Tatar WordNet: The Sources and the Component Parts

Alexander Kirillovich[1(✉)], Alfiya Galieva[2], Olga Nevzorova[1], Marat Shaekhov[2], Natalia Loukachevitch[1], and Dmitry Ilvovsky[1]

[1] Kazan Federal University, Kazan, Russia
alik.kirillovich@gmail.com, onevzoro@gmail.com, louk_nat@mail.ru, dilvovsky@hse.ru
[2] Tatarstan Academy of Sciences, Kazan, Russia
amgalieva@gmail.com, q-mir-bey@list.ru

Abstract. We describe an ongoing project of construction of the Tatar Wordnet. The Tatar Wordnet is being constructed on the base of three source resources, developed by us. The first source is TatThes, a bilingual Russian-Tatar Social-Political Thesaurus. TatThes, in turn, has been constructed by manual translation and extension of RuThes, a linguistic ontology for Russian. The second source is a Tatar translation of RuWordNet, a wordnet for Russian. This translation was carried out automatically on the base of a Russian-Tatar dictionary, and then was manually verified. The third source is a semantic classification of Tatar verbs, developed from scratch. We discuss the structure, methodology of compilation and the current state these source resources, and justify the choice of them as the initial resources for building the Tatar Wordnet. Our ultimate goal is to publish Tatar Wordnet on the Linguistic Linked Open Data cloud and integrate it to the Global WordNet Grid.

Keywords: Tatar language · WordNet · Linguistic Linked Open Data

1 Introduction

The Princeton WordNet thesaurus (PWN) [1, 2] is one of the most important language resources for linguistic studies and natural language processing. PWN is a large-scale lexical knowledge base for English, organized as a semantic network of synsets. A synset is a set of words with the same part-of-speech that can be interchanged in several contexts. Synsets are interlinked by semantic relations, such as hyponymy (between specific and more general concepts), meronymy (between parts and wholes), antonymy (between opposite concepts) and other.

Inspired by success of PWN, many projects have been initiated to develop wordnets for other languages across the globe. Nowadays wordnet-like resources are developed for nearly 80 languages, but Tatar language is not among them. To fill this gap, we lunched a project of construction TatWordNet, a wordnet-like resource for Tatar.

There are two main approaches for construction of wordnets for new languages: expand and merge [3]. The expand approach is to take the semantic network of PWN

A. Elizarov et al. (Eds.): DAMDID/RCDL 2019, CCIS 1223, pp. 195–208, 2020.
https://doi.org/10.1007/978-3-030-51913-1_13

and translate its synsets into the target language, adding additional synsets when needed. The merge approach is to develop a semantic network in the target language from scratch and then link it to PWN.

Since the merge approach is very labor-intensive and time consuming, the expand approach seems more appropriate for under-resources languages such as Tatar. However, in development of Tatar WordNet, the expand approach can't be directly applied either, due to the lack of large English-Tatar dictionaries, necessary for translation of PWN to Tatar. At the same time, there are several relatively large and high-quality Russian-Tatar dictionaries, so Russian thesauri can be used as the source resources instead of PWN.

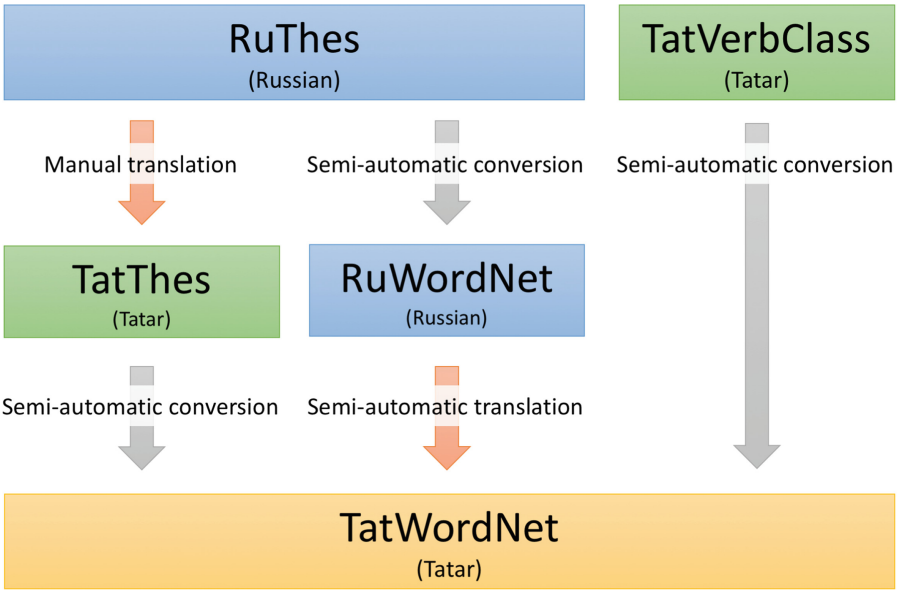

Fig. 1. The source resources of TatWordNet

With this consideration in mind we are constructing TatWordNet on the base of three source resources, developed by us (Fig. 1). The first source is TatThes, a bilingual Russian-Tatar Social-Political Thesaurus. TatThes, in turn, has been constructed by manual translation and extension of RuThes, a linguistic ontology for Russian. The second source is a Tatar translation of RuWordNet, a wordnet for Russian. RuWordNet, for its part, has been constructed by semi-automatic conversion of RuThes. The translation of RuWordNet to Tatar was carried out automatically on the base of a Russian-Tatar dictionary, and then was manually verified. The third source is a semantic classification of Tatar verbs, developed from scratch.

In this paper, we describe the methodology for constructing TatWordNet, and the source resources used in this constructing. The paper is an extended version of our short paper [4], and describes processing of all the source resources (TatThes, Tatar translation of RuWordNet and TatVerbClass), while in [4] processing of the only one source has been described.

The rest of the paper is organized as follows. Section 2 outlines the basic theoretical background of the study, and the main attention is paid to wordnet projects developed for the Turkic languages. Section 3 presents the methodology of compiling the Russian-Tatar socio-political thesaurus and its current state. Section 4 describes the most important aspects of implementing a wordnet-like resource using Tatar thesaurus synsets for Tatar nouns. Section 5 describes a Tatar translation of RuWordNet, and Sect. 6 describes a semantic classification of Tatar verbs. Section 7 discusses the conclusions and outlines the prospects of future work.

2 Related Works

At present time, there are various wordnets for some Turkic languages.

Two Turkish wordnet projects have been developed for the Turkish language. The first one [5, 6] has been created at Sabancı University as part of the BalkaNet project [7]. The BalkaNet project was built on the basis of a combination of expand and merge approaches. All wordnets contain many synonyms for Balkan common topics, as well as synsets typical for each of the BalkaNet languages. The size of Turkish Wordnet is about 15,000 synsets.

Another Turkish wordnet is the KeNet [8, 9]. This wordnet was built on the basis of modern Turkish dictionaries. To build this resource, a bottom-up approach was used. Based on dictionaries, words were selected and then manually grouped into synsets. The relationships between words have been automatically extracted from dictionary definitions and then the latter have been fixed between synsets. The size of this resource is about 113,000 synsets.

Unfortunately, lack of large Turkish-Tatar dictionaries (as well as English-Tatar ones) makes it impossible to translate Turkish resources into the Tatar language. In this respect, the Tatar language can be attributed to low-resource languages.

The Extended Open Multilingual Wordnet [10] resource is built from Open Multilingual Wordnet by replenishing the WordNet data automatically extracted from the Wiktionary and Unicode Common Locale Data Repository (CLDR). The resource contains wordnets for 150 languages, including several Turkic: Azerbaijani, Kazakh, Kirghiz, Tatar, Turkmen, Turkish, and Uzbek. The Tatar wordnet contains a total of 550 concepts, which covers 5% of the PWN core concepts.

The BabelNet [11] resource contains a common network of concepts that have text inputs in many languages. The BabelNet contains 90,821 Tatar text entries that refer to 63,989 concepts. However, due to the fact that this resource was built automatically, it has quality issues.

Thus, the development of a quality Tatar wordnet with an emphasis on the specific features of the Tatar language based on the existing lexical resources is very relevant.

3 Tatar Socio-Political Thesaurus: Methodological Issues of Compiling and Current State

The conceptual model of the Tatar socio-political thesaurus (hereinafter referred to as TatThes) and the general principles of displaying linguistic data are taken from the RuThes project (http://www.labinform.ru/pub/ruthes/) [12, 13]. The RuThes thesaurus is a hierarchical network of concepts with attributed lexical entries for automatic text processing.

In RuThes each concept is linked with a set of language expressions (nouns, adjectives, verbs or multiword expressions of different structures – noun phrases and verb phrases) which refer to the concept in texts (lexical entries). RuThes concepts have no internal structure as attributes (frame elements), so concept properties are described only by means of relations with other concepts.

Each of RuThes concepts is represented as a set of synonyms or near-synonyms (plesionyms). RuThes developers use a weaker term, ontological synonyms, to designate words belonging to different parts of speech (like stabilization, to stabilize); the items may be related to different styles and genres. Ontological synonyms are the most appropriate means to represent cross-linguistic equivalents (correspondences), because such approach allows us to fix units of the same meaning disregarding surface grammatical differences between them. For example, Table 1 represents basic ways of translating Russian adjective + noun phrases into Tatar.

Table 1. Examples of Russian A*dj* + *Noun* phrases and ways of translating them into Tatar

Russian unit	Corresponding Tatar unit	The structure of Tatar unit	English translation
Пенсионный возраст	Пенсия яше	$N + N_{POSS\_3}$	Retirement age
Рабочий класс	Эшчеләр сыйныфы	$N_{PL} + N_{POSS\_3}$	Working class
Консульская служба	Консуллык хезмәте	$N_{NMLZ} + N_{POSS\_3}$	Consular service
Сексуальное меньшинство	Сексуаль азчылык	ADJ + N	Sexual minority
Именная стипендия	Исемле стипендия	$N_{COMIT} + N_{PL}$	Nominal scholarship

TatThes is based on the list of concepts of RuThes, i.e. the Tatar component is based on the list of concepts of the RuThes thesaurus. The methodology of compiling the Tatar part of the thesaurus includes the following steps:

1. Search for equivalents (corresponding words and multiword expressions) which are actually used in Tatar as translations of Russian items.
2. Adding new concepts representing topics which are important for the sociopolitical and cultural life of the Tatar society and which are not presented in the original RuThes (for example, Islam-related concepts, designations of Tatar culture specific phenomena, etc.).

3. Revising relations between the concepts considering the place of each new concept in the hierarchy of the existing ones and, if necessary, adding new concepts of the intermediate level. So an important step is to check up the parallelism of conceptual structures between the languages.

TatThes is mainly being compiled by manual translation of terms from RuThes into Tatar; besides the Tatar language specific concepts and their lexical entries are added (about 250 new concepts). Search for equivalents in the Tatar language in many cases became a time-consuming task because available Russian-Tatar dictionaries of general purpose contain obsolete lexical data [14]. So when compiling the lists of concept names and lexical entries we manually browsed large arrays of official documents and media texts in Tatar. In the process of compiling the Thesaurus, data from the following available Tatar corpora is used:

1. Tatar National Corpus (http://tugantel.tatar/?lang=en);
2. Corpus of Written Tatar (http://www.corpus.tatar/en).

In the course of the project, we found that a distinguishing feature of contemporary Tatar lexicon is a great deal of absolute synonyms of different origin and structure, the main cause of the phenomenon being language contacts [15].

TatThes is implemented as a web application and has a special site (http://tattez.turklang.tatar/). Additionally, it has been published in the Linguistic Linked Open Data cloud as part of RuThes Cloud project [16]. Currently TatThes contains 10,000 concepts, 6,000 of them provided with lexical entries.

4 Tatar Thesaurus Data for Wordnet Implementation: Case of Nouns

Previously, the RuThes thesaurus has been semi-automatically converted to a wordnet-like structure, and a Russian wordnet (RuWordNet) has been generated [17, 18]. The conversion included two main steps:

1. automatic subdivision of RuThes text entries into three nets of synsets according to parts of speech;
2. semi-automatic conversion of RuThes relations to wordnet-like relations.

The current version of RuWordNet (http://ruwordnet.ru/eng) contains 110 thousand Russian unique words and expressions. The same approach can be used to transform TatThes to Tatar wordnet.

The TatThes data may serve as an initial basis for wordnet building for the following reasons:

1. The sociopolitical sphere covers a broad area of modern social relations. This area comprises generally known terms of politics, international relations, economics and finance, technology, industrial production, warfare, art, religion, sports, etc.

2. Currently TatThes, in addition to terminology, comprises some general lexicon branches representing lexical items which can be found in various domain specific texts.
3. Semantic relations in TatThes are necessary and sufficient to arrange the Tatar nominal vocabulary (nouns and noun phrases) as a wordnet-like network of synsets.

Thesaurus concepts unite synonymous items, so we have ready sets of synonyms as building blocks for the wordnet. The concepts are linked by semantic relations with each other. In the RuThes and in the TatThes there are four main types of relationships between concepts, see Table 2. Semantic relations, mapped in the wordnet, are not shared by all lexical categories, so converting thesaurus data into the wordnet format requires dissimilar ways for different parts of speech.

Table 2. Semantic relations between nouns in the thesaurus and in wordnets

Semantic relations in the Thesaurus	Semantic relations in wordnets
Hypernym—hyponyms	Hypernym—hyponyms
Holonym—meronym	Holonym—meronym
Symmetrical association (Asc)	
Asymmetric association (Asc1/Asc2)	

Asc and Asc1/Asc2 association relations need additional explanations. The Asc symmetrical association, distinguished in RuThes and inherited by the Tatar Socio-Political Thesaurus, connects very similar concepts, which the developers did not dare to combine into the same concept (for example, cases of presynonymy of items).

The Asc1/Asc2 asymmetric association connects two concepts that cannot be described by the relations mentioned above, but neither of them could exist without the other (for example, concept SUMMIT MEETING needs existence of the concept HEAD OF THE STATE). In studies of ontologies this relation may be mapped as the ontological dependence relation.

Nevertheless, basic semantic relations which we need to group noun concepts into the wordnet are presented in TatThes.

The core of TatThes is made up of nouns and noun phrases (see Table 3), so the bulk of thesaurus data may be used for Tatar wordnet building without significant changes (synonymous items are yet joined into synsets and the required relations between them are selected).

An important issue is reflecting Tatar specific word usage features in the resource. Mere presence of the shared concepts in languages does not necessarily evidence the same ways of usage of individual words or of usage of words of individual semantic classes. Consider this in the following example. A specific feature of the Tatar language is using hypernyms before a corresponding hyponym, and such usage is not regarded as pleonasm in many cases (examples 1–3):

Table 3. Number of noun concepts and noun phrase concepts in TatThes (on the data of the Russian part)

Structure of TatThes items	Number of items
Noun	3387
Adj + Noun	3135
Noun + Noun_{GEN}	352
Other	3126
Total	10000

(1) *Париж шәһәрендә* ‘in the city of Paris’ (instead of ‘in Paris’);
(2) *кыз кеше* ‘girl human’ (instead of ‘a girl’);
(3) *май аенда* ‘in the month of May’ (instead of ‘in May’).

In cases when such usage is conventionalized and corpus data evidences that the usage has a high frequency, we include such hyponym-hypernym items into the list of lexical entries of the concept. Such manner of designating is a feature of using toponyms and some classes of general lexicon, so it should be considered in Tatar wordnet building. For example, lexical entries of month names include such conventionalized noun phrases, composed of the month name and the hyponym designating month in general, see Table 4.

Table 4. Representing lexical entries of month names in the Thesaurus

Russian concept name	Russian lexical entries	Rus POS	Tatar concept name	Tatar lexical entries	Tat POS
ДЕКАБРЬ	Декабрь ‘December’ Декабрьский ‘of December’	N ADJ	Декабрь	Декабрь ‘December’ Декабрь ае ‘month of December’	N NP
ЯНВАРЬ	Январь ‘January’ Январский ‘of January’	N ADJ	Гыйнвар	Гыйнвар ‘January’ Гыйнвар ае ‘month of January’ Январь ‘January’ Январь ае ‘month of January’	N NP N NP
ФЕВРАЛЬ	Февраль ‘February’ Февральский ‘of February’	N ADJ	Февраль	Февраль ‘February’ Февраль ае ‘month of February’	N NP

Because RuThes concepts assemble ontological synonyms, RuThes lexical entries bring together words of different parts of speech. Therefore in standard cases a Russian synset joins a noun (often we use it as a concept name) and a relative adjective derived

from the noun (Table 5; only core items of synsets are represented). In Tatar, like in other Turkic languages, there are no original relative adjectives (and existing ones are borrowed from European or Oriental languages), so in many cases TatThes synsets are composed of items of the same part of speech, mainly of nouns. This circumstance greatly facilitates cleaning thesaurus synsets data for wordnet developing.

Table 5. Typical arrangement of Russian and Tatar Thesaurus synsets

Basic lexical entries of Russian concept	Part of speech of Russian words	Basic lexical entries of Tatar concept	Part of speech of Tatar words
Река ‘river’	N	Елга ‘river’	N
Речной ‘of river, fluvial’	ADJ		
Факультет ‘faculty’	N	Факультет ‘faculty’	N
Факультетский ‘of faculty’	ADJ		
Преподаватель ‘teacher	N	Укытучы ‘teacher’	N
Преподавательский ‘of teacher’	ADJ		
Больница ‘hospital’	N	Хастаханә ‘hospital’	N
Больничный ‘of hospital’	ADJ	Сырхауханә ‘hospital’	N

So the core of TatThes is made up of nouns and noun phrases (69% of total number of concepts). At the moment semantic relations between nouns mapped in the thesaurus, are necessary and sufficient to convert the Tatar thesaurus data into the wordnet format.

5 Tatar Translation of RuWordNet

In this section we describe Tatar translation of RuWordNet.

At first stage we performed automatic translation of RuWordNet resource with the help of the Russian-Tatar dictionary edited by F.A. Ganiev.

The next main task was manual verification of automatically obtained data. Using the data on hyponyms and hyperonyms, as well as the glossary, we checked the word meaning since the priority was not to evaluate the correct translation of individual words, but to the translation of the concepts of the original words into the target language. By analyzing and editing the text input in the Tatar language, one can see the following language situations (cases):

1) Noun synsets in the Russian language contain items derived from words of different parts of speech, for example, deverbal nouns naming actions and states. Words of different meaning and derivation models may be translated into Tatar differently. For example, often Russian deverbal nouns are conveyed in Tatar as verbal nouns – a hybrid grammatical class sharing some features of nouns and verbs (verbal nouns are the standard way to fix verbs in Tatar dictionaries). As a result, Russian noun synsets may correspond to Tatar synsets contacting items of dissimilar grammatical classes:

- *величие* (ru) 'greatness' - *бөеклек, олылык* (tat) – nouns in both languages;
- *бездействие* (ru) 'inaction' - *бер нәрсә дә эшләмәү* (one + what + PART + do-NEG, VN), *чара күрмәү* (measure + see-NEG, VN) (tat) – a noun in Russian and phrases with a verbal noun as a node word in Tatar;
- *гегемония* (ru) 'hegemony'/ - *гегемония, җитәкчелек итү* (leader-NMLZ + do), *өстенлек* (tat) – in the Tatar parts are nouns and a verbal noun phrase;
- *вескость* (ru) 'weightiness, validity' - *авыр булу* (heavy +be-VN*), саллы булу* (weighty + be-VN) (tat) - verbal noun phrases in Tatar.

Here and in examples below only the Tatar items with the grammatical structure differing from Russian correspondences are glossed.

2) There are many words (about 375) translated into the Tatar language by using a descriptive construction because these words are not presented in Tatar dictionaries:

- *коренник* (ru) 'shaft-horse'– *төпкә җигелгән ат* (bottom-DIR + harness- PASS, PCP_PS + horse) (tat);
- *выскочка (ru)* 'upstart' - *сикергәк, ялагай, сәнәктән көрәк булган кеше* (pitchfork-ABL + shovel + be-PCP_PS + man) (tat) – nouns and a noun phrase in Tatar.

Such descriptive phrases can be divided into 4 groups, depending on the lexical meaning and source word parts:

A) Root words that do not have a corresponding version in the Tatar language due to the fact that these concepts are not characteristic of the mode of life and the culture of this people. E.g.,

- *именинник* (ru) 'birthday boy' - *исем бәйрәмен* (name + fete-POSS-3, ACC + perform-VN, PCP_PR + man) (tat);
- *клюка (*ru) 'crooked top stick' – *кәкре башлы таяк* (crooked + head-ATTR_MUN + stick) (tat);

B) Terms and concepts that do not have equivalents in the Tatar language, transferred borrowed-words and/or descriptive phrases: *дротик* (ru) 'dart' – *дротик,* 'dart' – *кыска саплы сөңге* (short + handle-ATTR_MUN + spear) (tat).

C) Compound words that do not have equivalents identical in structure in the target language. E.g. many Russian two root words are conveyed in Tatar by means of compounds:

- *водосток* (ru) 'gutter' – *су агып төшә торган торба* (water + pour-CONV +go down-PCP_PR + be-PCP_PS + tube) (tat)*;*
- *двустволка* (ru) 'shotgun' – *ике көпшәле мылтык* (two +barrel- ATTR_MUN + gun) (tat);
- *естествоиспытатель* 'naturalist' – *табигать фәннәре белгече* (nature + science-PL, POSS_3, specialist-POSS_3) (tat).

D) Many Tatar synsets contain in addition phrases with a hypernym, in particular, names of months, plants, trees, nationalities, and other classes:

- *январь* (ru) 'January' – *гыйнвар, гыйнвар ае* (January + month-POSS_3) (tat);
- *вяз* (ru) 'elm'– *карама, карама агачы* (elm + tree-POSS_3) (tat);
- *липа* (ru) 'linden' – *юкә, юкә агачы* (linden+ tree-POSS_3) (tat);
- *японец* (ru) 'Japanese man' – *япон, япон кешесе* (Japanese + man-POSS_3) (tat);
- *девочка (ru)* 'female child' – *кыз бала (*girl + child*)* (tat);
- *иноходец* (ru) 'pacer horse' – *юрга, юрга ат* (pacer + horse) (tat).

3) The Tatar language has no morphological category of grammatical gender, and to convey this category, lexical means are used. So in Tatar synsets corresponding to Russian synsets gathering words denoting females, words specifying the age and the marital status is added to such text entries for the Tatar language (*кыз*'girl' or *хатын* 'woman'). This applies to translation names of nationalities, professions, social status, etc.:

- активистка (ru) 'activist woman or girl' – активистка, актив хатын, актив кыз;
- караимка (ru) 'karaite' woman or girl – караим хатыны, караим кызы;
- купальщица (ru) bather woman or girl' – су коенучы хатын, су коенучы кыз;
- манекенщица (ru) 'mannequin, fashion model woman or girl' – манекенчы хатын, манекенчы кыз.

4) As we mentioned above, a problematic area to translate is synsets in Russian for which there are no corresponding concepts in the Tatar culture. A significant portion of them make up the concepts of Orthodox Christianity absent in Islam (the latter is the religion of the most part of Tatars). We found currently 32 such items. For example:

- *ересь* (ru) 'heresy' – *ересь* (tat);
- *молебен* (ru) 'prayer service' – *молебен* (tat);
- *миропомазание* (ru) 'anointing'– *миро белән майлап чукындыру* (chrism + with + oil-CONV + baptize-VN) (tat).

Religious items are translated in three ways:

A) by using words borrowed from Russian (however the origin of words may be different, for example Greek);
B) by using explanatory translation;
C) by using words denoting close concepts from the Muslim terminology.

6 Database of Semantic Classes of Verbs

In this section, we describe TatVerbClass, a database of semantic classes of Tatar verbs [19].

The classification scheme is based on the following parameters of verbal lexemes:

1. thematic feature, linked with the verb's thematic class, which allows us to mark up the verb's denotation sphere;

2. grammatical feature, linked with the valency changing operations of voice affixes (possibility of producing grammatical voice derivatives and particular meanings of voice forms);
3. syntactic feature, related to the allowable predicate-argument structure and thematic roles of arguments;
4. derivational feature, related to the verb's derivation pattern (grammatical class of the stem, derivational meaning of the verb forming affix).

Each verb is provided with a semantic tag (or with a set of the latter), there have been distinguished 59 basic semantic (ontological) classes, such as movement verbs, speech verbs, etc.). Semantic classes may join items with dissimilar individual meanings, grammatical properties, syntactic behavior, etc. So in a semantic class we distinguish a set of individual subclasses including verbs of similar structure, features and behavior.

In spite of rather formal criteria when selecting subclasses (ability to produce the same grammatical voice derivatives and sharing argument structure of verbs are in the foreground), the words of similar meaning fall into the same subclass. In most cases subclasses join synonyms, antonyms and hyponyms related to the same hypernym (see Tables 6, 7 – examples of subclasses of the physiological verbs).

Table 6. Subclass of verbs related to the hypernym 'to feel sensations in the body'

Verbs	English translation (the main senses)	Thematic tags in DB
авырт	to ache (on physical pain)	t:physiol, t:perc
сызла	to ache (on intensive pain)	t:physiol, t:perc
әрне	to ache (on acute pain)	t:physiol, t:perc
ачыт	to ache (on burning pain)	t:physiol, t:perc
кычыт	to itch and tickle	t:physiol, t:perc
кызыш	to feel fever	t:physiol, t:perc
кымыржы	to itch	t:physiol, t:perc
чымырда	to feel goosebumps	t:physiol, t:perc
чемердә	to feel goosebumps	t:physiol, t:perc

All the verbs in Table 6 share the features:

– all the verbs have a basic meaning 'to feel some sensations in the body/part of the body' and they are provided with the same semantic tags;
– all the verbs are intransitive and express a state;
– all the verbs may have causative derivatives and can not produce passive and reciprocal derivatives;
– as a standard syntactic subject they have nouns denoting body or parts of body.

Table 7. Verbs denoting disease states

Verbs	English translation (the main senses)	Thematic tags in DB
авыр	to be ill	t:physiol:disease
чирлә	to be ill	t:physiol:disease
сырхала	to be ill	t:physiol:disease
хастала	to be ill	t:physiol:disease

Another example is a subclass of verbs denoting disease states (Table 7), where all the items are synonyms.

The verbs самсыра 'to ill, be down at health' despite the semantic affinity with the verbs from Table 7, is set outside the scope of the subclass, because it does not take arguments with белән 'with' postposition, unlike the verbs represented in the Table 7.

7 Conclusion

In this paper, we described the methodology for constructing TatWordNet on the base of the three resources: Russian-Tatar Social-Political Thesauru (TatThes), Tatar translation of RuWordNet and the Database of Semantic Classes of Tatar Verbs (TatVerbClass). Currently, TatWordNet consists in the three components, corresponding to these sources. Our immediate goal is to complete development of these components and merge them into single unified resource.

After that we are planning to continue our research in the following directions:

1. checking the quality and representativeness of the data obtained through comparison with frequency dictionary created on the basis of the "Tugan tel" Tatar National corpus and adding missing senses;
2. comparing the core data of TatWordNet with the core of Princeton WordNet and adding missing senses;
3. developing hierarchies for adjectives and other parts-of-speech.

Our ultimate goal is to publish Tatar Wordnet on the Linguistic Linked Open Data cloud [20] and integrate it to the Global WordNet Grid [21] via the Collaborative Interlingual Index.

Acknowledgments. This work was funded by Russian Science Foundation according to the research project no. 19-71-10056.

References

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

2. Fellbaum, C.: WordNet. In: Poli, R., et al. (eds.) Theory and Applications of Ontology: Computer Applications, pp. 231–243. Springer, Heidelberg (2010). https://doi.org/10.1007/978-90-481-8847-5_10
3. Vossen, P. (ed.): EuroWordNet: General Document. University of Amsterdam (2002). http://www.illc.uva.nl/EuroWordNet/docs.html
4. Galieva, A., Kirillovich, A., Loukachevich, N., and Nevzorova, O.: Towards a Tatar WordNet: a methodology of using tatar thesaurus. In: Elizarov, A., et al. (eds.) Selected Papers of the XXI International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019). CEUR Workshop Proceedings, vol. 2523, pp. 316–324. CEUR-WS (2019)
5. Çetinoğlu, Ö., Bilgin, O., Oflazer, K.: Turkish Wordnet. In: Oflazer, K., Saraçlar, M. (eds.) Turkish Natural Language Processing. TANLP, pp. 317–336. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90165-7_15
6. Bilgin, O., Çetinoğlu, Ö., Oflazer, K.: Building a Wordnet for Turkish. Romanian J. Inf. Sci. Technol. **7**(1–2), 163–172 (2004)
7. Tufis, D., Cristea, D., Stamou, S.: BalkaNet: aims, methods, results and perspectives. A general overview. Romanian J. Inf. Sci. Technol. **7**(1–2), 9–43 (2004)
8. Ehsani, R.: KeNet: a comprehensive Turkish wordnet and using it in text clustering. Ph.D. thesis. Işık University (2018)
9. Ehsani, R., Solak, E., Yildiz, O.T.: Constructing a WordNet for Turkish using manual and automatic annotation. ACM Trans. Asian Low-Resource Lang. Inf. Process. **17**(3), Article No. 24 (2018). https://doi.org/10.1145/3185664
10. Bond, F., Foster, R.: Linking and extending an Open Multilingual WordNet. In: Schuetze, H., Fung, P., Poesio, M. (eds.) Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Volume 1: Long Papers, pp. 1352–1362. ACL (2013)
11. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012). https://doi.org/10.1016/j.artint.2012.07.001
12. Loukachevitch, N., Dobrov, B.: RuThes linguistic ontology vs. Russian wordnets. In: Orav, H., Fellbaum, C., Vossen, P. (eds.) Proceedings of the 7th Conference on Global WordNet (GWC 2014), pp. 154–162. University of Tartu Press (2014)
13. Loukachevitch, N.V., Dobrov, B.V., Chetviorkin, I.I.: RuThes-Lite, a publicly available version of thesauru of Russian language RuThes. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", pp. 340–349. RGGU (2014)
14. Galieva, A., Kirillovich, A., Khakimov, B., Loukachevitch, N., Nevzorova, O., Suleymanov, D.: Toward domain-specific Russian-Tatar thesaurus construction. In: Bolgov, R., et al. (eds.) Proceedings of the International Conference on Internet and Modern Society (IMS-2017), pp. 120–124. ACM Press (2017). https://doi.org/10.1145/3143699.3143716
15. Galieva, A., Nevzorova, O., Yakubova, D.: Russian-Tatar socio-political thesaurus: methodology, challenges, the status of the project. In: Mitkov, R., Angelova, G. (eds.) Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017), pp. 245–252. INCOMA Ltd. (2017). https://doi.org/10.26615/978-954-452-049-6_034
16. Kirillovich, A., Nevzorova, O., Gimadiev, E., Loukachevitch, N.: RuThes Cloud: towards a multilevel linguistic linked open data resource for Russian. In: Różewski, P., Lange, C. (eds.) KESW 2017. CCIS, vol. 786, pp. 38–52. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69548-8_4
17. Loukachevitch, N.V., Lashevich, G., Gerasimova, A.A., Ivanov, V.V., Dobrov, B.V.: Creating Russian WordNet by conversion. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue", pp. 405–415. RGGU (2016)

18. Loukachevitch, N., Lashevich, G., Dobrov, B.: Comparing two thesaurus representations for Russian. In: Bond, F., Kuribayashi, T., Fellbaum, C., Vossen, P. (eds.) Proceedings of the 9th Global WordNet Conference (GWC 2018), pp. 35–44. GWA (2018)
19. Galieva, A., Vavilova, Z., Gatiatullin, A.: Semantic classification of Tatar verbs: selecting relevant parameters. In: Čibej, J., Kosem, I., and Krek, S. (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (Euralex 2018), pp. 811–818. Ljubljana University Press (2018)
20. Cimiano, P., Chiarcos, C., McCrae, J.P., Gracia, J.: Linguistic Linked Open Data Cloud. In: Cimiano, P., et al. (eds.) Linguistic Linked Data: Representation, Generation and Applications, pp. 29–41. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-30225-2_3
21. Vossen, P., Bond, F., McCrae, J.P.: Toward a truly multilingual Global Wordnet Grid. In: Barbu Mititelu, V., et al. (eds.) Proceedings of the 8th Global WordNet Conference (GWC 2016), pp. 419–426. GWA (2016)