Semantic Annotation of Tatar Verbs for Linguistic Applications

Alfiia Galieva Yakubova
Research Institute of Institute of Applied Semiotics Semiotics of the Tatarstan Academy of Sciences, Sciences
Kazan Federal University
Kazan, Russia
amgalieva@gmail.com

Ayrat Gatiatullin
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences,
Kazan Federal University
Kazan, Russia
agat1972@mail.ru

Olga Nevzorova
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences,
Kazan Federal University
Kazan, Russia
onevzoro@gmail.com;

Dilyara
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences
Kazan Federal University
Kazan, Russia
suleymanovad@gmail.com

Abstract

The paper discusses the problem of meta-language for linguistic applications and proposes a tag set for semantic annotation of verbs for Tatar National Corpus. The approach is based on data from explanatory dictionaries of Tatar and Russian languages, bilingual Russian-Tatar dictionaries and Russian National Corpus.

1 Introduction

The development of meta-language for semantic annotation for linguistic applications and corpora is one of actual problems in applied linguistics. Since there is no common semantic theory, semantic tags given to words and word combinations denote different semantic classes which specify the word meanings. Usually, morphological annotation, which gives basic lexical and grammatical classes of words, is used as a foundation of semantic annotation for vocabulary.

Grammatical annotation uses a fixed set of grammatical classes. The number of semantic attributes depends on generalization level: the more abstract attributes are, the less their number is; the more concrete semantic attributes are, the larger their number is. There are some problems in semantic describing of vocabulary such as the absence of distinct boundary between taxons, the necessity to process very large sets of attributes and semantic features, the complexity of delineation of semantic components in lexical unit and the inability to unambiguously define their features. As a general rule, separate semes do not have special formal indices, so it is difficult, if not impossible, to locate them in the language and to describe them in comprehensive and consistent way. The matching of meanings is a complicated problem, as the question of whether vocabulary, beyond the boundaries of separate groups, such as relationship terms, is systemic or not still open [1], and the meanings of words are very individualized.

Gönderme ve kabul tarihi: 15.09.2014-25.10.2014

Semantic annotation scheme assumes the existence of set of tags, their meanings and rules for application of tags to units of text or vocabulary. At present, there are no standards for semantic annotation creation and there is no semantic annotation in most of the developed corpora of Turkic languages (Tatar - http://web-corpora.net/TatarCorpus/search/?interface_language=ru, Crimean Tatar - http://korpus.juls.savba.sk/QIRIM/#id9, Turkish - www.tnc.org.tr/index.php/en/, Kazakh - http://kazcorpus.kz/klcweb/, Bashkir - http://mfbl.ru/bashkorp/korpus, Tuvan - http://www.tuvancorpus.ru/, Yakut - http://adictsakha.nsu.ru/corpora/corp/).

Quantitative and qualitative features of sets of tags, used in thesauri, electronic corpora and lexicographic databases, are varying. It is obvious, that the larger set of tags is, the more comprehensive analysis of linguistic material can be performed. On the other side, there are some advantages in simple encodings – they are more error-prone, more consistent during the process of annotation and require less handwork. So, it is important to work out the system with balance between available level of detail and simplicity for developers and-users.

Explanatory dictionaries of the Tatar language, bilingual Russian-Tatar dictionaries, thesauri of Russian language and data from Russian National Corpus were used during the development of classification.

There were no integral description of the Tatar language lexical system as a complex hierarchical network of units of different layers and types, so at present there are no ideographic dictionaries of the Tatar language. Therefore, not only general principles of representation of Tatar verbs in corpus annotations have to be developed, but the Tatar vocabulary has to be classified or real content of lexical-semantic groups has to be extracted from raw alphabetical word-list.

The problem of ideographic classification creation (the extraction of thesaurus from semantically unordered alphabetical word-list) comes as applied one, but the process of its solution leads to necessity of general theoretical analysis for systems in the vocabulary, to questions of language nomination and to necessity of revisiting some aspects of field theory of linguistics and vocabulary's structural features and properties [2].

Thesaurus is a specific object which allows an ability to research the systemic properties of language, various relational features, different significative and logical-semantic relationships and relationships of given lexeme to others.

2. Features of semantic annotation of Tatar verbs

The following basic principles of arrangement of lexical data (these principles are used in creation of ideographic dictionaries) were used during the development of semantic annotation of verbs in the Tatar language: system principle, hierarchy principle, variability principle, overlapping of word classes principle.

The ability to consider the overlapping of word classes, when lexeme is described by different independent tags (examples of such lexemes are given in Table 1), is an important advantage of corpora annotation, which is difficult to implement in printed ideographic dictionaries.

Table 1. A fragment of semantic annotation of lexemes

Tatar	English	Tags
aldau	*deceive*	t:speech, t:behav

zarlanu	*resent*	t:speech, t:psych:emot
yavu	*fall (on atmospheric precipitates)*	t:move, t:nat
gırlau	*snore*	t:sond, t:phys

The development of classification is connected to extraction of different lexical-semantic groups (LSG) of verbs, e.g. in well-known research of Levin [5], 57 basic semantic classes of verbs for English language are distinguished. These evaluations are considerably lower for Turkish languages – there are, at average, 10 lexical-semantic classes of verbs.

F. Ganiev [6] distributes verbs of the Tatar language into following 11 LSG:

1. Movement verbs;
2. Action verbs;
3. Process verbs;
4. State verbs;
5. Relationship verbs;
6. Behavior verbs;
7. Sound verbs;
8. Speech verbs;
9. Thought process verbs;
10. Perception verbs;
11. Imitative verbs.

M. Orazov [7] distinguishes the following LSG for Kazakh language:

1. Action verbs;
2. Movement verbs;
3. Relationship verbs;
4. Subjective evaluation verbs;
5. Nature related verbs;
6. Emotional verbs;
7. Sense-describing verbs;
8. Verbs with meaning of creation and appearance;
9. Thought process verbs;
10. Speech verbs;
11. Sound verbs;
12. State verbs.

The semantic annotation solutions from the Russian National Corpus (RNC) were used during the development of semantic annotation for the Tatar Corpus with s for lexical and word-derivational systems of the Tatar language. For example, the following tags were adopted from RNC:

t:move — movement (Example: *cabu (Tat) 'to run*');

t:move:body — change of position of body or a body part (Example: *utıru (Tat)* 'to sit,to sit down');

t:put — object placement (for example: *töyäw (Tat)* 'to load up', *quyu (Tat)* 'to put smth. on/in');
t:impact — physical influence (for example: *sugu (Tat)* 'to hit');

t:impact:creat — object creation (for example: *tözü (Tat)* 'to build')

t:impact:destr — object destruction (for example: *yandıru (Tat)* 'to burn smth. down').

II Basic and additional tags

The system principle assumes the reuse of the same tags for different grammar classes with common meanings. There are some differences in semantic annotation of different parts of speech with common meanings in RNC. For example, during the semantic annotation of nouns, tag t:temper – temperature (Example: cold, chill, heating) is used, but the same tag is not used in semantic annotation of verbs. The verb 'to heat' is only annotated with t:changest [4] – only change of feature is specified.

It is assumed, that in many cases t:chagest tag (state or feature change) can be further clarified with parameter describing the

change (if corresponding tags describing LSG, which may or may not belong to the same part of speech, exist), as in:

T:changest:size – change of size (for example: *zurayu (Tat)* ‘to grow’);

T:changest:form – change of shape (for example: *yäncü (Tat)* ‘to flatten’, *tügäräkläw (Tat)* ‘to make round’);

T:changest:color – change of color (for example: *sargayu (Tat)* ‘to become yellow’);

T:changest:humq – change of human’s mood (for example: *yavızlanu (Tat)* ‘to become exasperated’).

Any tag, which used as main tag describing nouns or adjectives, can be used as clarifying in course of the verb annotation.

It is assumed, that common designation, if possible, should be used when annotating part of speech in lexicographic base of the Tatar language. In Turkic languages there is a special grammar category between nouns and verbs – verbal nouns. The verbal noun describes an action (state or process) in most generalized form (without respect to mood and tense) and has certain grammar features of the verb (aspect, voice, raritivity forms) and the noun (case, plurality, possession) [8]. As such, there are little formal differences between nouns and verbs, and it is a reason for supporting the maximum possible commonality in semantic feature systems for nouns and verbs. For example, in modern grammar dictionaries of the Tatar language many verbs ending in *–u* are tagged as noun and verb at the same time.

Another feature of the Tatar language is a presence of many verbs describing physical influence, for their description tags, missing in RLC, are used, for example:

T:impact:tool – instrumental influence (for example: *boraulau (Tat)* ‘drilling’, *pıcaqlau (Tat)* ‘to cut with knife’, *ütükläw (Tat)* ‘to iron’).

Possessive domain (t:poss) in the Tatar language is clarified using tags describing possession relationship, e.g.:

T:poss:acquire – acquiring (for example: *tabu (Tat)* ‘to find’, *qorallanu (Tat)* ‘to arm with smth.’);

T:poss:deprive – depriving (for example: *yugaltu (Tat)* ‘to loss smth.’, *qoralsızlandıru (Tat)* ‘to disarm’).
There is a relationship domain in the Tatar language (t:relat) with following types of relations:

T:relat:interp – interpersonal relations (for example: *hörmätläw (Tat)* ‘to respect’);

T:relat:social –social relations (for example: *čiŋüü (Tat)* ‘to win’, *yaqlau (Tat)* ‘to defend’).

The following tags, which used to describe semantic in different part of speech, can be used for clarifying the semantics of derived verbs:

T:poss:acquire, pt:part & pc:plant (for example: *botaqlanu (Tat)* ‘to branch’), - here pt:part & pc:plant are related to parts of plants (for example: *yafraq (Tat)* ‘leaf’, *sabaq (Tat)* ‘stem).

In the Tatar language the special tags are used for phase and auxiliary verbs:
Aux: phase – phase verbs (for example: *baślaw (Tat)* ‘to begin’);

Aux – auxiliary verbs (for example: *itü (Tat)* ‘ to do, to make’).

Table 2. Example of semantic annotation of Tatar verbs

	Causation	Taxonomy
Sabaqlanu (Tat) 'to make stems'	ca:noncaus	t:poss:acquire, pt:part & pc:plant
Qaraltu (Tat) 'to darken'	ca:caus	t:changest:color
Käbäkhätlänü (Tat) 'to become sneaky'	ca:noncaus	t:changest: humq

III Conclusion

The proposed system of semantic annotation of verbs can be used for various linguistic applications for the Tatar language. The work for development of semantic annotation tag system for the Tatar National Corpus is in progress. Currently 170 semantic tags are described, the resulting tag set is used in linguistic databases developed at Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences, for example for annotating of multilingual lexicographic databases.

The work is supported by the Russian Foundation for Humanities and the Government of the Republic of Tatarstan, (project # 14-14-16031 a(r)/2014).

IV References

[1] RAHILINA, E. V., PLUNGAN, V.A. On lexical-semantic typology // Verbs describing movement in water: Lexical typology / Edited by. T. A. Mysac, E. V. Rahilina. — M.: Indirk, 2007. - pp. 11-26. In Russian.

[2] KARAULOV, Y.N. General and Russian ideography / U. N. Karaulov.— Moscow: Science, 1976.—355 p. In Russian.

[3] Exploratory dictionary of Russian verbs: Ideographic description / Edited by L.G.Babenko. - M.ASG-Press, 1999. - 704 p. In Russian.

[4] Russian Language Corpus. Semantics // http://www.ruscorpora.ru/corpora-sem.html

[5] LEVIN, B. English Verb Classes and Alternations: a Preliminary Investigation. Chicago: University of Chicago Press, 1993.

[6] *GANEEV, F.A. Semantic classes for verbs in Tatar language*. - Kazan: IALI, 1984. pp.75-84. In Russian.

[7] ORAZOV, M. Kazakh verb semantics (an experience in semantic classification). Alma-Ata, 1983, 56p. In Russian.

[8] Tatar grammar in 3 vol.. Kazan: TBP, 1993. Vol. 2. Morphology. — 398 p. In Tatar.