Semantic Classification of Tatar Verbs: Selecting Relevant Parameters

Alfiia Galieva[1], Ayrat Gatiatullin[1], Zhanna Vavilova[2]
[1]Tatarstan Academy of Sciences, [2]Kazan State Power Engineering University
E-mail: amgalieva@gmail.com, agat1972@mail.ru, zhannavavilova@mail.ru

Abstract

This paper describes the methodology and current results of the ongoing classification of the Tatar lexicon in the process of developing databases of semantic classes of verbs. Our previous work included a semantic classification of Tatar verbs according to their basic meaning and thematic class. As a result, there have distinguished 59 basic semantic classes, with a semantic tag, or a set of tags, attributed to each of 3,200 verbs.

If the thematic classification is universal and may be applied to any language, the currently developed classification described in this paper is based on the parametric principle and includes a set of morphological, syntactic, semantic, and derivational characteristics that are relevant for Tatar grammatical and semantic systems. In a sense, the work is aimed at creating a Tatar analogue of B. Levin's verb classes, taking into account language-specific features.

In the database, each semantic class, or subclass, is supposed to be provided with a set of admissible diathesis alternations and syntactic descriptions, depicting the verb valency, thematic roles of the arguments and semantic restrictions on them. By now we have created a detailed classification of speech, behavior, sound emissions, weather, emotions, mental states and actions verbs; when selecting the pertinent parameters and verifying their relevance, verbs of other classes were also considered.

Keywords: Tatar verb, semantics, corpus, semantic classes

1 Introduction

In recent decades, the emergence of search engines and annotated linguistic corpora has significantly expanded the toolkit of linguistic studies. Computer dictionaries, in particular, are of crucial importance in natural language processing which is aimed at data interpretation. Such lexicographic resources that provide researchers with examples from existing texts are being developed to help us obtain the latest information on semantics and the distribution of words of different parts of speech, on the models of lexical government, on variations of lexical units of different classes and on other important aspects.

Developing a semantic classification of the vocabulary of any low-resource language is a difficult and time-consuming task, when the main challenge is a deficiency of appropriate lexicographic resources, which results in the necessity to process "raw" linguistic material. This paper describes an approach to developing databases of semantic classes of Tatar verbs. Tatar, which falls into this category of low-resource languages, belongs to the Turkic family and has a rich agglutinative morphological system. Among other parts of speech, the Tatar verb is one of the most complicated, semantically intricate and grammatically sophisticated categories, which is distinguished by a multiplicity of senses, forms and requirements to the argument structure.

Generally speaking, verbs constitute a nucleus of lexical and grammatical systems of any language. The semantic structure of any verb is usually a complex of ontological and relational meaning components, which may find a formal expression on different levels of the linguistic structure. Their complicated semantic organization requires an integrated approach to their classification, a work that acquired a new élan with the rise of computational linguistics in the recent decades.

The classification which is currently being developed by the authors of the paper, relies upon the results of our previous work of selecting primary classes of Tatar verbs according to their basic meaning and thematic properties. That preliminary classification was carried out with the purpose of developing a corpus semantic dictionary, basing on the available explanatory dictionaries of the Tatar language (1977-1981; 2005), bilingual Russian-Tatar dictionaries, and the data from the Tatar National Corpus (2018). As a result, 59 basic semantic (ontological) classes (such as movement verbs, speech verbs, etc.) have been distinguished, marked by two types of tags: constructional (categorical) and semantic (thematic). Constructional components of meaning are identical for all semantic classes and subclasses. The semantic annotation tag system is currently being developed for the Tatar National Corpus, and 3,200 Tatar verbs have already been streamed into semantic classes. The peculiarity of this work is that semantic classes were defined exclusively on a thematic basis, without taking into account individual grammatical behavior or syntactic alternations of verbs, whereupon the resulting classes include items with different structural and syntactic properties (Galieva & Nevzorova 2016).

Thus our next step is to achieve syntactic and semantic coherence among members of classes by refining the lexical material and separating individual subclasses within basic classes. A crucial point in this work is determining a set of relevant grammatical and semantic language-specific parameters to refine the previously defined classes and to distribute verbs with similar syntactic behavior into subclasses. In a way the project is aimed at creating a Tatar analogue of B. Levin's verb classes (Levin 1993), which would focus on examining the distribution of syntactic frames of a verb in order to establish its class – in our case taking into account the language-specific features of Tatar verbs.

2 Related Work

Due to corpus studies, it can be argued that present-day English is rather well provided with semantic verb classifications, the most famous being WordNet (Miller 1995; Fellbaum 1998; Vossen 2002), VerbNet (Kipper et al. 2006; Kipper et al. 2008), FrameNet (Fillmore et al. 2004; Boas 2009) and some others. However, even in the English-speaking domain, experiments in search for the best features for verb classifications proceed. For instance, a research also based on Levin-style verb classification (Li & Brew 2008) is aimed at discovering the optimum combination of syntactic and lexical features for verb classification, a method which has yet to be elaborated.

The methodology of semantic classification of verbs is thoroughly examined in a case study on Italian verb features (Lenci 2014). The author juxtaposes two basic approaches towards classifying verbs: the ontology-based paradigm (like FrameNet) which considers the extra-linguistic situation where the verb's meaning unfolds, and the distribution-based approach (like the above-mentioned Levin's verb classification) which is focused on the verb's linguistic behavior and thus provides a more objective and linguistically relevant methodological framework. The distributional perspective offers a powerful instrument for studying the verb's behavior; taking this perspective, the methods of computational linguistics applied to the study of large-scale corpora are invaluable, as they allow linguists to obtain a plethora of evidence about verbal distributions (Lenci 2014).

Automatic acquisition of information is extremely helpful in compiling a distributional profile of a verb, which would embrace a set of its distributional properties. However, unlike English and other high-resource languages, this is not often the case with languages that are spoken by minor communities within states with a different dominant language. Thus the available corpora of the Turkic languages spoken on the territory of the Russian Federation are not yet provided with any system of semantic annotation (the Tatar National Corpus (2018), the Crimean Tatar Corpus (2018), the Bashkir Corpus (2018), the Tuvan Corpus (2018), the Yakut Corpus (2018), etc.). Its development for a number of corpora is currently underway: the electronic Khakass-Russian lexical database is being provided with an inventory of semantic tags based both on paradigmatic and syntagmatic characteristics of the word forms (Dybo et al. 2015); the corpus of the Tuvan language is also being equipped with the system of semantic annotation (Oorzhak & Khertek 2015).

In the framework of our project, an electronic database of Tatar verbs is being compiled to obtain the distributional profiles of verbal classes. It is supposed to be used in the system of semantic annotation of the Tatar National Corpus, as well as for textual analysis, information retrieval, or machine translation. The following section will be devoted to the approach towards classifying verbs for this project.

3 Selecting Relevant Parameters for Classification

The currently developed classification is based on the parametric principle and includes a set of morphological, syntactic, semantic and derivational characteristics which were considered relevant for Tatar grammatical and semantic systems and which allow distinguishing various semantic groups of verbs. The new classification is based on the following parameters of verbal lexemes:

- thematic features, linked with the verb's thematic class, which allows us to mark up the verb's denotation sphere;
- derivational features, related to the verb's derivation pattern (grammatical class of the stem, derivational meaning of the verb forming affix);
- grammatical features, linked with the valency changing operations of voice affixes (possibility of producing grammatical voice derivatives and particular meanings of voice forms);
- syntactic features, related to the allowable predicate-argument structure and thematic roles of arguments.

Tatar is an agglutinative language characterized by a regular morphology, and the derivational structure of a verb (taking into account its stem's grammatical and semantic class and the verb-forming affix) predicts in many respects the verb's basic semantic and grammatical properties. For example, stems referring to tools join the regular verb-forming affix *-la/-lä* and produce transitive verbs with the basic derivational meaning 'to operate with an instrument named by the noun stem' (Example 1). Adjectives may join the *-lan /-län* affix and produce intransitive inchoative verbs (Example 2). So in many cases verbs of the same derivational structure are characterized by similar grammatical properties and basic syntactic behavior.

(1) *Pıçkı* 'saw' – *pıçkılaw* 'to saw'*; boraw* 'drill' – *borawlaw* 'to drill'*; ütük* 'iron'– *ütükläw* 'to iron'.
(2) *Matur* 'beautiful '– *maturlanu* 'to become beautiful'; *yäşel* 'green' – *yäşellänü* 'to become green'; *turı* 'straight' – *turılanu* 'to become straight, to straighten'.

Another significant feature is valency changing operations of voice affixes – the possibility to produce grammatical voice derivatives and particular meanings of voice forms. The Tatar verb has five grammatical voices (basic, passive, causative, reflexive, and reciprocal). Voice affixes are joined in

a strict order and modify the verb's meaning in a direction which depends on its semantic structure. For example, the reciprocal affix is ambiguous and may accept cooperative (associative), assistive or reciprocal meaning; when joined to verbs of different types, it modifies their meaning in different ways. Thus, with labor verbs the reciprocal affix actualizes the assistive meaning, with emotion verbs – the cooperative (associative) meaning, etc. (Example 3). So joining voice affixes is restricted by the verb's semantics, and the fact of producing particular meanings, in combination with other characteristics, serve as a marker for distinguishing verb classes and subclasses.

(3) *Kuanu* 'to rejoice' – *kuanışu* 'to rejoice together'(the reciprocal affix *-şu* actualizing the cooperative meaning of the emotion verb)*;*
kazu 'to dig'– *kazışu* 'to help somebody dig' *(*the reciprocal affix *-şu* actualizing the assistive meaning of the labor verb);
tayanu ''to lean'– *tayanışu* 'to lean on each other' (the reciprocal affix *-şu* actualizing the reciprocal meaning of the contact verb).

The predicate-argument structure and its surface realization is a very important classification parameter. The forms of verb control and the thematic roles of arguments are to a large extent determined by peculiarities of verb meaning. So mapping the verb's argument structure is an important step in distinguishing verbs of certain semantic types.

The combination of parameters may be represented in the form of a grid (see Table 1). For example, Tatar fear verbs (*kurku* 'to fear, to be afraid'*, örkü* 'to experience a sudden strong fear'*, şölläw* 'to experience a slight fear'*, şürläw* 'to experience a slight fear') are all non-derivative and intransitive, and may join the reciprocal affix (with the cooperative meaning) and the causative affix (with the factitive causative meaning), but cannot join the passive affix. The source (causer) of fear is marked with the ablative affix (Examples 4-6).

Table 1. Basic distribution parameters for Tatar fear verbs

Examples of verbs	Semantic class	Derivational structure	Grammatical forms of main arguments		Joining voice affixes		
			Agent	Causer	Passive	Reciprocal	Causative
kurku	Emotion, fear	Non-derivative	NOM	ABL	-	+	+
örkü	Emotion, fear	Non-derivative	NOM	ABL	-	+	+
sölläw	Emotion, fear	Non-derivative	NOM	ABL	-	+	+
sürläw	Emotion, fear	Non-derivative	NOM	ABL	-	+	+

(4 Bala büredän kurka.
Child wolf-ABL to be afraid of-PRES
'The child is afraid of the wolf'.

(5) Büre balanı kurkıta.
Wolf child-ACC be afraid of-CAUS, PRES
'The wolf scares the child' (the verb of fear with the causative affix).

(6) Balalar bik kurkışalar.
Child-PL very be afraid of-COOP, PRES
'Children are scared (together)' (the verb of fear with the cooperative affix).

Such basic semantic and formal characteristics of verbs denoting fear are sufficient grounds for creating a separate subclass within the thematic class of emotions. The developed parameters of classification are used to represent Tatar verbs in a special database which is presented in the next paragraph.

4 Database of Tatar Verbs

The database of Tatar verbs is implemented by means of the Microsoft SQL Server database management system and is filled manually after carrying out a careful analysis of linguistic data. Figure 1 illustrates the structure of the database, which consists of eight interconnected tables, including

- a list of verbs' semantic classes (subclasses) which consists of basic forms of verbs provided with semantic tags;
- lists of possible voice derivatives (causative, cooperative, reflexive verbs) provided with semantic tags;
- expertly selected examples of verbs' usage provided with descriptions of possible surface realizations of the argument structure (the grammatical forms and thematic roles of the arguments required, and the semantic restrictions on them);
- a brief description of the typical derivational structure of each semantic subclass member with the indication of the possibility of producing voice derivatives;
- a list of semantic tags.

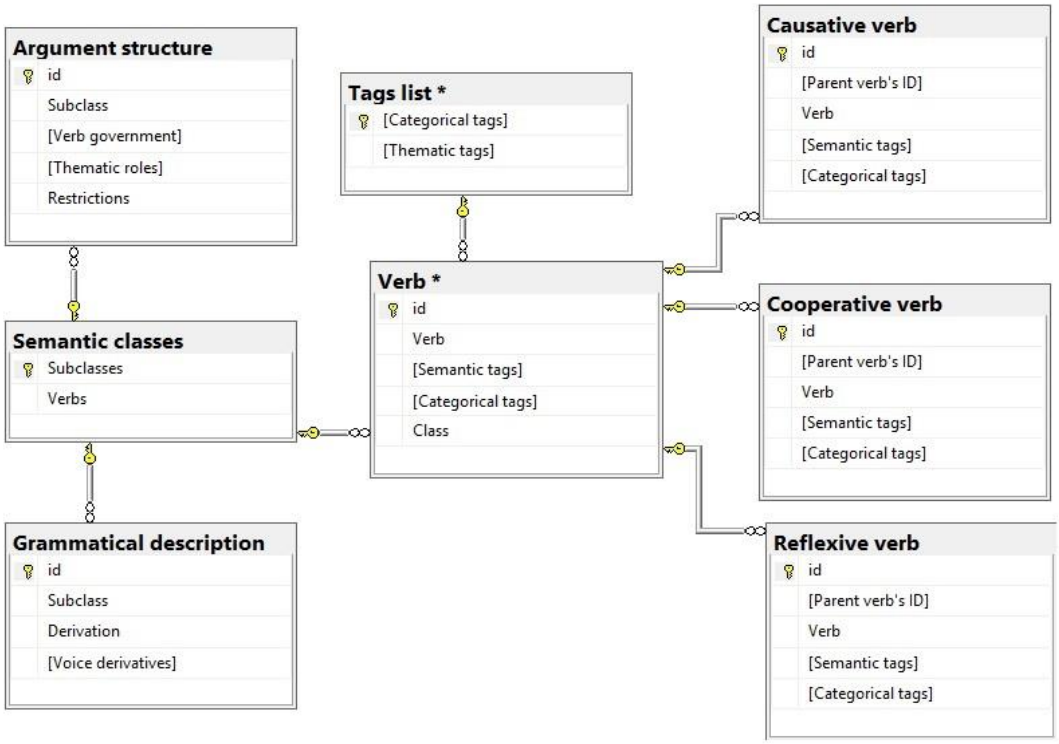

Figure 1: The structure of the database of Tatar verbs.

The database search system provides search for individual verbs by applying semantic tags, represents their semantic class and subclass, as well as enables users to retrieve the list of verbs related to the same subclass. Besides, the database stores information about derivation patterns and possible voice forms. For each subclass there has been chosen a prototypical representative which is provided with typical examples of its usage, information on its syntactic environment and thematic classes of arguments. The information about the typical arguments is taken from contexts of the Tatar National Corpus (2018). Table 2 presents a part of the database table describing the syntactic environment of Tatar fear verbs (whose list is given in Table 1), with the verb *kurku* 'to fear, to be afraid of' as the prototypical verb of the whole fear verbs subclass.

Table 2. Syntactic environment of the verb *kurku* 'to fear, to be afraid of'

1	*Et büredän kurka.* 'The dog is afraid of the wolf'.		
	Verb government	**Thematic role of the argument**	**Semantic constraints to the argument**
	N(NOM)	subject of the emotional state	Living being / creature
	N(ABL)	source/cause of the emotional state	
2	*Bala uramga çıgarga kurka.* 'The child is afraid of going outside'.		
	Verb government	**Thematic role of the argument**	**Semantic constraints to the argument**
	N(NOM)	subject of the emotional state	Living being / creature
	INF_1	cause of the emotional state	
3	*Xatın-kız üz balası öçen kurka.* 'The woman fears for her child'.		
	Verb government	**Thematic role of the argument**	**Semantic constraints to the argument**
	N(NOM)	subject of the emotional state	Living being / creature
	N(NOM) + PSP *öçen* 'for'	cause of the emotional state	

Therefore each subclass contains a list of related verbs and is provided with a set of linguistic descriptions depicting the verb's derivational structure, possible voice derivatives and admissible surface realizations of the argument structure. The items of the same thematic class with a similar meaning, derivation pattern and syntactic behavior fall into the same subclass, while synonyms with different argument structures fall into different subclasses.

In the current version of the database of Tatar verbs words denoting speech, behaviour, sound emissions, weather, mental states and actions, are presented (see Table 3). When selecting the pertinent parameters and verifying their relevance, verbs of other semantic classes were also considered. The study of verbs of other semantic classes on the basis of the developed description model is in prospect, which will considerably extend the scrutinized lexical material.

Table 3. Distribution of verbs and subclasses within thematic classes.

Thematic class	Number of verbs	Number of subclasses	Semantic tag	Examples
Emotion verbs	234	31	t:psych:emot	*yılaw* 'to cry' *moŋayu* 'to sorrow'
Speech verbs	157	17	t:speech	*söyläw* 'to tell' *maktaw* 'to praise'
Behaviour verbs	233	9	t:behav	*aldaw* 'to deceive' *maymıllany* 'to ape'
Mental verbs	119	11	t:ment	*aŋlaw* 'to understand' *kartsınu* 'to consider somebody too old'
Sound emission verbs	230	2	t:sound	*ulaw* 'to howl' *miyawlaw* 'to meow'
Weather verbs	22	3	t:weather	*buranlaw* 'to storm' (of the weather) *bolıtlaw* 'to cloud' (of the sky)
Total	995	73		

5 Conclusion

The presented classification of Tatar verbs is based on the parametric principle, involving a set of semantic, morphological and syntactic characteristics which were considered relevant for Tatar grammatical and semantic systems. This allows us to mark up the denotation sphere of the verb and to consider the grammatical class of the stem and meaning of the verb forming affix, the possibility of producing voice derivatives and particular meanings of voice forms, as well as the verb's valency and the thematic roles of its arguments.

These criteria for classifying verbs are still being revised. By now we have discovered that verbs of the same subclass share some other properties; for example, they may (or may not) be used in certain types of grammaticalised converb constructions. So when adding new items and classes (subclasses) we are planning to update the database with their descriptions considering these new parameters.

The database of Tatar verbs is developed for both professional linguists and Tatar language learners; it can also be used in text processing systems. At present it has a local intranet version; in the future we are planning to develop an online database under a CC-BY-SA copyright license. Currently the database uses a Russian interface and provides linguistic descriptions in Russian. The issue of translating it into English is to be addressed in the nearest future; in particular, with the purpose of extending the field of scientific cooperation in linguistics, it is being planned to develop an English interface and to provide linguistic information on verbs and semantic classes in English.

6 Abbreviations

ABL – Ablative, ACC - Accusative, CAUS - Causative, COOP - Cooperative, INF_1 – Infinitive 1, N – noun, NOM – Nominative, PL - Plural, PRES - Present, PSP – Postposition.

References

Bibliographical Bashkir Corpus. Accessed at: http://mfbl.ru/bashkorp/korpus [29/03/2018].

Boas, H. C. (ed.) (2009). *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Walter de Gruyter.

Crimean Tatar Corpus. Accessed at: http://korpus.juls.savba.sk/QIRIM/#id9 [29/03/2018].

Dybo, A., Sheymovich, A. & Krylov, S. (2015). Some Possibilities of Semantic and Etymological Tagging of Corpora for Turkic Languages. In *Proceedings of the International Conference "Turkic Languages Processing" (TurkLang-2015)*, 17-19 September 2015. Kazan, pp. 304-327.

Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. Cambridge, Mass: MIT Press.

Fillmore, C. J., Baker, C. F. & Sato, H. (2004). FrameNet as a "Net". In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC2004)*, 26-28 May 2004. Lisbon, vol. 4, pp. 1091-1094.

Galieva, A. & Nevzorova, O. (2016). Semantic Annotation of Verbs for the Tatar Corpus. In *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. 6-10 September, 2016. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 340-347.

Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, 24-26 May 2006. Genoa, Italy, pp. 1027-1032.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A Large-Scale Classification of English Verbs. In *Language Resources and Evaluation*, 42(1), pp. 21-40.

Lenci, A. (2014) Carving Verb Classes from Corpora in Word Classes: Nature, Typology and Representations. In R. Simone, F. Masini (eds.) *Word Classes: Nature, typology and representations (Current Issues in Linguistic Theory 332).* John Benjamins Publishing, pp. 17-36.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Li, J. & Brew, C. (2008). Which Are the Best Features for Automatic Verb Classification. In *Computational Linguistics: Human Language. Technologies. Proceedings of the Conference*. 15-20 June 2008. Columbus: The Ohio State University, pp. 434-442.

Miller, G. A. (1995). WordNet: A Lexical Database for English. In *Communications of the ACM*, 38 (11), pp. 39-41.

Oorzhak, B. & Khertek, A. (2015). Development of Semantic Markup for the Corpus of Tuvan Language. In *Proceedings of the International Conference"Turkic Languages Processing" (TurkLang-2015)*, 17-19 September 2015. Kazan, pp. 351 – 373.

Tatar National Corpus. Accessed at:http://tugantel.tatar/?lang=en [29/03/2018].

Tatar Explanatory Dictionary. In 3 volumes (1977-1981). Kazan (In Tatar).

Tatar Explanatory Dictionary. In 1 volume (2005). Kazan (In Tatar).

Tuvan Corpus. Accessed at: http://www.tuvancorpus.ru [29/03/2018].

Vossen, P. (ed.) (2002). *EuroWordNet General Document*. Version 3. Accessed at: http://vossen.info/docs/2002/EWNGeneral.pdf [29/03/2018].

Yakut Corpus. Accessed at: http://adictsakha.nsu.ru/corpora/corp [29/03/2018].