Towards Building Wordnet for the Tatar Language: A Semantic Model of the Verb System

Alfiya M. Galieva, Olga A. Nevzorova, and Ayrat R. Gatiatullin

Kazan Federal University
Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences
{amgalieva,onevzoro}@gmail.com, agat1972@mail.ru
http://ips.antat.ru/

Abstract. Wordnet is a lexical database where nouns, verbs, adjectives, and adverbs are organized in a conceptual hierarchy linking semantically and lexically related concepts to each other. This paper reports on the prototype of the Tatar Wordnet which currently contains about 5,500 Tatar verbs. Within our project we are creating a model of the semantic system of Tatar verbs as a hierarchical structure considering specifics of the Tatar language. For this purpose we use the entries of available Tatar dictionaries (explanatory dictionaries and those of synonyms). As the first step the extraction of available verbal synonyms from the dictionary of synonyms of the Tatar language was carried out. Then the most frequent 5156 Tatar verbs were selected and classified into several groups (synsets) according to their dominant semantic components with the purpose of adding new synsets and enriching those already existing (currently about 1,500 core synsets were distinguished). Then semantic relations between synsets were mapped (the verbs were linked according to their troponymy, entailment, and causality). The paper presents the results obtained, and discusses some problems encountered along the way.

Keywords: Wordnet, synset, Tatar language, verb.

1 Introduction

Developing semantic networks of various types for different languages is an issue of current importance in Natural Language Processing. WordNet [1,2,3] is a lexical database where words marked as belonging to a certain part of speech are linked via semantic relationships. Wordnet-like thesauri are organized around the notion of synset (synonym set).

Wordnets for many languages vary in the degree of development. Wordnets for Turkic languages have not been developed yet. The Turkish wordnet project has been initiated by the Human Language and Speech Technologies Laboratory at the Sabanci University (Kemal Oflazer group) [4], but unfortunately it has not been completed. One of the undertakings in Turkic languages is building the Tatar Wordnet prototype, which is presented in this paper. This project

P. Klinov and D. Mouromtsev (Eds.): KESW 2014, CCIS 468, pp. 57–66, 2014.

is carried out at the Research Institute of Applied Semiotics of the Tatarstan Academy of Sciences. We are going to create a model of the semantic system of Tatar verbs, considering specifics of the Tatar language. Our aim is to build the Tatar Wordnet with modeling of the Tatar verb system using Princeton WordNet core synsets and EuroWordNet Basic Concepts [5].

The paper is organized as follows: Section 2 presents the morphological complexity of the Tatar language and the resources used. Section 3 discusses the process of creating the Tatar Wordnet, limitations of the approach, problems encountered along the way, and proposes some plans for further refinement of the Tatar Wordnet. Section 4 reflects the results of the preliminary research of the Tatar corpus data.

Tatar verbs are given in Turkish-oriented graphics.

2 Challenges

2.1 Morphological Complexity of the Tatar Language

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. The Turkic verb is characterized by the following:

- a complicated negative form (often corresponding to English single word or collocation: *däşű* — to speak, *däşmäw* — to keep silence;
- a complex system of tenses and moods, including synthetic and analytical forms;
- a developed and polynomial system of verbal names: deverbal names (names of actions), adverbial verbs, adjectival and participle forms;
- a complex system of grammatical voices (active, passive, reciprocal (cooperative), causative, reflexive), the ability to combine voice affixes with each other within a word form (*yuu* — to wash, *yuılu* — to be washed, *yuıŝu* — to help wash, *yuınu* — to wash oneself, *yuındıru* — to make somebody wash; *kölü* — to laugh, *kölderü* — to make somebody laugh);
- various forms of expression of causative category; a word form may contain two, three or even more causative indices modifying the action expressed by the word stem to the left of the causative affixes (*qaytu* — to return, *qaytaru* — to bring, *qaytartu* — to make somebody return, to make somebody bring something).

In Tatar the same verb may denote:

- an action: *yatu* 'to lie down',
- a state of being: *yatu* 'to be down'.

As a result, it may enter multiple synsets.

2.2 Available Resources in Tatar Language

Let us give a brief description of the available linguistic resources appropriate for building the Tatar Wordnet.

In the wordnet building basically four kinds of resources have been used:

1. English WordNet as an initial skeleton (lexical database, types of synsets, super-subordinate relations of synsets),
2. already existing taxonomies of the language (both at word and sense level),
3. bilingual dictionaries (English and the target language),
4. monolingual dictionaries [6].

In the development of the Tatar Wordnet we used all these kinds of resources. Moreover, additional data was obtained from published online dictionaries and the Tatar National Corpus.

We have at our disposal only one specialized dictionary – the printed dictionary of synonyms of the Tatar language (1999), compiled by S. S. Khanbikova and F. S. Safiullina [7]. It contains 25,000 words of different parts of speech united in 4,500 entries. The portion of verbal lexis in the dictionary is not large. The main difficulty of working with the dictionary is that criteria for considering words to be synonyms are unclear, and as a consequence the dictionary synonyms series contain numerous descriptive expressions rather than synonyms, so dictionary entries for wordnet building require critical analysis and error correction. The dictionary does not reflect the diversity of the Tatar language, as it contains a small number of entries. A large part of synsets consists of basic vocabulary and they have to be enriched. Besides, being a classical dictionary of synonyms, the dictionary compiled by S. S. Khanbikova and F.S. Safiullina merely consists of a list of synonyms, and it does not contain information on semantic relations between the synonyms series and it does not include concepts that are expressed by single words.

Tatar lexicons in the form of explanatory disctionaries [8,9] provide entries and senses for synonyms extraction and synset construction. Such data are especially important for synsets that are not represented in the dictionary of synonyms or for synsets requiring enriching. The explanatory dictionaries contain data that have been selected for the purpose of representation of the Tatar language's inventory of lexemes; these dictionaries keep a small number of strings of synonyms only as a means of word definition mapping; nevertheless they can be a great help in clarifying concepts and filling synsets.

The entries of Russian-Tatar electronic bilingual dictionaries can also be used as a resource for synonyms extraction, since bilingual dictionaries offer a translation of a number of synonyms for basic meanings of words. The Russian-Tatar electronic dictionary ABBYY Lingvo X3 contains 47,000 words (7896 verbs) [10].

The Tatar National Corpus [11] includes writings of all sorts from literary novels and popular scientific literature and educational texts to everyday newspapers and magazines, texts of Internet publications on informative, social and political topics and official documents. The corpus is an open system, therefore it

permits expansion of the annotation system (currently only grammatical annotation is used). In the current version of the corpus the texts are divided into two types: fiction (71,5 %) and non-fiction (28,5 % of the total volume). In future a more detailed classification of genres of texts will be introduced [12].

The system of morphological annotation of the National Corpus of the Tatar Language is mainly oriented at presenting all the existing grammatical word-forms. In the model used for formal representation of the Tatar agglutinative morphology a word-form is built by consecutive adding regular word-formative and inflectional affixes to the root. As a rule, each grammatical meaning is expressed by a separate affix, and the affixes are unambiguous and regular. Thereby, in order to mark up a word, it is necessary to analyze the structure of its affixal chain, using stems dictionaries. Grammatical annotation of a Tatar word includes the information about the part of speech of the word and a set of morphological features (parameters). The Corpus as the most reliable source of linguistic information is used for revealing frequency distributions of words and senses.

Thus we have lexicographical sources of different types and the corpus text collection for wordnet building. First we extract available verbal synonyms from the dictionary of synonyms of the Tatar language. Then we supplement manually derived synsets and add new ones using the words automatically extracted from other dictionaries and the corpus data.

3 Methodology

3.1 General Principles of Wordnet Development

Despite existing general principles of development of wordnets and Wordnet-like thesauri and depending on the fact that these thesauri may or may not be combined into a system of interconnected semantic networks as EuroWordNet or BalkaNet, a set of resources and methods of their usage varies greatly in different projects. The standard method of constructing national Wordnet-like thesauri includes a conceptual and definitional analysis, an analysis of collocations, corpus studies, processing statistic data, methods of formalization.

There are two basic approaches to the development of Wordwet-like thesauri [14]. The first—the widespread Expand Model—assumes that the selection is done in Princeton WordNet [1] and the WordNet synsets are translated automatically (using bilingual dictionaries) into equivalent synsets into the other language. The WordNet relations are taken over and where necessary adapted to the new wordnet. Possibly, monolingual resources are used to verify the wordnet relations imposed on non-English synsets. In such projects adding synsets which do not exist in Princeton WordNet is often considered as a future plan [13].

Another approach known as the Merge Model sets a task to define synsets and relations in particular language and then align new wordnet with the Princeton WordNet using equivalence relations. The Merge Model results in a wordnet that is independent of Princeton WordNet, which enables to represent and maintain the language-specific properties.

Relations of synonymy, linking words on similarity of the meaning, are basic to all types of Wordnet-like thesauri. By the synset we understand a string of words of the same part of speech that can be interchanged in a certain context.

In EuroWordNet, developers mark two words that denote the same range of entities as semantically equivalent, irrespective of the morpho-syntactic differences, differences in register, style or dialect or differences in pragmatic use of the words. Another, more practical, criterion which follows from the homogeneity principle is that two words which are synonymous cannot be related by any other semantic relation defined [14].

3.2 Language Specific Features of the Tatar Verbs in a Wordnet-Like Thesaurus

Our project's aim 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). The Expand Model is impossible for us to use in default of an English-Tatar dictionary containing the real wealth of the Tatar language both at the word and sense level (available Tatar-English and English-Tatar dictionaries contain only basic vocabulary and can be used only for educational purposes).

The Tatar language has a complex morphology and one of the main reasons for this complexity is the wide use of various combinations (agglutination) of verbal inflectional affixes of different types.

Because of the specificity of the grammatical system of the Tatar language the same synset may contain verbs of the basic voice as well as of other voices (especially causative), for example:

The synset 'to throw': {*taşlau, atu, atıp bärü, ırgıtu*}.

The verbs *taşlau, atu, atıp bärü* are in the form of the basic voice, and the verb *ırgıtu* is in the form of the causative voice.

In many cases adding an affix and affixes combination to the verb stem modifies noticeably the verb meaning and even leads to a change in its semantic class. Some examples are given in Table 1.

A polysemantic word can belong to multiple synsets (Table 2).

Every synset contains a group of synonyms of different type: 1) one-word synthetic verbs (for example, *uqu* - to read, 2) analytical verbs consisting of a notional word expressing the lexical meaning and an auxiliary verb (for example, *yärdäm itü* - to help, *gıybädät kılu* - to pray, 3) word-combinations which include a word expressing the lexical meaning and a notional verb as an auxiliary verb (for example, collocations like *aşıysı kilü* — to feel hungry).

Monolingual wordnets had to have their synsets aligned with the translation equivalent synsets of the Princeton WordNet. We set a task to create our original model of semantic system of Tatar verbs as the hierarchical structure which would be relevant to the lexical system of the Tatar language. In doing that we rely upon Global Base Concepts [5].

Linguistic specificity of the lexical system causes some difficulties at the stage of alignment of synsets.

Table 1. Meanings of Tatar verbs with different voice affixes

Tatar verb (stem+ affixes)	Voice	English translation	Verb class	transitivity
aldaw (alda+w)	basic	deceive, cheat; trick; swindle	behavior verb	transitive
aldanu (alda+n+u)	reflexive	to be deceived	behavior verb	intransitive
aldatu (alda+t+u)	causative	allow to deceive oneself	behavior verb	transitive
aldaşu (alda+ş+u)	reciprocal (cooperative)	deceive, cheat; trick; swindle	behavior verb	intransitive
räncü (ränc[e]+ü)	basic	to take offense	emotion verb	intransitive
räncetü (ränc[e]+t+ü)	causative	to give umbrage to smb.	behavior verb	transitive

Table 2. Tatar polysemantic verb

Tatar polysemantic verb	sense	synonyms
karaw	to look	karaw, bagu
karaw	to look after	karau, küzätü, saklaw, küz-kolak bulu
karaw	to follow smb.'s example	karaw, ürnäk alu
karaw	to repair	karaw, remontlaw, remont yasaw, tözätü

One of the features of the Tatar language is a large number of lower-level synsets consisting of words of particular meaning, while more general higher-level concepts are often not lexicalized. For example, there are in abundance sound verbs characterizing sound in many particular aspects (type of sound source, timbre, pitch, homogeneousness or heterogeneousness of the sound, etc.), but there is a lacuna as to a verb denoting sound emission in general (no analogous to English verb *to sound* (Table 3)). Most Tatar sound emission verbs have no equivalents in English, for example, verbs in Table 3 may be translated roughly as 'crash; peal; rumble'.

A serious problem for us in the Tatar Wordnet building is the imperfection of word definitions given in the Tatar lexicons. For example, the descriptions of meanings of most sound verbs in the Tatar lexicon look like the following:

daŋgıldaw 'to emit a sound resembling *daŋg*';

dıŋgırdaw 'to emit a sound resembling *dıŋgır*'[9].

So the lexicon entries contain only imitative words, and no description of sound type and character. Such definitions are often unsuitable or deficient for synset construction, thus we intend to offer our original definition for concepts within the framework of our project.

Table 3. Example of sound emission verbs synsets

troponym 'to emit a sound (= to sound)' — non verbalized	troponym 'to cause sound emission' — non verbalized
basic voice	causative voice
specific manner of sound emission' {daňgıldaw, daňgırdaw, dıňgıldaw, dıňgırdaw}	'to cause specific manner of sound' {daňgıldatu, daňgırdatu, dıňgıldatu, dıňgırdatu }

A set of non-lexicalized concepts may be revealed in the course of a semantic analysis on the step of synset building as well as construction and alignment of the hierarchy of synsets.

The table of Verbal Base Concepts selected in the English, Dutch, Italian and Spanish Wordnets includes a concept *to have* as a basic concept of high level [5]. The Tatar language has no verbalized concept of '*to have*'; possessive relations in Turkic languages are expressed by means of the verb *to be*:

Minem maşinam bar.

My car is/exists (word by word translation).

I have a car.

Nonetheless many Tatar verbs contain the concept to have in a bound form:

- *Tamırlanu* 'to take roots',
- *Sabaqlanu* 'to form a stalk',
- *botaqlanu* 'to form branches',
- *börelänü* 'to form buds'.

The semantic structure of these verbs includes the following integral semes: 'beginning', 'proper possessivity', 'meronymy relations' and 'characterization'. So, a large number of Tatar possessive verbs with the meaning component 'part of plant', may be interpreted as 'starts to have what is named a deriving stem', i.e. the interpretation of such verbs can look like 'S starts to have Sm', where Sm is stem (motivating) word. The meaning component 'to have' in a bound form is contained in the semantic structure of many other groups of verbs.

The category of possessivity, as well as that of space and time, can be referred to as a universal category, reflecting typical extra-linguistic relations of possessivity. The basic universal category of possessivity has its real implementation in every language, its unique set of expressive means and its place in a special model of the world. In languages of different types, possessive verbs have different semantic organization, and they are characterized by different features of collocability. Besides, the structure of the category of possessivity is not homogeneous for different lexical classes, that is why in order to determine the boundaries, the composition and peculiarities of implementation of this category, it is necessary to analyze the conceptualization of possessive relations in different lexico-semantical groups.

The main characteristic of the Base Concepts is their importance in wordnets. From our point of view, the concept *to have* is very important in the semantic system of the Tatar language, and its importance is caused by the ability of this concept to function as an anchor in attaching other concepts with possessive meaning. Although the concept *to have* is not lexicalized in Tatar, nonetheless the meaning component 'to have' is to be distinguished in the semantic structure of some groups of verbs and to be used in the constructing hierarchy. So the structure of the thesaurus should take into account the lexicalized concepts as well as non lexicalized ones.

If entries of lexicographic resources seem arbitrary we search the corpus data for information. Let us take, for example, the synset *to help*; the dictionary compiled by Sh.S. Khanbikova and F.S. Safiullina represents it as {*yärdäm itü* (headword), *yärdämgä kilü, yärdäm kürsätü, bulışlık itü, bulışka kilü*} [7].

The corpus data give evidence that noun *bulışlık* 'help' combines with auxillary verb *itü* as well as *kürsätü* (roughly 50% of documents contain *bulışlık itü*, 50% — *bulışlık kürsätü*), so the synset with headword *yärdäm itü* 'to help' must contain the collocation *bulışlık kürsätü*. Whereas the study of frequency distribution shows that the collocation *bulışka kilü* is characterized by low frequency (3 occurrences only) and may be excluded.

As a result the synset with headword *yärdäm itü* 'to help' looks like the following:

{*bulışu, yärdäm itü, yärdäm kürsätü, bulışlık itü, bulışlık kürsätü*}.

So our task consists of extracting synsets from available dictionaries, enriching these synsets, adding other semantic links to the taxonomic structure, and aligning this structure with other existing ontologies (Princeton WordNet and EuroWordNet).

One of the biggest problems facing the developers of the Tatar Wordnet is representing actual distribution of meanings of Tatar verbs. To achieve this goal the contexts of lexemes under consideration are extracted from the Tatar National Corpus. The set of extracted contexts for each lexeme is annotated regarding the scheme of meanings given in the explanatory dictionary.

In selecting the optimal number of corpus contexts for the analysis we have relied on the results obtained by I. Azarova and her colleagues during the creation of Russian Wordnet (RusNet)[15]. According to these data, the selective annotation of 100-150 contexts taken randomly from different works gives the same distribution scheme of contexts as a complete set, including 1500-2000 contexts. Thereby a set of meanings that should be represented in the thesaurus is established through the context analysis of the corpus data. The isolated (single) instances of realization of meanings are considered occasional. For delimitation of occasional and usual meanings we introduce a threshold in 1% of the total number of contexts. The experiments carried out on the corpus data demonstrate that this value is relevant for selecting common usage senses.

If necessary, headwords in synsets are also established by means of using the statistical method of research of the corpus data.

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

- constructing new verbal synsets and enriching the existing ones;
- constructing the hierarchical network of Tatar verbal synsets;
- including analytical forms in synsets;
- correlating causative pairs;
- improving word definitions on the corpus data in cases where the definitions given in the vocabularies are incomplete;
- revealing non-lexicalized hyperonyms;
- considering corpus frequency information for synset construction.

The feasible application of the developed resource lies in the textual analysis of the Tatar language (i.e. disambiguation), machine translation, semantic annotation of the Tatar National Corpus, and systematization of Tatar verbal lexis in building new dictionaries, in particular, the semantic dictionary of the Tatar verbs.

4 Preliminary Evaluation

As the first step the extraction of available verbal synonyms from the dictionary of synonyms of the Tatar language was carried out (about 1,000 synsets). Then the most frequently used 5,156 Tatar synthetic (one-word) verbs were selected automatically from Tatar lexicon and Tatar-Russian dictionary and manually classified into several groups according to their dominant semantic components. Also the list of most frequently used (common) analytical verbs (compound verbs) in Tatar was compiled from the corpus data, and frequency distribution of these verbs was determined. We have obtained 250 compound verbs having the auxiliary component *itü* (to do, to make) and 100 compound verbs having the auxiliary component *kılu* (to do, to make), for example, *säyähät itü* – to travel, *häräkät itü* – to move, *hökem kılu* - to sentence, to condemn.

We enriched the verbal synsets from the dictionary of synonyms of the Tatar language by manually deriving synsets and adding the words automatically extracted from other dictionaries and the corpus data. The next step is the construction of the hierarchical semantic network of synsets as wordnet requires, which is done manually. Currently about 1,500 core synsets are compiled, with the semantic relations between them mapped according to the verbs' troponymy, entailment, and causality relations.

Preliminary experiments on the corpus data verify that the developed prototype of Tatar Wordnet represents the most significant structural relations of Tatar verbal vocabulary. We have selected 50 sound emission verbs of different types from 25 synsets, then extracted from the Tatar National Corpus and studied 1000 contexts containing these verbs. The context analysis prompts a conclusion that selected synonyms satisfy the criterion of interchangeability. Almost all of the sound verbs have causative correlates. Lexicalized and non-lexicalized concepts at the higher, more abstract levels of hierarchies correspond to their English analogues. Nevertheless, the synonyms of the low level reflect language-specific lexicalization patterns.

5 Conclusion

Our goal is to combine the experience of traditional Tatar lexicography, the reliable corpus data and the advantages of the Wordnet thesauri standard that will enable us to represent the Tatar language in a way that would meet the demands of contemporary computational linguistics.

The presented methods enable us to represent adequately the specific features of the Tatar lexicon, and to minimize the subjectivity of lexical data differentiation, thus to make them open for verification and to maintain language-specific relations in wordnets. The current Tatar Wordnet is still being actively developed so the numbers reported are expected to change soon.

Acknowledgments. The work is supported by the Russian Foundation for Humanities (project #14–14–16031).

References

1. WordNet. A lexical database for English, http://wordnet.princeton.edu
2. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 3(11), 39–41 (1995)
3. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
4. Bilgin, O., Özlem, C., Kemal, O.: Building a Wordnet for Turkish. Romanian Journal of Information Science and Technology 7(1-2), 163–172 (2004)
5. Piek, V., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Bertagna, A., Peters, W.: The EuroWordNet Base Concepts and Top Ontology. Deliverable D017 D 34:D036 (1998)
6. Farreres, X., Rigau, G., Rodriguez, H.: Using WordNet for Building WordNets. In: COLING-ACL Workshop on Usage of Wordnet in Natural Language Processing Systems, Montreal, Canada (1998)
7. Khanbikova, S. S., Safiullina, F. S.: Dictionary of synonyms of Tatar language. Kazan (1999) (in Tatar)
8. The Tatar explanatory dictionary in 3 volumes. Kazan (1977-1981) (in Tatar)
9. The Tatar explanatory dictionary in 1 volume. Kazan (2005) (in Tatar)
10. ABBYY Lingvo, http://www.abbyy.ru/lingvo
11. Tatar National Corpus, http://web-corpora.net/TatarCorpus/search/?interface_language=en
12. Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., Khakimov, B.: National corpus of the Tatar language "Tugan Tel": Grammatical Annotation and Implementation. Procedia — Social and Behavioral Sciences 95, 68–74 (2013)
13. Isahara, H., Bond, F., Uchimoto, K.: Development of the Japanese WordNet. In: 6th International Conference on Language Resources and Evaluation, Marrakech (2008)
14. Vossen, P. (ed.): EuroWordNet General Document. Version 3 (2002), http://vossen.info/docs/2002/EWNGeneral.pdf
15. Azarova, I.V., Sinopalnikova, A.A., Yavorskaya, M.V.: Guidelines for RussNet structuring, http://www.dialog-21.ru/Archive/2004/Sinopalnikova.htm (in Russian)