Alfiya Galieva
Tatarstan Academy of Sciences, Tatarstan
amgalieva@gmail.com

Olga Nevzorova
Tatarstan Academy of Sciences, Tatarstan
onevzoro@gmail.com

Dzhavdet Suleymanov
Tatarstan Academy of Sciences, Tatarstan
dvdt.slt@gmail.com

Keywords: bilingual thesaurus, concept, terminology, the Tatar language, socio-political vocabulary, lexical synonymy.

Tatar Socio-Political Terminology in a Bilingual Thesaurus

Introduction

Tatar belongs to the Turkic language family, and is regarded as the second language in Russia – according to its geographic distribution and number of speakers (Information materials, 2010). Since 1992, the Tatar language has the status of the official language of the Republic of Tatarstan alongside with the Russian language, which is protected by the Constitution of Tatarstan and the Law on the Language. The current requirement to use the Tatar language in the government and administration conditions the need to translate the official documents and information materials. In particular, the texts of laws of the Republic of Tatarstan and other regional legal acts are to be published both in Tatar and Russian; state authorities and institutions are to use both state languages. Therefore, building new lexicographic resources that contain present-day linguistic data related to the socio-political domain, fills the needs of Tatar society.

However, despite the official status, Tatar can be referred to the so called low-resource languages, because it has few lexicographic resources and lacks relevant description in the existing grammar books (being the latter a shared challenge in a large number of non-Indo-European languages).

Very few of the world's languages currently enjoy the latest achievements of modern NLP technologies such as text processing, information retrieval, or machine translation. Very few have managed to assemble the basic resources that are necessary for building advanced end-user technologies, among them monolingual and bilingual corpora, machine-readable dictionaries, thesauri, part-of-speech taggers, morphological analyzers, parsers (Scannell 2007). Building new lexicographic projects for low-resource languages is a task of current importance both of theoretical interest and in terms of its practical implementation. In many cases the main obstacle to it is the lack of parent resources, the lack of required volumes of lexical data of a language in an appropriate format, and the problem is more significant if the language under study is morphologically rich (Tachbelie, Abate, Besacier 2011; Doborjginidze, Lobzhanidze 2016).

Researchers follow three main paradigms to create resources for underresourced language: (1) using crowdsourcing to produce a small resource rapidly and relatively cheaply; (2) using existing gold-standard collections and machine translation tools to translate an existing gold-standard resource that is easy to create but of lower quality, with more "noise" and (3) using manual effort with appropriately skilled human participants to create a resource that is more expensive but of high quality (El-Haj, Kruschwitz, Fox 2015).

Developers of new wordnets often use the so called *Expand Model* (Vossen 1997; Vossen 2002) when available wordnets serve as a network of mapped linguistic relations between the items, and synsets of a source language are translated using bilingual dictionaries into equivalent synsets in the target language, wherein the level of automation of the process may be different. Most of the wordnets existing for today are created by translating the English Princetone WordNet (Miller 1995; Fellbaum 2010). The work on Turkish wordnet started with translating the base concepts of the EuroWordNet project (Bilgin, Çetinoğlu, Oflazer 2004).

We use the above mentioned *Expand* method to develop a bilingual Russian-Tatar Thesaurus, with the Tatar part built upon the concepts of the Russian RuThes thesaurus (http://www.labinform.ru/pub/ruthes/). The Thesaurus comprises lexical data related to the socio-political (Galieva, Nevzorova, Yakubova 2017) and the IT (Khakimov 2018) subject areas. This paper focuses on the main aspects of compiling the Socio-Political Thesaurus (Section 2), the current state of the project and the main results (Section 3), and discusses some crucial linguistic issues on representing some data in the Thesaurus (Section 4).

2. Methodology issues of compiling Thesaurus

The conceptual model of the Russian-Tatar Thesaurus and general principles of mapping linguistic data are borrowed from the Russian RuThes project (http://www.labinform.ru/pub/ruthes/). RuThes thesaurus is built as a hierarchical network of concepts with attributed lexical entries which name concepts in a language. The main units in RuThes are concepts – mental objects that designate classes of entities distinguished by human beings. 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 this 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 (Loukachevitch 2011; Loukachevitch, Dobrov 2014). RuThes was developed as a resource for automatic text processing and is used as a corporate resource for conceptual information retrieval, automatic query expansion, automatic rubrication and annotation, and automatic clustering (Loukachevitch 2011; Loukachevitch, Dobrov 2015).

So Russian-Tatar bilingual Thesaurus maintains the basic conceptual structure of RuThes, nevertheless, structural modifications, demanded by Tatar language data, are admitted. This modification are related to adding single concepts on a branch, as well as the adding new thematic branches reflecting the specific features of the Tatar language and culture. As a result, each part of the Thesaurus – the Russian and Tatar ones – represents a unique language-internal system of lexicalizations. At the same time, the languages are interconnected, so it is possible to proceed from the concepts and lexical entries in one language to the corresponding items in the other.

The methodology of compiling the Tatar part of the thesaurus includes the following main steps:

1. Search for equivalents (corresponding words and multiword expressions) which are actually used in Tatar as a translations of the Russian items.

2. Adding new concepts which refer to the important issues in terms of the socio-political 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).

3. Revising relations between the concepts, defining the place of each new concept in the hierarchy of the existing ones and, if necessary, adding new concepts of the intermediate level Galieva, Nevzorova, Yakubova 2017).

It is important to keep in mind that conceptual structures may vary from language to language. So an important step is to check up the parallelism of conceptual structures between the languages.

The compilation of the thesaurus is carried out by manual translation of terms from the Russian RuThes into Tatar. This translation encounters many hindrances, in particular, Russian-Tatar dictionaries are outdated in many respects and do not contain vocabulary designating present-day realities or offer obsolete translations (Galieva, Nevzorova, Yakubova 2017). Using machine translation, such as Yandex Translate (https://translate.yandex.com/) does not provide the required results either, because the available Russian-Tatar machine translation systems are trained on rather limited data.

Our goal is to detect and to fix the actually used Tatar socio-political terminology, so we browsed manually large arrays of official documents and media texts in Tatar to find the correct items. In many cases, searching manually for equivalents in the target Tatar language was a time-consuming task, and the use of data from the Tatar corpora was very helpful. We made use of:

1. “Tugan Tel” Tatar National Corpus (http://tugantel.tatar/?lang=en);

2. Socio-Political Subcorpus of Tatar National Corpus (http://tugantel.tatar/corpus/op/);

2. Corpus of Written Tatar (http://www.corpus.tatar/en).

The data gathered from these corpora allowed us to acquire reliable information on the frequency of use and the typical contextual environment of Tatar words and multiword expressions denoting the socio-political realities, which, in turn, allowed to check the meanings of lexical items and to determine their place in the hierarchy of concepts. So, using corpus data became a necessary stage for collecting lexical data as well as disclosing semantic relations between terms. We also used corpus data as empirical material to study current processes and trends in Tatar socio-political vocabulary (to trace out ways of translating the Russian terms, to find new loanwords and semantic calques, to examine synonymy of terms, etc.), and this information became a reliable guideline to suggest our translation variants of terms in cases when Tatar variants of term were not found in any source.

The project of building the Russian-Tatar Thesaurus is aimed at compiling the whole body of modern Tatar socio-political vocabulary in real use, therefore, we pay much attention to working on arrays of Tatar texts. Currently the Thesaurus covers the vocabulary related to state government, economy, social life, justice, warfare, culture, religion, sports and some other basic topics. It also comprises some general lexicon branches representing lexical items which can be found in a large number of various texts, regardless of their subject area.

The Thesaurus system is implemented as a web application (http://tattez.turklang.tatar/) and has an open and a closed part. Functionality of the open part includes search for and navigation on concepts and lexical entries. An authorized editor is provided with the options of adding, editing and removing concepts and lexical entries, for comparing the conceptual structures of the languages, and reestablishing relations between concepts.

3. Main results

Currently, the Thesaurus contains 10,000 concepts, and 6,000 of them are provided with lexical entries. The design of the Tatar component of the Thesaurus preserves the basic structure of the RuThez thesaurus, and the Tatar component overlies the list of original RuThez concepts.

Some individual concepts may link to numerous lexical entries. For example, Table 1 represents the concept PUBLIC SYSTEM in the Russian and Tatar parts of the Thesaurus, and the lexical entries of both parts are linguistic items designating the social (public) structure. Hereinafter we use uppercase to represent concept names and the lowercase to represent lexical entries of concepts.

Table 1. PUBLIC SYSTEM concept in the Russian-Tatar thesa-urus.

Concept name in Russian	*Lexical entries in Russian*	*Concept name in Tatar*	*Lexical entries in Tatar*
ОБЩЕСТВЕННАЯ СИСТЕМА	общественный строй, общественно-политическая система, общественное устройство, общественная система, социальный строй, общественно-политический строй, общественно-по-литическое устройство, социально-политическое устройство, социально-политический строй, политико-экономическая система, общественно-экономическое устройство, общественно-экономическая система, социально-экономическая система, социально-экономическое устройство, социально-экономический строй,	ИҖТИМАГЫЙ СИСТЕМА	иҗтимагый строй, җәмгыять строе, социаль строй, иҗтимагый-сәяси система, иҗтимагый-политик система, җәмгыять корылышы, җәмгыять төзелеше, иҗтимагый система, җәмгыять системасы, иҗтимагый-сәяси строй, иҗтимагый-сәяси корылыш, социаль-сәяси төзелеш, социаль-сәяси строй, иҗтимагый-икътисади система, социаль-экономик система, социаль-икътисади система, иҗтимагый-политик система, җәмгыять бәйләнешләре системасы, иҗтимагый бәйләнешләр системасы

	общественно-экономический строй, система общественных отношений, строй		

Table 2 contains brief information on concepts and lexical entries of the Russian and Tatar parts of the thesaurus. It only considers the concepts of the socio-political area (items of general lexicon are removed).

Table 2. Quantitative characteristics of basic thesaurus units.

Basic quantitative characteristics	*Russian*	*Tatar*
Number of concepts in the sample	5,830	5,830
Average number of lexical entries per concept	4.61	3.70
Average length of a lexical entry, in words	1.79	2.23

A larger number of lexical entries per concepts in the Russian part of the Thesaurus is mainly due to the peculiarities of the Russian grammar and word formation system which lead to the existence of more surface realizations of concepts (see the next section). The larger average length of Tatar lexical entries may be explained rather by the direction of translating (as we translate Russian lexical items into Tatar and not vice versa, in many cases we use elements of explanatory translation) than by proper linguistic reasons.

Table 3 represents the distribution of the most frequent Tatar word forms used in concept names and lexical entries; functional words are excluded from the sample. The data are given without lemmatization of items (so the variety of word forms is not considered, and the most frequently used word forms only are represented). As a matter of fact the data in Table 3 outline the main thematic blocks of the Tatar socio-political lexicon reflected in the Thesaurus.

Table 3. The most frequent wordforms used in Tatar Thesaurus

Word form	*Part of speech*	*English translation*	*Number in concept names*	*Number in lexical entries*
дәүләт	N	state, country, nation	130	235
хәрби	ADJ	military, martial	105	131
хезмәт	N	labor, work, duty	95	166
эш	N	work, job	68	157
суд	N	court of law	58	153
җинаять	N	crime	55	129
медицина	N	medicine	55	134
чыгару	V	to produce, to issue, to publish	52	101
уку	V	to study, to learn	50	101
кеше	N	man, human being	48	143
финанс	N, ADJ	finance, financial	48	28
саклау	V	to guard, to care, to protect	47	70
хокук	N	law, right	47	92
сәяси	ADJ	political	46	100
социаль	ADJ	social	46	139
органы	N, Poss_3	organ	45	105
федераль	ADJ	federal	45	69
хезмәте	N,	labor, work, duty	44	85
сәүдә	POSS_3	commerce, trade	43	73
халык	N	people	40	101
акча	N	money	39	32
иҗтимагый	N	social, public	39	108
халыкара	ADJ	international	39	49
	ADJ			

The quantitative distribution of 100 most frequent word forms is presented in Figure 1 as a word cloud; in this case, functional words, including auxiliary verbs and participles, were kept in the sample, so data differ from Table 3.

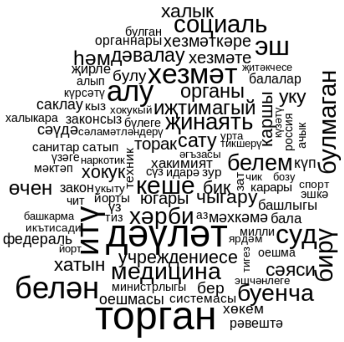

Figure 1. The most frequently used word forms among lexical entries.

More detailed discussion on some discrepancies between the Russian and the Tatar languages in terms of cross-linguistic mapping of vocabulary, and some features of the present-day Tatar socio-political terminology is given in the section below.

4. Discussion

In this section we focus on main linguistic issues concerning representing data in the Thesaurus and language interaction.

4.1. Fixing ontological synonyms.

The importance of the issue of semantic equivalence of cross-linguistic synonyms for various aspects of general linguistics, lexicological, and lexicographic research and related activities is on the focus of researchers (Margalitadze, Meladze 2016).

The RuThes thesaurus framework is based on compiling the so called ontological synonyms, which enables to designate linguistic items denoting concepts regardless of their surface realization. Such an approach allows us to embrace within the same concept words of different parts of speech (like the noun *stabilization,* the verb *to stabilize, the* participle *stabilized,* etc.) (Loukachevich, Dobrov 2014).

The advantage of this approach is easy to show on the example of relative adjectives. Relative adjectives name attributes by designating their relation to the subject, action or to another attribute, and in

Russian relative adjectives make up the bulk and the constantly replenished class of adjectives (Shvedova 1980: 539). In Tatar, like in other Turkic languages, relative adjectives are loan words that were borrowed from different languages (Russian, Arabic, Persian, etc.), and many Russian adjectives in the Tatar language have no equivalents of the same part of speech. In RuThes, as a rule, Russian nouns and their adjectival derivatives occur as lexical entries of the same concept, while Tatar concepts may or not include include relative adjectives, depending on their existence. Table 4 illustrates this approach with examples of some Russian and Tatar concepts and lexical entries, where Russian concepts are instantiated by nouns and adjectives, whereas Tatar concepts are instantiated by nouns only.

Table 4. Arrangement of Russian and Tatar concepts: representing relative adjectives

Basic lexical entries of Russian concepts	*Part of speech of Russian words*	*Basic lexical entries of Tatar concepts*	*Part of speech of Tatar words*
факультет 'faculty' факультетский 'of faculty'	N ADJ	факультет 'faculty'	N
преподаватель 'teacher, instructor' преподавательский 'of teacher, of instructor'	N ADJ	укытучы 'teacher, instructor'	N
больница 'hospital' больничный 'of hospital'	N ADJ	хастаханә 'hospital' сырхауханә 'hospital'	N

Lack of relative adjectives in Tatar influences the creation of multiword terms. In particular, in many cases Russian *adjective + noun* two-component terms correspond to *noun + noun* phrases in Tatar, and the grammatical pattern of such items may be different. For example, Table 5 represents basic ways of translating Russian *adjective + noun* phrases into Tatar.

Table 5. Examples of Russian *adjective + noun* phrases and ways of translating them into Tatar.

Russian unit	*Corresponding Tatar unit*	*The structure of the Tatar unit*	*English translation*
государственный контракт	дәүләт контракты	N + N, POSS_3	'government contract'
детский сад	балалар бакчасы	N, PL + N, POSS_3	'kindergarten'
авторское право	авторлык хокукы	N, NMLZ +N, POSS_3	'copyright'
сексуальное меньшинство	сексуаль азчылык	ADJ + N	'sexual minority'
именная стипендия	исемле стипендия	N, COMIT + N, PL	'nominal scholarship'

Presence of feminine correlates of the terms denoting persons in Russian also increases the amount of lexical entries in the Russian part of the Thesaurus (see examples in Table 6). In Russian, such feminine variants are formed by means of special suffixes and may be produced from words denoting job titles, positions, lifestyle, human features, etc. In Tatar such gender marking suffixes are absent, very few loanwords have feminine correlates, and gender marking may be expressed in texts explicitly by means of words denoting women or girls.

Table 6. Arrangement of Russian and Tatar concepts: representing feminine correlates of person names

Basic lexical entries of Russian concepts	*Part of speech of Russian words*	*Basic lexical entries of Tatar concepts*	*Part of speech of Tatar words*
капиталист 'capitalist' капиталистка 'woman capitalist'	N N, F	капиталист 'capitalist'	N
стажер 'trainee' стажерка 'woman trainee	N N, F	стажер 'trainee'	N
пенсионер 'pensioner' пенсионерка 'woman pensioner'	N N, F	пенсионер 'pensioner'	N

поэт 'poet' поэтесса 'poetess'	N N, F	шагыйрь 'poet' шагыйрә 'poetess'	N N, F
Лжец 'liar' лгунья 'woman liar'	N N, F	ялганчы 'liar'	N N

So, ontological synonyms are the most appropriate means to represent cross-linguistic equivalents (correspondences) in the Socio-Political Thesaurus, because such approach allows us to join units of the same meaning disregarding surface grammatical and gender differences between them.

4.2. Synonymy in Tatar socio-political vocabulary

When working on the project, we discovered that distinguishing feature of the contemporary Tatar lexicon is the presence of absolute synonyms of different origin in and structure, which is mainly caused by language contacts. In the Soviet era many words denoting socio-political realities were borrowed from Russian and European languages (as a rule, via Russian) or were coined by component-by-component translation of Russian items. The movement for the revival of the Tatar language since late 80s led to active language renewal, which caused crowding out of Russian and European words and enriching the Tatar vocabulary with many items of Tatar and Oriental (Arabic and Persian) origin. Many Arabic and Persian words and their derivatives that were considered as obsolescent in the Soviet period (and marked correspondingly in the dictionaries of the time) were entered into the active vocabulary fund. Thus in the modern Tatar, words of different roots (Turkic, Russian, Arabic, Persian, Greek, Latin, and English) denoting the same referent are competing. Selection of homosemous words is not completed yet, and the concurrence brought dissimilar results for different lexemes (Galieva, Nevzorova, Yakubova 2017; Galieva 2018).

Most of the synonyms appear when Russian items are translated into Tatar, and we gather all of them, even if only a single variant of the term is represented in Russian (Table 7).

Table 7. TECHNICAL REGULATION concept in the Russian-Tatar thesaurus.

Concept name in Russian	*Lexical entries in Russian*	*Concept name in Tatar*	*Lexical entries in Tatar*
ТЕХНИЧЕСКОЕ РЕГУЛИРОВАНИЕ	техническое регулирование	ТЕХНИК КӨЙЛӘҮ	техник көйләү, техник җайлау, техник яктан җайга салу, техник җайга салу, техник яктан тәртипкә салу

The presence of many variants for translating the item represented in Table 7 is due to the absence of the direct equivalent of the Russian word *регулировать* 'to regulate' in the Tatar language. Another reason is the abundance of equivalents; for example, in Tatar there are two words denoting court of law: the Russian loanword *суд* 'court of law' and the Arabic one *мәхкәмә* 'court of law'. So, synonymous items, especially compound ones, may be of different composition and structure. For example, Table 8 represents the lexical entries of the concept CONSTITUTIONAL COURT in the Tatar part of the thesaurus, where it becomes clear that a Russian item has at least four Tatar translation variants (and corpus data evidences that all of them are in actual use).

Table 8. Tatar lexical entries of the CONSTITUTIONAL COURT concept in the Thesaurus

Items denoting CONSTITUTIONAL COURT	*Structure of the item*	*Number in Corpus of Written Tatar*	*Number in Tatar National Corpus*	*Number in Socio-Political Subcorpus*
Конституция суды	N + N, POSS_3	1675	119	436
Конституцион суд	ADV + N	364	66	6
Конституция мәхкәмәсе	N + N, POSS_3	517	126	5
Конституцион мәхкәмә	ADV + N	73	19	1

Unlike the Russian-Tatar dictionaries, which offer, as a rule, one variant only of translating the term (and often the latter is not the most frequently used, according to corpus data, if any dictionaries contain a needed term at all), the Thesaurus gives sets of synonymous items that we managed to compile from dictionaries, corpora and official texts, etc. So a user can make his/her own choice in favor of an item.

In general, one can conclude that synonymy in Tatar socio-political vocabulary has diverse dimensions and it should be examined in special works. Concepts in the bilingual Thesaurus contain synonymous items of different origin and structure, and this phenomenon speaks for the unstable language usage rather than the richness of the Tatar language.

4.3. Influence of the Russian language

Data of the bilingual thesaurus provide us with reliable linguistic data which makes it possible to examine correspondences between the languages as well as to study current processes in Tatar vocabulary in terms of structure and content. Currently the Tatar language is greatly influenced by Russian, and this influence extends to vocabulary (loan-words, loan translation, partial calques, etc.) and grammar (for example, choosing the so called izafe constructions or adjective + noun constructions, verb control options). The Thesaurus provides us with information on borrowed vocabulary, and allows to make its quantitative evaluation. Table 9 represents the number of borrowings from Russian in the Tatar socio-political (data on concept names only is used).

Table 9. Loanwords and partial calques in Thesaurus.

Loanwords and partial calques	*Number in the sample*	*Examples*		
		Russian item	*Tatar item*	*Translation*
Loanwords	769	банк доцент	банк доцент	'bank' 'assistant professor'
Partial calques	1120	сотрудник банка кодекс законов	банк хезмәткәре законнар кодексы	'bank employee' 'code of laws'
Number in the sample	5830			

So the Thesaurus data enables us to trace main features and actual trends in Tatar socio-political vocabulary. The Republic of Tatarstan is a multinational region of the Russian Federation, and language contacts, especially the influence of the dominant Russian language, is not to be underestimated. The Russian language maintains in Tatarstan its higher social position and has a significant impact on the current state and development trends of the lexical-semantic system of the Tatar language. Bilingual Socio-Political Thesaurus provides a user with translation equivalents (correspondences) of Russian terms, and in many cases, the user has a choice of which variant of the term to prefer.

Conclusion

The Russian-Tatar Socio-Political Thesaurus is a wordnet-like lexical resource, with a hierarchical structure based on concepts and lexical entries. Unlike a wordnet, concepts in the Thesaurus embrace ontological synonyms regardless of their part of speech characteristics.

We consider the value of the Socio-Political Thesaurus is bidimensional: (1) from the practical viewpoint, as a lexicographic resource that contains relevant data on basic domains of the present-day social life; (2) as a collection of arranged lexical data for further research of actual processes in Tatar vocabulary and language interaction.

The addressees of the bilingual Socio-Political Thesaurus are government and public figures, administrative officers, journalists, Tatar language teachers, students, etc. This resource also may be used to improve Russian-Tatar machine translation systems and in Tatar texts processing for various purposes.

Future plans embrace expanding the volume of the Thesaurus, including new subject areas, as well as adding newly found lexical entries to available concepts and, if necessary, revising relations between concepts.

Abbreviations

ADJ – adjective, COMIT – comitative, F – feminine, N – noun, NMLZ – nominalizer, PL – plural, POSS_3 – possessive, 3d person, SG – singular, V – verb.

References

Bilgin, O., Çetinoğlu, Ö., Oflazer, K. 2004 - 'Building a wordnet for Turkish'. Romanian Journal of Information Science and Technology, 7(1-2), 163-172.

Corpus of Written Tatar - URL: http:/corpus.tatar/.

Doborjginidze, N., Lobzhanidze, I. 2016 - Corpus of the Georgian language. In Margalitadze T., Meladze G. (eds.). *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*. 6 – 10 September, 2016 — Tbilisi, Ivane Javakhishvili Tbilisi State University, 328 – 334.

El-Haj, M., Kruschwitz, U., Fox, C. 2015 - 'Creating language resources for under-resourced languages: methodologies, and experiments with Arabic'. *Language Resources and Evaluation*, 49(3), 549-580.

http://eprints.lancs.ac.uk/71289/1/ELHAJ_LREV.pdf

Fellbaum, C. 2010 - 'Wordnet'. In *Theory and applications of ontology: computer applications*. Springer, 231–243.

Galieva, A. 2018 - 'Synonymy in Modern Tatar Reflected by the Tatar-Russian Socio-Political Thesaurus'. In: Čibej J., Gorjanc V., Kosem I., Krek S. (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana, 585 — 994.

Galieva, A., Nevzorova, O., Yakubova, D. 2017 - 'Russian-Tatar Socio-Political Thesaurus: Methodology, Challenges, the Status of the Project'. In: Angelova, G. et al. (eds). *International Conference Recent Advances in Natural Language Processin*g, Varn, 245 - 252.

Information materials on the final results of the 2010 All-Russian Population Census - (2010) [Informatsionnye materialy ob okonchatel'nyh itogah Vserossijskoj perepisi naseleniya 2010 goda]. URL:

http://www.gks.ru/free_doc/new_site/perepis2010/perepis_itogi1612.htm.

Khakimov, B.E. 2018 - 'The Experience of Thesaurus Modeling of Tatar Information Technologies Terminology' [Opyt tezaurusnogo modelirovaniya tatarskoy terminologii informatsionnyih tehnologiy]. *Kazanskaya nauka*, 11, 193-198.

Loukachevitch, N. 2011 - Thesauri in information retrieval problems [Tezaurusy v zadachakh informatsionnogo poiska]. Moscow: Moscow State University Press.

Loukachevitch, N., Dobrov, B. 2014 - 'RuThes Linguistic Ontology vs. Russian Wordnets'. In: *Proceedings of the Seventh Global Wordnet Conference.* Tartu: University of Tartu Press, 154-162.

Loukachevitch, N., Dobrov B. 2015 - 'The Socio-Political Thesaurus as a Resource for Automatic Document Processing in Russian'. *Terminology*, vol. 21(2), 237-262.

Margalitadze, T., Meladze, G. 2016 - 'Importance of the Issue of Partial Equivalence for Bilingual Lexicography and Language Teaching'. In: Margalitadze T., Meladze G. (eds.) Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity. 6 – 10 September, 2016 — Tbilisi, Ivane Javakhishvili Tbilisi State University, 2016, 787-797.

Miller, G. A. 1995 - 'Wordnet: a Lexical Database for English'. *Communications of the ACM*, 38(11):39–41.

Russian-Tatar Socio-Political Thesaurus. URL: http://tattez.turklang.tatar/

Scannell, K. P. 2007 - 'The Crúbadán Project: Corpus Building for Under-resourced Languages'. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, Vol. 4, pp. 5-15.

Shvedova, N. Yu. (ed.) 1980 - Russian Grammar [Russkaya grammatika], V. 1. Moskow: Nauka.

Socio-Political Subcorpus of Tatar National Corpus/ URL: http://tugantel.tatar/corpus/op/.

Tachbelie, M. Y., Abate, S. T., Besacier, L. 2011 - 'Part-of-Speech Tagging for Underresourced and Morphologically Rich Languages—the

Case of Amharic'. In *Conference on Human Language Technology for Development, Alexandria, Egypt, 2-5 May*, 50-55.

Tatar National Corpus - http://tugantel.tatar/?lang=en

Vossen, P. 1997 - 'Eurowordnet: a Multilingual Database for Information Retrieval'. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, 5–7.

Vossen, P. 2002 - EuroWordNet: General Document. URL: http://vossen.info/docs/2002/EWNGeneral.pdf.

Yandex Translate - URL: https://translate.yandex.com/.

ალფია გალიევა
თათრეთის რესპუბლიკის მეცნიერებათა აკადემია, თათრეთის რესპუბლიკა
amgalieva@gmail.com

ოლგა ნევზოროვა
თათრეთის რესპუბლიკის მეცნიერებათა აკადემია, თათრეთის რესპუბლიკა
onevzoro@gmail.com

ჯავდეთ სულეიმანოვი
თათრეთის რესპუბლიკის მეცნიერებათა აკადემია, თათრეთის რესპუბლიკა
dvdt.slt@gmail.com

საკვანძო სიტყვები: ორენოვანი თეზაურუსი, ცნება, ტერმინოლოგია, სოციო-პოლიტიკური ლექსიკა, ლექსიკური სინონიმია.

თათრული სოციოპოლიტიკური ტერმინოლოგია ორენოვან თეზაურუსში

რეზიუმე

სტატიაში განხილულია რუსულ-თათრული სოციოპოლიტიკური თეზაურუსის (http://tattez.antat.ru/) შედგენის ძირითად მეთოდოლოგია. აღნიშნული თეზაურუსი იქმნება რუსული რუტესის თეზაურუსის ფორმატის საფუძველზე (http://www.labinform.ru/pub/ruthes/index.htm). პროექტის მიზანია, შეავსოს სოციოპოლიტიკურ სფეროებთან დაკავშირებული თანამედროვე თათრული ლექსიკა, რომელიც შეეხება შემდეგ სფეროებს: სახელმწიფო მმართველობას,

ეკონომიკას, სოციალურ ცხოვრებას, მართლმსაჯულებას, საომარ მოქმედებებს, კულტურასა და რელიგიას. თათრული თეზაურუსი მოიცავს ზოგიერთ ისეთ ლექსიკურ ერთეულს, რომელიც შეიძლება ხშირად დადასტურდეს სხვადასხვა დარგობრივ ტექსტშიც. თითოეული ცნება უკავშირდება და ტექსტშივე მიემართება (ლექსიკურ ჩანაწერს) ენობრივ გამოთქმათა გარკვეულ ჯგუფს (ერთსიტყვიან და რამდენიმესიტყვიან გამონათქვამს). თეზაურუსის თათრული ნაწილი ეფუძნება რუტესის ცნებათა ნუსხას, შესაბამისად, შენარჩუნებულია რუტესის ძირითადი აგებულება.

თათრული თეზაურუსი ეყრდნობა შემდეგი კორპუსებს:

1. თათართა ეროვნული კორპუსი (http://tugantel.tatar/?lang=en);
2. სამწერლობო თათრული ენის კორპუსი (http://www.corpus.tatar/en).

სტატიაში ასევე განხილულია თათრული ენის ტერმინოლოგიური პრობლემები და რუსული და თათრული ენების იმ სემანტიკურ მიმართებათა შორის განსხვავება, რომლის გამოყენებითაც იქმნება ტერმინები.

პროექტზე მუშაობის მთავარი ამოცანაა ლექსიკური მონაცემების მოპოვება და თათრულ ენაში არსებული სოციოპოლიტიკური ლექსიკის შეძლებისდაგვარად სრულად წარმოდგენა, მათ შორის – მოჭარბებული სინონიმებისაც. ვინაიდან თათრული კულტურა დასავლური და აღმოსავლური ცივილიზაციების გზაგასაყარზეა, ენაში ხშირია ლექსიკური ნასესხობანი როგორც არაბულ-მუსლიმანური, ისე ევროპული კულტურებიდან. ევროპული ენებიდან ნასესხები სიტყვები თათრულში მკვიდრდება შუამავალი ენის – რუსულის – მეშვეობით, სწორედ რუსულიდან სესხულობს თათრული ენა ბევრ სიტყვასა თუ კონსტრუქციას. გარდა ამისა, სინონიმთა მნიშვნელოვანი ნაწილი თურქული და თათრული ლექსიკური მასალის საფუძველზეა შექმნილი. სწორედ ამიტომ თანამედროვე თათრულ ენაში სხვადასხვა წარმო-მავლობის სინონიმებია (თურქული, რუსული, არაბული, სპარსული, ბერძნული, ლათინური და ინგლისური), რომლებიც მდიდარ ლექსიკურ მასალას ქმნის. დაბოლოს, თათრული ენის გრამატიკული სისტემა საშუალებას იძლევა ტერმინები სიტყვაწარმოებითი თუ სინტაქსური სტრუქტურის საშუალებით შეიქმნას. შედეგად, სოციოპოლიტიკურ ტერმინებს აქვთ სხვადასხვა ლექსიკური შედგენილობისა და სტრუქტურის მქონე სინონიმები, რომლებიც უკლებლივ უნდა აისახოს თათრულ თეზაურუსში.

ამჟამად რუსულ-თათრული სოციოპოლიტიკური თეზაურუსი მოიცავს 9000 ცნებას და მუდმივად ივსება აღნიშნული პროექტისათვის საგანგებოდ შექმნილი პროგრამული საშუალებებით[1].

[1] აღნიშნული კვლევა ფინანსურად უზრუნველყო რუსეთის სამეცნიერო ფონდმა, კვლევითი პროექტი № 16-18-02074.