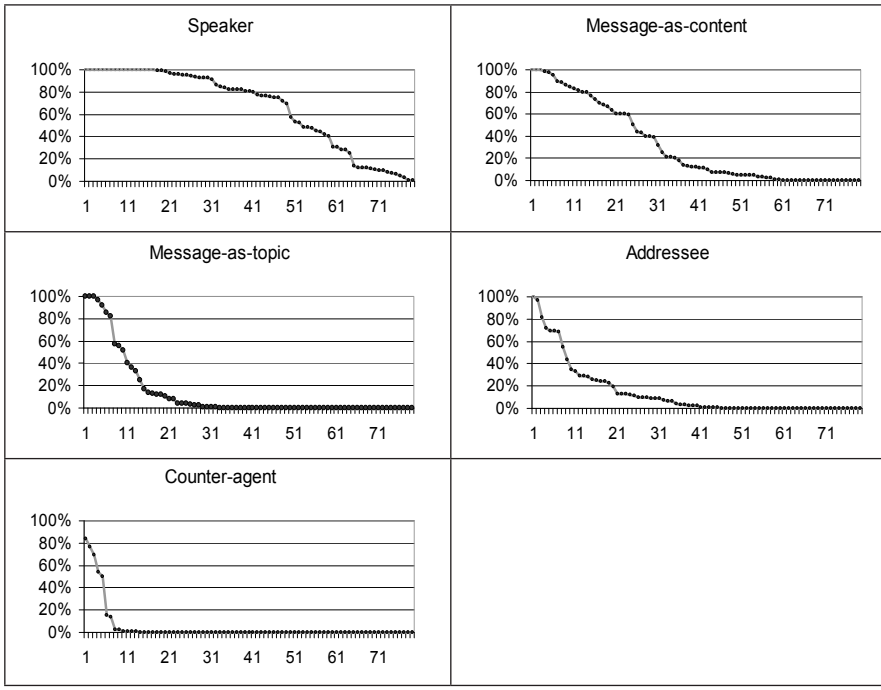

Fig. 2. Verbs sorted by the ratio of top-5 semantic role tags in their context (Speaker, Message-as-content, Addressee, Message-as-topic, Counter-agent), in % of tagged examples

Three types of vectors presenting morpho-syntactic tags (e. g. *o* + NPloc), semantic roles (e. g. Topic), or their matches (e. g. *o* + NPloc|Topic) are gathered taking frequencies from annotated corpus as coordinates. The data set includes vectors for 80 verbs which refer to speech in their primary meaning. As a result, there is a 34-dimensional vector space for morhosyntactic tags (tags that occur less than 5 times such as *v kachestve* + NPgen 'qua', *ot imeni* + NPgen 'on behalf of' are removed from the data set), 20-dimensional space for semantic roles (roles that occur less than 5 times such as Result and Direction are removed as well), and a 71-dimensional space for the combined features of morpho-syntax and semantic roles (also pruned with the threshold of 5).

Table 1 shows the comparison of k-means-based clustering[4] results against four variants of gold standard. The metrics of Purity (PU), Collocation (CO), and F1 are understood in accordance with (Lang, Lapata 2011). PU is calculated as (3), where

[4] K-means is a traditional algorithm which finds the best partition of points in n-dimensional vector space (in our case, verbs) into k clusters such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized (for an overview and discussion, see Jain 2010). K is a fixed positive integer number specified by the researcher. K-means starts with a (random) initial partitition with K seed points selected as cluster centers and initial assignment of data points to clusters. After that, the data points are reassigned to its closest cluster center and then new cluster centers are calculated, and these two steps are repeated until cluster membership stabilizes.

n denotes the total number of instances, G_j is the set of instances belonging to the *j*-th gold class and Ci is the set of instances belonging to the i-th cluster. CO measures how well the procedure meets the goal of clustering all gold instances with the same label into a single predicted cluster and is computed according to (4). F1 is the harmonic mean of Pu and CO.

$$\mathrm{Pu} = \frac{1}{n}\sum_{i=1}^{n_C} \max_{j=1,\ldots,n_G} \left|C_i \cap G_j\right| \tag{3}$$

$$\mathrm{Co} = \frac{1}{n}\sum_{j=1}^{n_G} \max_{i=1,\ldots,n_C} \left|C_i \cap G_j\right| \tag{4}$$

It is important to keep in mind that there cannot be an ideal gold classification due to different principles that could be applied to data (e. g. thematic proximity, event structure, pragmatic goals, etc., cf. the classifications of Wierzbicka 1983; Bogdanov 1990; Glovinskaya 1993; Era Kuznetsova 1989; Shvedova 1998–2007; Babenko 2007, among many others). Rather, it is better think of probability to which a linguistic community would agree to assign the verb C to the same cluster as A and B. Given that, we built four variants of classification: 1) based on Babenko 2007's classes (our verbs fall into 23 Babenko's classes including six classes of speech verbs and some classes outside the speech domain like behavior, emotion, etc.); 2) based on the role of the 2nd participant: addressee-like verbs, counteragent-like verbs, patient-like verbs, benefactor-like verbs, no-addressee verbs; 3) based on the goals of the speaker (7 classes); 4) its more detailed version with 19 classes of sharing information, getting information, symmetric communication, and various types of speech affect like asking, abusing, etc. The number of verbs accumulated in classes in each gold standard is reported in Table 2.

According to F-measure, Roles generally overperform Forms with the only exception of the best split in cross-validation against the last gold standard (Goal33). Interestingly, Forms perform better at smaller *k*-s while Roles work better at larger *k* values. Roles & Forms optimizes the split in three cases of four but there are many cases there Roles demonstrate higher scores than Roles & Forms at the same *k* value. Thus, we conclude that the hierarchy of features predicting speech verb classes looks like the following: Roles & Forms ≥ Roles > Forms.

In the second trial, we use the same vector datasets (Forms, Roles, Roles & Forms) to compare three hierarchical cluster trees based on cosine distances and to follow the verbs changing (or not changing) their position in clusters. At $k = 7$, there are eight clusters of size 4 to 14 where the verbs group together under all three conditions (e. g. {*besedovat', zdorovat'sja, obschat'sja, prostit'sja*}, {*blagodarit', informirovat', pozdravit', privetstvovat', khvalit', zvat', klikat', oprosit'*}, etc.; 63 verbs in total). Due to the small number of verbs in clusters, it is easy to inspect the homogeneity of clusters manually. The indisputable errors include such pairings as *obvinit'* and *ugovorit'* ('accuse' and 'persuade'), *prokl'ast'* and *obosnovat'* ('imprecate' and 'justify'), *podkhvatit'* and *ugrozhat'* ('play along' and 'threaten').

Table 1. Evaluation of k-means clustering by Purity (PU), Collocation (CO) and their harmonic mean F1 (Lang, Lapata 2011). The best trials are bold-faced. PU tends to increase and CO tends to decrease as the number of clusters *k* increases

Babenko (k(gold)=23, elements in the largest class = 26)									
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.3418	**0.9114**	0.4971	0.3544	**0.8861**	0.5063	0.3418	**0.9241**	**0.4990**
5	0.3544	**0.9114**	0.5104	0.3797	0.6709	0.4850	0.3418	0.7342	0.4664
13	0.4430	0.6329	**0.5212**	0.4557	0.5949	0.5161	0.4304	0.5570	0.4856
20	0.4810	0.4937	0.4873	0.5316	0.5190	0.5252	**0.4557**	0.4430	0.4493
21	**0.5063**	0.5190	0.5126	0.5316	0.5190	0.5252	0.4430	0.4810	0.4612
23	0.4810	0.4684	0.4746	0.5570	0.5063	**0.5304**	**0.4557**	0.4430	0.4493
25	0.4304	0.4684	0.4486	**0.5823**	0.4557	0.5113	0.4177	0.4304	0.4240
Role2 (k(gold)=7, elements in the largest class = 33)									
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.4051	**0.9114**	0.5609	0.4430	**0.8228**	**0.5759**	0.4051	**0.9241**	**0.5632**
5	0.4177	**0.9114**	0.5729	0.4937	0.5949	0.5396	0.4051	0.6076	0.4861
7	0.5570	0.6582	**0.6034**	0.4430	0.5696	0.4984	0.4937	0.5063	0.4999
10	0.4937	0.6076	0.5447	0.5696	0.5190	0.5431	0.4684	0.4304	0.4486
SpeakerGoals ((k(gold)=7, elements in the largest class = 30)									
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.4304	**0.9114**	0.5847	0.4557	**0.8354**	0.5897	0.4304	**0.9241**	**0.5872**
5	0.4430	**0.9114**	0.5962	0.5190	0.5696	0.5431	0.4430	0.6076	0.5124
7	**0.5823**	0.6962	**0.6342**	0.4810	0.5316	0.5051	0.4937	0.5190	0.5060
8	0.5696	0.5570	0.5632	**0.6456**	0.5443	**0.5906**	0.4810	0.5063	0.4933
9	0.5696	0.5570	0.5632	0.6203	0.4810	0.5418	**0.5316**	0.4177	0.4678
Goals19 (k(gold)=19, elements in the largest class = 26)									
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.3418	**0.9241**	0.4990	0.3671	**0.8354**	0.5101	0.3418	**0.9241**	0.4990
5	0.3544	**0.9241**	0.5123	0.3924	0.6582	0.4917	0.3418	0.6835	0.4557
11	0.3924	0.6076	0.4768	0.4684	0.5823	0.5191	0.4810	0.6709	**0.5603**
19	0.5316	0.6076	**0.5671**	0.5443	0.5190	0.5313	0.5190	0.5063	0.5126
22	**0.5696**	0.4684	0.5140	0.5823	0.4937	0.5343	0.5316	0.4557	0.4907
23	0.5443	0.4430	0.4885	**0.5949**	0.4810	0.5319	0.5316	0.4430	0.4833
24	0.5443	0.4810	0.5107	**0.5949**	0.4937	**0.5396**	**0.5443**	0.4430	0.4885
25	**0.5696**	0.4557	0.5063	**0.5949**	0.4684	0.5241	**0.5443**	0.4430	0.4885

Table 2. The size of classes in four gold standards, sorted from larger to smaller groups

Babenko	26	10	9	7	4	4	2	2	1	1	1	1 in each other class
Role2	30	20	16	9	2	1	1					
SpeakerGoals	33	25	8	5	3	3	2					
Goals19	26	9	8	6	5	4	3	3	3	2	2	1 in each other class

For example, *obvinit'* and *ugovorit'* are similar in terms of Roles vectors (Speaker—Addressee—Message-as-content available in context), but partly different in terms of their main morpho-syntactic patterns, cf. NPnom NPacc *v*+NPloc and NPnom NPacc VPinf. The cosine measure shows very small distance between their vectors due to orthogonality effects where the following situation takes place:

dimensions	1	2	3	4	5	6	7	8	*i*
vector 1	a	b	-	c	-	-	-	1	-
vector 2	a	b	c	-	-	1	-	-	-

In this particular case, the number of *v*+NPloc and VPinf tags which occur in context is 55 and 54, respectively, while other tags (e. g. Message-as-content|Conj + CL, reason|Sins, etc.) occur not more than 1 time. If the verb is somewhat unique (cf. the high ratio of *v*+NPloc tags in the context of *objasnit'*), the probability that the hierarchical clustering will produce an error becomes even greater.

4. Case study 2: possessive verbs

Our second case study deals with verbs that refer to change of possession (frames of buying, stealing, giving, etc). We created a boolean table with the verbs in the rows and with the possible pairings of morpho-syntactic tags and semantic roles (e. g. NPnom|Eventual Possessor) in the columns. It was marked in the table which clusters of morpho-syntactic patterns and semantic roles are compatible with each verb.

Further, all the verbs were manually divided into clusters based on a set of heuristics involving their morpho-syntactic patterns and semantic roles. The results were verified with an agglomerative clustering method[5]. As a result, we found four main clusters of verbs, most of them are further subdivided into several smaller clusters. The structure of the possessive domain is represented below (the verbs of each subclass are also included into all the parent classes). The figures in brackets show the number of verbs which have fallen into a particular class.

1. Take (34 verbs): verbs with patterns where NPnom is an Eventual Possessor (e. g. *brat' 1* 'to take', *otn'at' 1* 'to take away').

1.1 Buy (4 verbs): verbs having a pattern NPnom V NPacc *za* + NPacc (Eventual Possessor—Patient—Price)[6], e. g. *kupit'* 1 'to buy', *ar'endovat' 1* 'to rent'

1.2 Steal (10 verbs): verbs having a pattern NPnom V NPacc {ADV / PRfrom_where + NPx} (Eventual Possessor—Patient—Starting Point). Interestingly, they

[5] hclust() method in R, package 'stats', see Langfelder, Horvath 2012. The agglomerative starts with one cluster for each verb and merges the pair of clusters with the minimum intercluster distance.

[6] The verbs of Buying may obviously occur in some other patterns, e. g. in those expressing an Initial Possessor or a Place. However, these patterns are not specific for this class of verbs and cannot serve as a diagnostics for it, so we do not refer to them when defining the class of Buying. The descriptions of the other classes follow the same principle: what is specially mentioned is only the diagnostic patterns for each class.

tend to describe the events of theft[7]. There are 10 verbs with this pattern in the database, and 7 of them obviously refer to stealing: *vorovat' 1* 'to steal', *krast' 1* 'to steal', *pohitit'* 1 'to steal, to kidnap', *taskat' 4* 'to pinch', *taščit'* 6 'to pinch', *t'anut' 11* 'to swipe', *uv'esti 2* 'to steal (usually cattle or a car)' The other three verbs with this formal pattern are *brat' 1* 'to take', *zabrat' 1* 'to take away', and *hvatat' 1* 'to snatch', which may all be used in neutral possessive contexts outside the domain of stealing. Thus, our protocol suggests here a broader result than necessary, but it is important that all the verbs of stealing are inside this formal class, the only possible "lost" verb is *otn'at' 1* 'to take away, to deprive'

1.3 Receive (3 verbs): verbs having a pattern NPnom V NPacc *ot* + NPgen (Eventual Possessor—Patient—Initial Possessor). There are three verbs—*polučit' 1* 'to receive', *prin'at' 1* 'to accept (e. g. a present)', *prin'at' 4* 'to accept (e. g., an advertisement)', — where the Initial possessor is marked not by the more frequent PP *u* + NPgen, but by *ot* + NPgen. This opposition seems to be related to the degree of agentivity: verbs with *u* + NPgen like *brat' 1* 'to take' imply a more active behavior of the Agent than verbs with *ot* + NPgen argument.

1.4 Earn (2 verbs): verbs expressing Price as NPacc. This is a subclass of taking verbs which corresponds to the events of earning money. In FrameBank these are the verbs *zarabotat' 1* and *polučit' 2* meaning 'to earn (money)'.

1.5 Borrow (3 verbs): verbs of taking which have Time Period among their participants. They refer to the events of borrowing: *ar'endovat' 1* 'to lease sth. from sb.', *zan'at' 8* 'to borrow', *sn'at' 13* 'to rent'

2. Give (90 verbs): verbs with patterns where NPnom is an Initial Possessor, cf. *dat' 1* 'to give', *vozvratit' 1* 'to return sth. to sb.', *pr'epodn'esti 1* 'to present sth. to sb.', etc.

2.1 Sell (6 verbs): verbs having a pattern NPnom V NPacc *za* + NPacc (Initial Possessor—Patient—Price), e. g. *prodat' 1* 'to sell', *sdat' 4* 'to lease sth. to sb.'

2.2 Pay (2 verbs): verbs of giving (*platit' 1* 'to pay', *ustupit' 3* 'to take off (a price)') which have a direct object expressing Price, similarly to the verbs of earning.

2.3 Give somewhere (8 verbs): verbs having a pattern NPnom V NPacc {ADV / PRwhere(to) + NPx} (Initial Possessor—Patient—Point of destination). We put them together under a technical label "Give somewhere" This subclass doesn't appear to be homogenous, but the verbs included there follow some semantic tendencies. First, these are the verbs *vernut' 1* and *vozvratit' 1* meaning 'to return sth. to sb.' Second, this subclass includes the verbs *vyslat' 1*, *poslat' 2* both meaning 'to send' and *p'er'edat' 1* 'to pass sth. to sb.' which presume an intermediary in the change of possession. Third, there are verbs *podat' 2* 'to submit', *sdat' 1* 'to return, to surrender', and *sdat' 2* 'to submit, to hand in' also belonging to this subclass and referring to change of possession as a part of social relationship between the Initial Possessor and some kind of authorities being the Eventual Possessor.

---

[7] Here our results are in line with Apresjan 1967: 176–177, where the class of theft is singled out on the basis of its constructional properties.

2.4 Give with some goal (12 verbs): verbs having a pattern NPnom V NPacc *na* + NPacc (Initial Possessor—Patient—Goal) or a pattern NPnom V NPacc NPdat *na* + NPacc (Initial Possessor—Patient—Eventual Possessor—Goal). All of them conceptualize giving as an action intended to achieve some goal, cf. *assignovat'* 1 'to allocate', *žertvovat' 1* 'to donate', *tratit' 1, 2* 'to spend', *darit' 1* 'to make a present', *pr'epodn'esti 1* 'to present (a gift)' etc.

2.5 Supply (8 verbs): verbs (*balovat' 2* 'to make sb. glad by giving sth.', *vooružit' 1* 'to arm', *nagradit' 1* 'to award', *ob'esp'ečit' 1* 'to provide', *(n'e) obid'et' 2* 'not to stint sb. of sth.', *obogatit' 2* 'to enrich', *ssudit' 1* 'to loan') with a pattern NPnom V NPacc NPins (Initial Possessor—Eventual Possessor—Patient), a verb *od'et' 2* 'to provide clothes for sb' with a pattern NPnom V NPacc (Initial Possessor—Eventual Possessor), and a verb *obogatit' 1* 'to enrich' with a pattern NPnom V NPacc (Method—Eventual Possessor). The core of this set (*vooružit' 1*, *ob'esp'ečit' 1, obogatit' 1, 2, ssudit' 1, od'et' 2*) describes supplying sb. with sth. necessary. However, there are three peripheral verbs (*balovat' 2, nagradit' 1, (n'e) obid'et' 2*) falling into this subclass.

2.6 Lend (2 verbs): verbs of giving which have Time Period among their participants (*odolžit'* 1 'to lend', *sdat' 4* 'to rent out').

3. Exchange (1 verb): a verb *men'at's'a 1* 'to exchange', as its NPnom is Possessor

4. Other types

4.1 Owe (1 verb): a verb *sl'edovat' 9* 'to owe (lit.: to follow)' with its specific patterns, e. g. *Skol'ko* (ADV) *s nih* (s + NPgen) ***sl'edujet*** *za r'emont* (za + NPacc) 'How much do they owe for the repair (lit.: How much follows from them for the repair)?'

4.2 Go to somebody (2 verbs): two verbs with patterns where NPnom is a Patient—*dostat's'a 1* and *otojti 10*—meaning 'to go to sb.'

As has been stated above, our classification of the verbs is based both on their morpho-syntactic patterns and on the sets of semantic roles. If treated separately, these two criteria appear to be less fruitful than their combination. The semantic roles without the syntactic patterns fail to produce an adequate classification, since most possessive verbs are conversives (in the terminology of Apresjan 1974/1995: 256–283) and involve the same set of participants getting different syntactic ranks. The morpho-syntactic structure is more successful for clustering the possessive domain, e. g. the verbs of Taking and Giving complementary fit the patterns Snom V Sacc *u* + Sgen and Snom V Sacc Sdat respectively, the pattern Snom V Sacc *ot* + Sgen is unique for the verbs of Receiving. In many cases, however, the syntactic clustering lacks information on semantic roles and therefore produces too broad classes (as is sometimes the case in Apresjan 1967, along with a great deal of reliable correlations between semantic classes and constructional patterns). Thus, the classes of Buying and Selling admit the same syntactic pattern Snom V Sacc *za* + Sacc and are differentiated due to different correspondence between the semantic roles and syntactic participants. The classes of Paying and Earning encounter a similar problem, being marked out on the grounds

of Price being their direct object. In the pattern Snom V Sacc *na* + Sacc, the PP may perform 7 different roles in different verbs (Patient, Resource, Period, Eventual Possessor, Point of Destination, Price, Goal), so its specification as Goal is necessary for defining the class of "Giving with some goal". The patterns Snom V Sacc Sins and Snom V Sacc {ADV / PRwhere(to) + Sx} mostly correspond to the classes of Supplying and "Giving somewhere", but the syntactic clustering without the semantic roles produces 2 and 1 false results respectively (including *brat' 1* 'to take' and *kupit' 1* 'to buy' into the former domain and *vyslat' 1* 'to send' into the latter).

We have compared our results with the data of Berkeley FrameNet as the gold standard. The latter contains a frame of Giving with 6 subframes (Commerce_pay, Commerce_sell, Lending, Submitting_documents, Supply, Surrendering_possession), and the frame of Getting with 8 subframes (Amassing, Commerce_buy, Commerce_collect, Kidnapping, Receiving with the subframe of Borrowing, and Taking, further inherited by Theft).

The basic distinction between Giving and Getting is the same in FrameNet and in our survey. The frames of Commerce_pay, Commerce_sell, Lending, Commerce_buy, Receiving, Borrowing, Taking, and Theft transparently correspond to our classes. The frame of Amassing (e. g., *Bogs accumulate carbon for thousands of years*) is outside the possessive domain in FrameBank. The domain of kidnapping seems to be much more elaborated in English than in Russian, including even such specific verbs as *to shanghai* defined in the Oxford Dictionary as 'to force (someone) to join a ship lacking a full crew by drugging them or using other underhand means', therefore we haven't revealed this verb class in FrameBank. The frame of Commerce_collect has a bit strange definition in FrameNet ('Subframe of Commerce_money-transfer in which the Seller comes to have the Money', e. g. *The man at the counter collected payment from Lee for his dry-cleaning*), as the grounds for focusing on the motion of the Seller are not quite clear, but it roughly corresponds to our subtype "Earn"

The frames of Submitting_documents and Supply are both included into broader classes with some periphery ("Give somewhere" and "Supply", respectively). The only frame present in FrameNet but missing in our clustering is Surrendering_possession ('A Surrenderer is compelled to transfer a Theme to a Recipient', e. g. *Shortly after the boy surrendered the gun, the three remaining warriors made a rush for liberty*): we suggest that these verbs are treated separately due to rather fine-grained semantic reasons and do not seem to have their specific constructions.

Interestingly, our protocol has shown the subclass of Giving with some goal ('to allocate', 'to donate', etc.). These verbs do not form a single class in FrameNet, but intuitively they form a homogenous semantic class with a common set of participants, therefore our method seems to have been more successful here.

5. Conclusions

There is a great many statistical approaches to clustering word vectors which have been developed over the past decades. With access to ever growing corpora and handy script libraries and ready-made services, the task of clustering verbs

in different languages and domains seems pretty straightforward. As Alessandro Lenci pointed out in the draft of his paper (Lenci 2014), 'We have no doubt that verbs can be grouped into classes, since almost everything can be classified'. In this paper we have argued for the need to draw more attention to the input data structure while using the same algorithms and to involve the multiple gold standard approach.

Our case studies of speech verbs and possessive verbs have shown that unsupervised clustering performs better if semantic roles are taken into account, either as the only input (in the case of speech verbs) or together with the morpho-syntactic patterns. These observations have been made on rather small experimental dataset available for Russian and imply that the future development of this approach would require enhancing large corpora with SRL annotations. Given the recent success in deep learning and semantic role labeling in general (see Lang, Lapata 2014; Hermann et al. 2014; Täckström et al. 2015 etc.) and in Russian SRL parsing (Smirnov et al. 2014; Kuznetsov 2015), this does not sound as an unrealistic challenge.

References

1. *Apresjan, Juri.* 1967. Experimental'noe issledovanie semantiki russkogo glagola. Moscow.
2. *Apresjan, Juri.* 2000. Systematic lexicograpy. Oxford: Oxford University Press.
3. *Apresjan, Juri.* 2002. Principles of systematic lexicography. In Marie-Heléne Corrèard (ed.), Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins. Euralex, Grenoble, pp. 91–104.
4. *Apresjan, Juri.* 1995. Integral'noe opisanie iazyka i sistemnaia leksikographiia [An Integrated Description of Language and Systematic Lexicography]. Moscow: Jazyki russkoj kul'tury.
5. *Apresjan, Jury D., Pall, Erna.* 1982. Russian verb—Hungarian verb. Government and combinability [Russkij glagol—vengerskij glagol. Upravlenie i sochetaemost'], Tankyonvkiado, Budapest.
6. *Babenko, Ludmila G.* 2007. Big Explanatory Dictionary of Russian Verbs: Ideographic description. Synonyms. Antonyms. English translation equivalents [Bol'shoj tolkovyj slovar' russkix glagolov: Ideograficheskoe opisanie. Sinonimy. Antonimy. Anglijskie Ekvivalenty]. Moscow: AST-press.
7. *Baker, Collin F., Charles J. Fillmore, and John B. Lowe.* 1998. The Berkeley FrameNet Project. In Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada. Pp. 86–90.
8. *Bogdanov V. V.* 1990. Speech communication: pragmatic and semantic aspects [Rechevoe obschenie: pragmaticheskie i semanticheskie aspekty]. Leningrad: LGU.
9. *Dorr, Bonnie J., Jones, Doug.* Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. Proceedings of the 16th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1996.

10. *Glovinskaya, Marina Ya.* 1993. Semantics of speech verbs from the point of view of the speech act theory [Semantika glagolov rechi s tochki zrenija teorii rechevykh aktov]. In The Russian language in its functioning [Russkiy jazyk v ego funkcionirovanii]. Moscow. Pp. 158–218.
11. *Gries, Stefan Th. & Anatol Stefanowitsch.* 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics 9 (1). Pp. 97–129.
12. *Hanks, Patrick.* 1996. Contextual Dependency and Lexical Sets. International Journal of Corpus Linguistics, Vol. 1(1), pp. 75–98.
13. *Hermann, Karl Moritz, Dipanjan Das, Jason Weston, and Kuzman Ganchev.* 2014. Semantic frame identification with distributed word representations. In Proceedings of ACL 2014.
14. *Jain, Anil K.* 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, Vol. 31 (8), pp. 651–666.
15. *Kashkin, Egor, and Olga Lyashevskaya.* Semanticheskie roli i set' konstrukcij v sisteme FrameBank [Semantic roles and construction net in Russian FrameBank]. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2013. Vol. 12 (19), 2013. Moscow: RGGU. Pp. 325–343.
16. *Korhonen, Anna.* 2002. Subcategorization Acquisition. Ph.D. thesis, University of Cambridge, Computer Laboratory. Technical Report UCAM-CL-TR-530.
17. *Kuznetsov, Ilya.* Semantic Role Labeling for Russian language based on Russian FrameBank // AIST-2015. CCIS, Springer (forthcoming).
18. *Kuznetsova, Era V.* (ed.). 1989. Lexico-semantic groups of Russian verbs [Leksiko-semanticheskie gruppy russkikh glagolov]. Irkutsk.
19. *Kuznetsova, Julia, and Olga Lyashevskaya.* Konstrukcii i transformacii [Constructions and transformations]. Electronic publication: Slovo i Jazyk, 2–4 February 2010, Moscow, Russia. IPPI RAN.
20. *Lang, Joel and Mirella Lapata.* 2011. Unsupervised semantic role induction with graph partitioning, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1320–1331.
21. *Lang, Joel, and Mirella Lapata.* 2014. Similarity-driven semantic role induction via graph partitioning. Computational Linguistics 40 (3). Pp. 633–669.
22. *Langfelder, Peter, and Steve Horvath.* 2012. Fast R functions for robust correlations and hierarchical clustering. Journal of statistical software, Vol. 46 (11).
23. *Lapata, Maria.* 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD. Pp. 397–404.
24. *Lapata, Mirella, and Chris Brew.* 2004. Verb class disambiguation using informative priors. Computational Linguistics, Vol. 30 (1). Pp. 45–73.
25. *Lenci, Alessandro.* 2014. Carving Verb Classes from Corpora. In Simone, Raffaele, Francesca Masini (eds.), Word Classes: Nature, typology and representations. Current Issues in Linguistic Theory 332. John Benjamins, Amsterdam, Philadelphia. Pp. 17–36. http://sesia.humnet.unipi.it/lexit/papers/lenciWordClasses.pdf
26. *Levin, Beth.* 1993. English Verb Classes and Alternations. University of Chicago Press, Chicago.

27. *Lyashevskaya, Olga, Kashkin, Egor.* Evaluation of frame-semantic role labeling in a case-marking language. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2014. Vol. 20. Pp. 362–378.
28. *Lyashevskaya, Olga and Julia Kuznetsova.* Russkij FrameNet: k zadache sozdanija korpusnogo slovarja konstrukcij [Russian FrameNet: towards a corpus-based dictionary of constructions]. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2009. Vol. 8 (15), 2009. Moscow: RGGU. Pp. 306–312.
29. *Palmer, Martha.* 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In Proceedings of the Generative Lexicon Conference. Sept. 2009. GenLex: Pisa, Italy.
30. *Pustejovsky, James.* 1995. The Generative Lexicon, Cambridge, Mass.: MIT Press.
31. *Schulte im Walde, Sabine.* 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. Computational Linguistics 32:2. Pp. 159–194.
32. *Schulte im Walde, Sabine.* 2009. The Induction of Verb Frames and Verb Classes from Corpora. Corpus Linguistics. An International Handbook ed. by Anke Lüdeling & Merja Kytö. Berlin: Mouton de Gruyter. Pp. 952–972.
33. *Shalyapina, Zoya M.* 2001. Co-occurence valencies as a universal tool for description of natural language syntagmatics [Strukturnye valentnosti kak universal'nyi instrument opisaniya yazykovoi sochetaemosti]. Moscow Journal of Linguistics, 2001, Vol. 5, № 2. Pp. 35–84.
34. *Shvedova, Natal'ya Ju.* 1998–2007. Russian semantic dictionary [Russkiy semanticheskiy slovar']. Moscow.
35. *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.
36. *Täckström, Oscar, Kuzman Ganchev, Dipanjan Das.* 2015. Efficient Inference and Structured Learning for Semantic Role Labeling. Transactions of the Association for Computational Linguistics, Vol. 3. Pp. 29–41.
37. *Wierzbicka, Anna.* 1983. Genry mowy. In T. Dobrzyńska, E. Janus (eds.) Tekst i zdanie: Zbiór studiów. Wrocław: Ossolineum. Pp. 125–137.