Corpus based Tatar lexicography: verbs in TatWordnet

Alfiya Galieva\textsuperscript{a1}, Olga Nevzorova\textsuperscript{a,b}, Dzhavdet Suleymanov\textsuperscript{a,b}

\textit{a Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Bauman Str. 20, Kazan, 420111, Russia, b Kazan Federal University, Kremlevskaya, 18, Kazan, 420008, Russia}

Abstract

The paper deals with the issue of presentation of Tatar verbs in a wordnet-like thesaurus relying on data of the Tatar National Corpus. In wordnet building we take into consideration such features of Turkic and Tatar verbs as presence of one-word and compound verbs, and a complicated system of grammatical voices when adding a voice affix modifies in many cases the verb meaning. Using data from the Tatar National Corpus we solve some key problems: improving word definition on corpus data in cases where the definitions given in vocabularies are incomplete; constructing new verbal synsets; constructing a hierarchical network of Tatar verbs.

© 2015 The Authors. Published by Elsevier Ltd.

Keywords: Tatar National Corpus, wordnet, Tatar verbs, synonyms, synset.

1. Introduction

The development of Turkic linguistic studies during the past years has been marked by deepening the theoretical basis of linguistic research, with the increasing emphasis on new directions and challenges of modern linguistics, including applied linguistics. The corpus technology is getting more and more wide application in various aspects of modern linguistics. As a result, corpus-based dictionary compilation became a common method in lexicography. However, corpus-based lexicology and lexicography of the Tatar language, which belongs to the Turkic group and is spoken mostly in the Russian Federation by over 5 million people (Tatar is the second language in the Russian Federation by the number of speakers and geographic distribution), still remains virtually undeveloped area.
Despite the fact that Tatar lexicography has a long history and traditions, it was unable to take into account the actual distribution of linguistic units and all their senses in speech, until today. Units for dictionaries have been selected on the basis of linguistic intuition of dictionary compilers, which is subjective in many respects. Even such relatively modern dictionary as The Tatar explanatory dictionary (2005) has been compiled on the basis of hand-picked card index examples. Insufficient attention has been paid to syntactic factors influencing the meaning (word combinations, verb diathesis, order of words, etc.).

This research focuses on the problems of development of Turkic and Tatar corpus-based lexicography on the example of building the Tatar Wordnet relying on data of the Tatar National Corpus. The development of the Tatar Wordnet is a presently on-going endeavour. In our study we also determine and evaluate some possible ways to improve the structure and content of Tatar lexicons using the corpus data. The main tasks in this aspect are searching for words not included in the dictionaries; searching for typical collocations and quotations to the already known senses of words, as well as for new collocations; searching for the senses of words which have not been included in dictionaries; searching for new senses of words; improving the current definitions of words; analysing the meanings of synonyms, etc.

The paper is organized as follows: Section 2 is devoted to the background of the research, Section 3 examines some morphological features of Tatar verbs that are significant within the wordnet project, Section 4 focuses on some aspects of the methodology, Section 5 shows the ways of use of data from the Tatar National Corpus in developing a wordnet-like thesaurus of Tatar verbs.

2. Related work

Semantic relations that were used to build hierarchical networks of nouns and adjectives cannot be applied for mapping verbs without revision, so many researches draw attention to the issue of presentation of verbs in semantic networks (Fellbaum, 1990, 2003). In Princeton WordNet verb synsets are arranged into hierarchies relying upon several entailment relations (Miller, 1995; Fellbaum, 1998), where prevalent is troponymy, which maps the presence of a ‘manner’ relation between two verbs.

The organization of verb synsets for languages typologically dissimilar to English raises new questions in front of developers. Although the structure of a wordnet thesaurus put on some fundamental restrictions on the possibility of mapping semantic relations (they must stay within the boundaries of semantic relations of Princeton WordNet), the specifics of the language is to be taken into proper account, and the special literature provides us with valuable examples. An issue of representing Indo-Aryan compound verbs is discussed in (Paul, 2010). An attempt to design a sufficiently rich set of semantic relations between verbs in the Polish wordnet is given in (Maziarz, Piaśecki, Szpakowicz, Rabiega-Wisniewska, and Hojka, 2011). Basic principles of building a wordnet-like thesaurus of Tatar verbs are presented in (Galieva, Nevzorova, and Gatiatullin, 2013).

Tatar linguistics has no special dictionaries (thesauri) containing words divided into semantic classes as yet. In the works that are available at the moment we can find only enumeration of several semantic classes and a limited number of the most typical representatives of the class (Ganiev, 1984; Tatar Grammar, 1993).

For wordnet building we have only one specialized dictionary at our disposal– the printed dictionary of synonyms of the Tatar language (1999), compiled by Sh. S. Khanbikova and F.S. Safiullina (Khanbikova and Safiullina, 1999). The volume of the dictionary is limited by 25,000 words of different parts of speech unified in 4,500 entries. Also for wordnet building we use entries of verbal lexis of the Tatar lexicons (The Tatar explanatory dictionary, in 3 volumes, 1977-1981; The Tatar explanatory dictionary, in 1 volume, 2005). In many cases the reasons for selecting the classes and synsets and the criteria of classification in dictionaries remain unclear or seem biased, so we rely upon corpus data for verification, improvement and supplement of lexicographic data.

The Tatar Corpus project was launched in March 2012. It was designed by a team of computational linguists
from Research Institute of Applied Semiotics of Tatarstan Academy of Sciences (www.ips.antat.ru). The Tatar National Corpus “Tugan Tel” (‘Mother Tongue’) at the moment consists of written texts of different genres and styles of the modern literary Tatar language. The main sources of texts for the corpus are fictional texts, educational and scientific literature, texts of Internet publications on social and political topics and texts of official documents (Suleymanov, Nevzorova, Gatiatullin, Gilmullin, and Khakimov, 2013). The corpus is grammatically annotated, and its estimated volume by the end of 2014 was 30 million words.

Grammatical annotation in the Tatar National Corpus provides two main layers of morphological information about the Tatar agglutinative word-form:

- part of speech and its characteristics;
- set of morphological characteristics for each tag denoting a part of speech.

3. Morphological features of the Tatar verbs

Verbs that function as dominant constructional part of speech, constitute the core of lexical and grammatical systems of any language. Verbs are characterized by conceptual complexity and complicacy of their semantic structure and collocation by playing sentence forming role in speaking.

The Tatar language belongs to the Turkic family; Tatar shares characteristic features of all Turkic languages, such as agglutination and progressive vowel harmony. Of all parts of speech the verb stands as the most complex and comprehensive, and the Turkic verb system has particularly complex and branched forms.

In wordnet building we take into consideration the following features of Turkic and Tatar verbs:

- Presence of synthetic (one-word) and analytical (compound) verbs.

Compound verb is a kind of complex predicate that frequently occurs in Turkic languages, and it is usually composed of two words of different types:

a) a notional word and an auxiliary verb: yârdäm (‘a help, an aid’) itü (‘to do’) - ‘to help’, sâyâhât (‘a travel, a voyage’) itü (‘to do’) - ‘to travel, to voyage’;

b) a converb and a verb (both components may express a lexical meaning): uylap (‘thinking’) tabu (‘to find’) - ‘to invent’; qiçqiri (‘crying’) uqu (‘to read’) - ‘to read aloud’;

c) a sound imitation or a image imitation word and an auxiliary verb: šûbir-šûbir (an imitative word) itü (‘to do’) - ‘to rustle, to patter (about rain or water falling from above)’, cem-cem (an imitative word) itü (‘to do’) - ‘to twinkle, to shimmer’;

d) a paired compound verb: borgalanu-sirgalanu - ‘to pose, to show off’, kiyenü-yasanu - ‘to smarten oneself up’.

Complex system of grammatical voices (active, passive, reciprocal (cooperative), causative, reflexive), the ability to combine voice affixes with each other within a word form; joining voice affixes often changes the word meaning:

kilü ‘to come’ - kiterü ‘to bring’ or ‘to make come’; basu ‘to stand up’ - basturu ‘to make stand up sb’, ‘to put vertically sth’ (causatives);
qiçqiru ‘to cry aloud’, qiçqirimu ‘to raise a cry’ (reflexive).

Voice derivatives are formed from each sense of polysemous verbs, so the entries of polysemous verbs may constitute an intricate interweaving of forms and senses.
Table 1. Distribution of voice forms in the Tatar National Corpus

<table>
<thead>
<tr>
<th>Verb</th>
<th>Meaning</th>
<th>Passive</th>
<th>Reflexive</th>
<th>Causative</th>
<th>Reciprocal</th>
<th>Total in the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>yörü</td>
<td>to go, to walk</td>
<td>107</td>
<td>832</td>
<td>44</td>
<td>782</td>
<td>64075</td>
</tr>
<tr>
<td>söylaw</td>
<td>to tell, to speak</td>
<td>0</td>
<td>895</td>
<td>98</td>
<td>7257</td>
<td>62879</td>
</tr>
<tr>
<td>saylaw</td>
<td>to choose</td>
<td>1354</td>
<td>0</td>
<td>43</td>
<td>5</td>
<td>10063</td>
</tr>
<tr>
<td>tañlaw</td>
<td>to throw</td>
<td>0</td>
<td>1293</td>
<td>33</td>
<td>37</td>
<td>20047</td>
</tr>
</tbody>
</table>

Table 2. Combination of voice affixes (data from the Tatar National Corpus

<table>
<thead>
<tr>
<th>Affix/affix</th>
<th>PASS</th>
<th>REFL</th>
<th>CAUS(+dir)</th>
<th>CAUS(+t)</th>
<th>RECP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PASS</td>
<td>-</td>
<td>6751</td>
<td>15313</td>
<td>27</td>
<td>16215</td>
</tr>
<tr>
<td>REFL</td>
<td>6751</td>
<td>-</td>
<td>4573</td>
<td>389</td>
<td>24239</td>
</tr>
<tr>
<td>CAUS(+dir)</td>
<td>15313</td>
<td>4573</td>
<td>-</td>
<td>1457</td>
<td>16285</td>
</tr>
<tr>
<td>CAUS(+t)</td>
<td>27</td>
<td>389</td>
<td>1457</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td>RECP</td>
<td>16215</td>
<td>24239</td>
<td>16285</td>
<td>14</td>
<td>-</td>
</tr>
</tbody>
</table>

- Complex forms of expression of causative category, containing a chain of various causative indices modifying the action expressed by the word stem (two, three or even more causative affixes to the right of the stem):

  basu 'to stand up' - bastru 'to put sth vertically', basörtu 'to make sb put sth vertically', etc.

- In Tatar the same verb may denote an action (yatu - 'to lie down') or a state of being (yatu - 'to be down'); as a result, it may enter multiple synsets.

So in wordnet building we are to solve some theoretical issues of Tatar and Turkic linguistics in relation to our material: demarcate homonymy and polysemy, word formation and grammatical forms for each verb on sense level and word level, and all this is carried out relying on corpus data.

4. Methodology

A wordnet is a computerized dictionary of synonyms and a lexical database, and wordnet projects are based on synonymy relations between words. In a wordnet, nouns, verbs, adjectives and adverbs are organized in a conceptual hierarchy, linking semantically and lexically related concepts to each other (Miller, 1995; Fellbaum, 1998). The prototype of the Tatar Wordnet currently contains about 5,000 Tatar verbs. Within our project we are going to create a model of the semantic system of Tatar verbs as the hierarchical structure considering specifics of the Tatar language.

In EuroWordNet project, two words are semantically equivalent if they denote the same range of entities, irrespective of the morpho-syntactic differences, differences in register, style or dialect or differences in pragmatic use of the words. A nother, more practical, criterion which follows from the homogeneity principle is that two words which are synonymous cannot be related by any of the other semantic relations defined (Vossen, 2002).

So Tatar verbs aqçu and b aççu ('to cry aloud') are included in the same synset, because these words are semantically equivalent and they may replace each other in the same contexts (in spite of stylistic differences).
The same way verbs äytü 'to say' and verbs mıgırdaw, bıdırdaw, lıdırdaw ('to mumble') should not be included into the same synset, because these lexemes are characterized by ‘manner’ relation between their meanings (troponymy relations in WordNet terms), and they belong to different levels of concepts.

Main approaches to wordnet building are the following:

Merge Model: the selection is done within a local resource and the synsets and their language-internal relations are first developed separately, and after that the equivalence relations to WordNet1.5 are generated.

Expand Model: the selection is done in WordNet1.5 and the WordNet1.5 synsets are translated (using bilingual dictionaries) into equivalent synsets of another language. The wordnet relations are taken over and where necessary adapted to EuroWordNet. Possibly, monolingual resources are used to verify the wordnet relations imposed on non-English synsets (Vossen, 2002).

The first reason for not using the latter model is that currently rich English-Russian dictionaries containing 50-100 thousand words (neither printed nor electronic) are not available. Another and maybe more important reason is that a wordnet constructed according to the Merge Model should reflect lexico-semantic relations closer to the specifics of the Tatar language, and a developing thesaurus will not copy decisions made for the lexical system of another language, firstly of English. This is very important for verbal lexis, since the Tatar verbal system differs essentially from the verbal system of English. The aim of our project is to develop a semantic classification of Tatar verbal lexis and to create a complex semantic model of the verbal system of the Tatar language by means of the WordNet technology (Merge Model) (Galieva, 2014).

As a first stage we classify Tatar synthetic verbs into semantic classes according to their main meaning components. Then we select synonymous words (concepts) in derived classes.

5. The use of the Tatar National Corpus information in developing a wordnet-like thesaurus of Tatar verbs

We solve some key problems in the course of the Tatar Wordnet project:

1. Improvement of word definition on Corpus data in cases where the definitions given in the vocabularies are incomplete.

In Tatar lexicons verb definitions are given with differing degrees of accuracy for different words and even semantic classes of words. Let us take the class of sound verbs. The Tatar language has a rich stock of verbs designating various types of sounds in their direct sense, and these verbs are closely attached to the world of sound emitting things with their objective physical and other characteristics. Such verbs imitate sounds of the world, and in Tatar lexicons almost all of the sound verbs are described by means of onomatopoetic words, imitating certain sounds. The following are examples of such definitions from Tatar Explanatory Dictionary (2005):

bezläw, bezeldäw “to emit high-pitched sounds, to emit a sound like bez-bez”; bızlaw “emit a sound resembling bız”; žułdaw “emit a sound resembling žu”.

All these verbs denote sounds of insects making monotonous vibrations, a humming sound during the flight (flapping its wings); nevertheless the definition, given in Tatar Lexicon, does not contain any information on this subject. And dealing with such definitions, a dictionary user cannot estimate the synonymity of verbs bezläw, bezeldäw, bızlaw, žułdaw, so definitions of this sort are almost useless for wordnet building. The corpus contexts containing these verbs usually also include words denoting bees, wasps, flies and some other insects.
So our task is to give a new definition to the verb using corpus data, for example, the synset \{bezlaw, bezeldaw, bezlaw, žulaw, żulaw\} may be defined as ‘to emit monotonous, vibrating, comparatively low sound, by flapping wings during the flight (about insects)’.

2. Constructing new verbal synsets and enriching the available ones.

The available Tatar dictionary of synonyms contains a very limited number of words (less than 1000 synonyms sets), so our task is to make new synsets and to enrich the available ones. For example, the Dictionary of synonyms contains almost no sound emission verbs discussed above. For all the missed verbs we are to build our own synsets using dictionaries and to study some typical collocations of verbs on corpus data to reveal whether the verbs are actually synonyms and whether they may replace each other in contexts.

3. Constructing hierarchical network of Tatar verbal synsets, revealing non-lexicalized hypernyms.

In many dictionaries of synonyms published in Russia concepts of differing levels are traditionally considered to be synonyms, so such dictionaries do not contain any information on hypernymy relations. In Tatar lexicons hypernymy relations are also fixed irregularly.

For example, verbs quurqu, şürlaw, şöllaw, örkü, in the Dictionary of Synonyms are given in the same synset (Khanbikova and Safiullina, 1999). In Tatar Lexicon they identify each other as synonyms (The Tatar explanatory dictionary, 2005). Meanwhile, corpus data make evidence that quurqu, örkü1 mean ‘be afraid of’ and şürlaw / şöllaw mean ‘be rather afraid of, be lightly afraid of’, örkü2 means ‘suddenly be very much afraid of (usually about animals)’. So we see concepts of higher and lower levels distributed to different synsets:

- \{quurqu, örkü1\} ‘be afraid of’
- \{şürlaw, şöllaw\} ‘be rather afraid of, be lightly afraid of’,
- \{örkü2\} ‘suddenly be very much afraid of (usually about animals)’.

In many cases Tatar synsets have no lexicalized hypernyms, for example sound verbs discussed above have no common hypernym ‘to sound’, or, to express it more precisely, the Tatar language has no special verb for designating the process of sound emission as such, regardless of the nature of the emitted sound. In such cases we build the necessary higher level concepts for filling gaps in hierarchy.

So, the construction of hierarchical network of Tatar verbal synsets is accomplished relying on data of the Tatar National Corpus.

4. Including analytical forms in synsets.

The synsets include one-word verbs and compound (analytical) verbs. For example, synset \{quw, qup çgaru, qup cibårů\} ‘to turf out, to drive away’ contains one-word verb quw and compounds qup çgaru and qup cibårů.

5. Correlating causative pairs.

The Tatar language has two types of causatives: morphological and lexical ones.

Morphological causatives are regularly formed by means of special affixes (-br / ter, -dur / -der, -qar / -kär, -ır / -er, -qız / -kez, -qırz / -gez) which are of a great variety, due to phonetic features of the verbal stem:

- alıdru ‘to make take’;
- ayaw ‘to eat’ - ayatu ‘to feed’.
Often the same stem may form derivatives by means of different causative affixes.

Table 3. Example of causative derivatives, formed of the same stem ki- ‘to put on’ (data from the Tatar National Corpus).

<table>
<thead>
<tr>
<th>Causative verb</th>
<th>Causative affix</th>
<th>Number of matches</th>
<th>Number of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>kiderū</td>
<td>-der</td>
<td>681</td>
<td>305</td>
</tr>
<tr>
<td>kigezū</td>
<td>-gez</td>
<td>149</td>
<td>68</td>
</tr>
<tr>
<td>kigerū</td>
<td>-ger</td>
<td>73</td>
<td>44</td>
</tr>
</tbody>
</table>

Lexical causatives have no special morphological index:

iltü 'to carry to sb'
cibärü 'to send to sb'.

The same synset may include morphological and lexical causatives:

{iltü, alıp baru, kiterū} ‘to carry to sb’, where iltü, alıp baru are lexical causatives and kiterū is a morphological causative.

6. Taking corpus frequency information into account for synset construction.

In our developing thesaurus we include verbs or senses only on condition that these words and senses are represented in Tatar National Corpus (at least in several different contexts).

Headwords in synsets are also established by means of statistical method on corpus data, so we are to do it especially in cases where we have some phonetic variants of a verb that bear the same meaning; in such cases it is the most frequent of the words of the synonyms series that is regarded as the headword:

{daŋıldaw, dıŋıldaw, dıŋırdaw, danırdaw} ‘to rumble, to rattle’ (approximate translation).

6. Conclusion

Developing a wordnet-like thesaurus of Tatar verbs allows us to combine the experience of traditional Tatar lexicography and modern information technologies. The Tatar National Corpus plays an important role in building the Tatar Wordnet. Using the corpus technology enables us to create a resource that reflects adequately the distribution of Tatar words and their lexical-semantic variants in real contextual environments.

One of the major objectives of the project is to map the language specific semantic system of Tatar verbs in resources, which would meet the needs of modern computational linguistics.

Thus we solve some key problems in the course of the Tatar Wordnet project:

- improving word definition of Corpus data in cases where the definitions given in the vocabularies are incomplete;
- constructing new verbal synsets and enriching the available ones;
- constructing a hierarchical network of Tatar verbal synsets;
- including analytical forms in synsets;
- correlating causative pairs;
- taking corpus frequency information into account for synset construction.
Acknowledgements

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

References

The Tatar National Corpus <http://web-corpora.net/TatarCorpus/search/?interface_language=en.>