Towards a Tatar Wordnet: a Methodology of Using Tatar Thesaurus

Alfiya Galieva[1], Alexander Kirillovich[2],
Natalia Loukachevich[3] and Olga Nevzorova[1,2]

[1] Tatarstan Academy of Sciences, Kazan, Russia
[2] Kazan (Volga region) Federal University, Kazan, Russia
[3] Lomonosov Moscow State University, Moscow, Russia

amgalieva@gmail.com, alik.kirillovich@gmail.com,
louk_nat@mail.ru, onevzoro@gmail.com

Abstract. For wordnet developing for a new language, the key problem is to find original resources that contain enough lexical data of the language in an appropriate format. This article discusses the structure, methodology of compilation and the current state of the bilingual Russian-Tatar Social-Political Thesaurus, which can serve as an initial resource for building the Tatar Wordnet. This thesaurus reflects the logical-semantic organization of lexical elements (synonymous, generic, and some other relationships) at the conceptual and lexical levels. Mainly, we focus on building synsets for *nouns* (single *nouns and noun phrases*).

Keywords: Tatar language, WordNet, Thesaurus, Linguistic ontology, Socio-political terminology.

1 Introduction

A great hindrance to develop linguistic ontologies for a new language and conceptual modeling is the lack of original lexicographic resources containing full and relevant linguistic data description.

Success of Princeton WordNet has determined emergence of wordnets and wordnet-like projects for different languages and multilingual wordnets. In wordnet building developers often use the Expand Model (Vossen 2002: 52) when available wordnets that serve as mapped linguistic relations between the items and ready synsets of a source language are translated using bilingual dictionaries into equivalent synsets in the target language. Most of the wordnets existing for today are implemented by translating Princeton English WordNet. The alternative approach is very laborious, time-consuming and difficult to implement, based on compiling synsets and mapping semantic relations between word senses directly on the data of the language for which a wordnet is developed. In this case good dictionaries of synonyms and other semantic dictionaries are required.

So the possibility of developing wordnets is largely determined by the presence of bilingual dictionaries or fairly complete descriptions of the semantic system of the language.

The absence of large English-Tatar dictionaries (the available ones are of very limited volume and may be used only for education purposes) makes it impossible to use the Expand Model to Tatar wordnet development, as well as absence of Tatar semantic dictionaries makes it almost impossible to develop original Tatar wordnet.

The objective of this article is to describe the methodology for constructing a Tatar wordnet based on a lexical resource such as the Tatar social-political thesaurus. This approach allows you to directly use the data of the thesaurus, primarily a set of synsets and relationships between synsets.

The body 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 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 (Çetinoğlu, et al, 2018; Bilgin, et al, 2004) has been created at Sabancı University as part of the BalkaNet project (Tufis, et al, 2004). 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 the Turkish Wordnet is about 15,000 synsets.

Another Turkish wordnet is the KeNet (Ehsani, 2018; Ehsani, et al, 2018). This wordnet was built on the basis of modern Turkish dictionaries. A bottom-up approach was used to build this resource. Based on dictionaries, words were selected and then they were manually grouped into synsets. The relationships between words have been automatically extracted from dictionary definitions and then these relationships have been created between synsets. The size of this resource is about 113,000 synsets.

Unfortunately, the 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 (Bond, 2013) 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, Uzbek. The Tatar wordnet contains a total of 550 concepts, which cover 5% of the PWN core concepts.

The BabelNet (Navigli and Ponzetto, 2012) 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 problems.

Thus, the development of a qualitative 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 Its Current State

The conceptual model of the Tatar socio-political thesaurus (hereinafter referred to as TatThes), the general principles of displaying linguistic data are taken from the RuThes project (http://www.labinform.ru/pub/ruthes/) (Loukachevitch and Dobrov, 2014; Loukachevitch, Dobrov and Chetviorkin, 2014). The RuThes thesaurus build as is a hierarchical network of concepts with attributed lexical entries for automatic text processing.

In the 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). The 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 the RuThes concept is represented as a set of synonyms or near-synonyms (plesionyms). The 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 2. 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

The TatThes is based on the list of concepts of the 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 the new concepts of the intermediate level. So an important step is to check up the parallelism of conceptual structures between the languages.

The TatThes is mainly being compiled by manual translation of terms from the 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 (Galieva, Kirillovich, et al., 2017). 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 distinguishing feature of the contemporary Tatar lexicon is a great deal of absolute synonyms of different origin in and structure, the main cause of the phenomenon is language contacts (Galieva, Nevzorova, et al., 2017; Galieva, 2018).

The 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 (Kirillovich, et al, 2017). Currently the TatThes contains 10,000 concepts, and 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 the WordNet-like structure, and Russian wordnet (RuWordNet) has been generated (Loukachevitch, et al, 2016; Loukachevitch, et al., 2018). The conversion included two main steps:

1. the automatic subdivision of the RuThes text entries into three nets of synsets according to parts of speech;
2. the 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 be serve as an initial basis for wordnet building by 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 the 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 the 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 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 wordnet, are not all shared by all lexical categories, so thesaurus data converting into wordnet format require dissimilar ways for different parts of speech.

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

Semantic relations in 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 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 not exist without the existence of the other (for example, a concept SUMMIT MEETING needs existing 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 nouns concepts into wordnet are presented in the TatThes.

The core of the 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).

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

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

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

(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’).

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

Rus 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

In cases when such a usage is conventionalized and corpus data evidences that the usage has a high frequency, we include such hyponym-hypernym items into a list of lexical entries of a concept. Such manner of designating is a feature of using topo-

nyms 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.

Because the RuThes concepts assemble ontological synonyms, the RuThes lexical entries bring together words of different part of speech. Therefore, in standard case a Russian synset joins a noun (often we use it as a concept name) and a relative adjective derived from a noun (Table 5; only core items of synsets are represented). In Tatar, like in other Turkic languages, there is no original relative adjectives (and existing ones are borrowed from European or Oriental languages), so in many cases the 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 a Russian concept	Part of speech of Russian words	Basic lexical entries of a 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 the TatThes is made up of nouns and noun phrases (69% of total number of concepts). At the moment semantic relations between nouns mapped in thesaurus, are necessary and sufficient to convert Tatar thesaurus data into the wordnet format.

4 Conclusion

When building a wordnet for a new language, in particular, for a low-resource one, a crucial issue is searching for appropriate sources. We are planning to use data of the TatThes as a base resource for developing Tatar wordnet.

The TatThes is being compiled by manual translation of terms from the RuThes into Tatar, with searching Tatar equivalents used in real texts, so the thesaurus contains relevant lexical data. In the TatThes each concept is linked with a set of language expressions (single words or multiword expressions) which refer to the concept in texts – lexical entries.

The analysis of thesaurus data shows that the bulk of the thesaurus synsets are formed around nouns or noun phrases. A mapping semantic relations of nouns in the thesaurus reproduces a mapping semantic relations in wordnets.

Future work includes adding material of verbs and other parts of speech. Also we are planning to develop some automatic approaches to mining terms and to asses the Tatar terminology coverage in Thesaurus on Tatar socio-political texts data.

Acknowledgements

This work was partially funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017 and partially by a subsidy assigned to the Institute of Applied Semiotics of the Tatarstan Academy of Sciences for the state assignment.

References

1. Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer: Building a Wordnet for Turkish. Romanian Journal of Information Science and Technology **7** (1–2), 163–172 (2004).
2. Francis Bond and Ryan Foster: Linking and Extending an Open Multilingual Wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), 1352–1362. ACL (2013).
3. Özlem Çetinoğlu, Orhan Bilgin, and Kemal Oflazer: Turkish Wordnet. In: K. Oflazer and M. Saraçlar (eds). Turkish Natural Language Processing. Springer (2018). doi:10.1007/978-3-319-90165-7_15
4. Razieh Ehsani. KeNet: A Comprehensive Turkish Wordnet and Using It in Text Clustering. PhD Thesis. Işık University (2018).
5. Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz: Constructing a WordNet for Turkish Using Manual and Automatic Annotation. ACM Transactions on Asian and Low-Resource Language Information Processing **17** (3), Article No. 24 (2018). doi:10.1145/3185664
6. Alfiya Galieva, Alexander Kirillovich, Bulat Khakimov, Natalia Loukachevitch, Olga Nevzorova, and Dzhavdet Suleymanov: Toward Domain-Specific Russian-Tatar Thesaurus Construction. In: R. Bolgov, N. Borisov, et al. (eds.) Proceedings of the International Conference on Internet and Modern Society (IMS-2017), pp. 120–124. ACM Press, New York (2017). doi:10.1145/3143699.3143716
7. Alfiya Galieva, Olga Nevzorova, and Dilyara Yakubova: Russian-Tatar Socio-Political Thesaurus: Methodology, Challenges, the Status of the Project. In: R. Mitkov and G. Angelova (eds.) Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017), pp. 245–252. INCOMA Ltd., Varn (2017). doi:10.26615/978-954-452-049-6_034
8. Alfiia Galieva: Synonymy in Modern Tatar Reflected by the Tatar-Russian Socio-Political Thesaurus. In: J. Čibej, V. Gorjanc, I. Kosem and S. Krek (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (Euralex 2018), pp. 585-994. Ljubljana University Press (2018).
9. Alexander Kirillovich, Olga Nevzorova, Emil Gimadiev, and Natalia Loukachevitch: RuThes Cloud: Towards a Multilevel Linguistic Linked Open Data Resource for Russian. In: P. Różewski and C. Lange (eds.) Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017). Communications in Computer and Information Science **786**, 38–52. Springer (2017). doi:10.1007/978-3-319-69548-8_4
10. Natalia Loukachevitch and Boris Dobrov: RuThes Linguistic Ontology vs. Russian Wordnets. In: H. Orav, C. Fellbaum and P. Vossen (eds.) Proceedings of the 7th Conference on Global WordNet (GWC 2014), pp. 154–162. University of Tartu Press (2014).

11. Loukachevitch, N.V., Dobrov, B.V., and 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)
12. Loukachevitch, N.V., Lashevich, G., Gerasimova, A.A., Ivanov, V.V., and 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).
13. Natalia Loukachevitch, German Lashevich, and Boris Dobrov: Comparing Two Thesaurus Representations for Russian. In: F. Bond, T. Kuribayashi, C. Fellbaum and P. Vossen (eds.) Proceedings of the 9th Global WordNet Conference (GWC 2018), pp. 35–44. Global Wordnet Association (2018).
14. Roberto Navigli and Simone Paolo Ponzetto: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, December 2012, 217–250 (2012).
15. Tufis, D., Cristea, D., and Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian Journal of Information Science and Technology **7** (1–2), 9–43 (2004).
16. Piek Vossen (ed). EuroWordNet: General Document (2002).