Development of Corpus-Based Tatar-Russian Socio-Political Dictionary of Collocations

Alfiya Galieva, Olga Nevzorova

***Abstract* — This paper discusses main sources and methodology of compiling the Tatar-Russian Socio-Political Dictionary of collocations. The area of collocations within the language system is of particular importance, and the well-known language-specificity of collocations suggests the need for bilingual collocation dictionaries. Socio-political domain is one of the most dynamically developing spheres of present-day life, with the socio-political vocabulary rapidly developing and being enriched with new lexical items and senses reflecting the realities of the time. The Dictionary is based on data of the available corpora of the Tatar language and is built as a collocation dictionary.**

The main criteria for selecting linguistic data are those of objective (frequency in the corpus) and subjective evaluation (evaluation of the word from the point of view of its thematic, stylistic and collocational value). The main unit in the Dictionary is the noun phrase formed by filling one of possible semantic-syntactic positions of the word and meeting the criteria of semantic completeness. As an exception, we also included certain combinations of header words with postpositions derived from nouns, as long as the corresponding collocations are typical for socio-political discourse. A special attention is paid to a distinguishing feature of the contemporary Tatar lexicon – synonymy.

***Keywords* — socio-political vocabulary, collocation dictionary, the Tatar language, bilingual dictionary, corpus**

I. Introduction

Development of new lexicographic resources for minority and low-resource languages is a task of current importance that has scientific and practical dimensions. This paper presents the project on developing the Tatar-Russian Socio-Political Dictionary of Collocations [1]. The area of collocations within the language system is of particular importance, and the well-known language-specificity of collocations suggests the need for bilingual collocation dictionaries.

Tatar, related to Turkic family, is the language of the ethnic majority in the Tatarstan Republic of the Russian Federation, where it co-exists with Russian as a state language under the current legislation that proclaims them as equal, according to the Law on Languages which was adopted in 1992 in Tatarstan. Future preservation and development of the Tatar language, which is functioning in parallel with the Russian language, is largely determined by its use in education and electronic communications, and one of the topical tasks is developing terminology in Tatar, fixing and representing it in open access resources.

Compiling the Tatar-Russian Socio-Political Dictionary of Collocations is carried out due to a combination of factors, such as:

- socio-political vocabulary is a significant formation of any language in active use, and it is undergoing permanent changes;
- available bilingual Russian-Tatar dictionaries for general purposes and special lexicons contain outdated data and are lacking new words and phrases, thus failing to reflect the current state of the language; besides they contain rather a limited number of collocations;
- Tatar corpora provide reliable information about Tatar socio-political vocabulary and lexical co-occurrences in actual use, so they are to be used in compiling new generation dictionaries.

The Tatar-Russian Socio-Political Dictionary of collocations is based on data of the available Tatar corpora. The use of corpus-based dictionaries is but a recent trend, especially as far as it concerns minor languages [2]. In the case of Tatar, the possibilities of compiling lexicographic resources of different types are virtually enormous, with the support of the available corpora.

This article discusses the main design decisions adopted in compiling the Tatar-Russian Socio-Political Dictionary of Collocations. The direction of compiling the dictionary is from Tatar to Russian (and not vice versa) because the dictionary is aimed at detecting and fixing main features of the present-day Tatar socio-political lexicon. The new Dictionary is demanded by linguists, journalists and professional translators as well as in education process both in secondary and high school.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of related work; Section 3 outlines main available resources that were used to compile the Dictionary; Section 4 presents the methodology of developing the Dictionary, and considerable attention is paid to the issue of presenting the challenges encountered. Section 5 sketches out an interesting trend concerning Tatar socio-political vocabulary that was detected when compiling the Dictionary – co-existence of a great number of synonymous items. Finally, Section 6 lists the conclusions that can be derived from this research.

II.

Manuscript received October 22, 2018. The reported study was funded by Russian Science Foundation, research project № 16-18-02074.

Alfiya Galieva is with the Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia (e-mail: amgalieva@gmail.com).
Olga Nevzorova is with the Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia and Kazan Federal University, Kazan, Russia (e-mail: onevzoro@gmail.com).

Related work

In recent decades the lexicographic and practical value of collocations has become evident and a large number of linguists and editors were involved in projects related to this issue. In [3] collocation is defined as "a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text" [3, p.24]. Collocations in this broad sense may include a wide range of heterogeneous sets of words. So there are some difficulties concerning understanding the nature, structure and definition of collocations. Alan Partington [4] classifies the definitions of collocations into textual (co-occurrence in a text), statistical (co-occurrence with a greater than random probability) and psychological (co-occurrence due to a psychological link between words).

Collocations are highly specific for a particular language, they are conventionalized and may have contextual restrictions. Competently composed collocation dictionaries are a valuable resource for translators and language learners (for beginners as well as advanced students). Practical interest in collocations is registered due to the fact that they are considered as an important source for producing naturally sounding speech, which is one of the primary goals in language teaching.

Language learners and users draw much of their vocabulary knowledge from context, apart from explicit instruction. Paul Nation [5, p.318] summarizes the discussions about the importance of collocations with the following arguments: (1) language knowledge is collocational knowledge; (2) fluent and appropriate language use requires collocational knowledge; and (3) many words are used in a limited set of collocations and knowing these is a part of what is involved in knowing the words.

In 1986 M. Benson, E. Benson and R. Ilson published the first monolingual English collocations dictionary, The BBI Combinatory Dictionary of English [6]. In 1999, LTP Dictionary of Selected Collocations was issued [7]. These dictionaries present essential collocations of English in an easily accessible form that shows which word combinations exist in English and which grammatical constructions are possible. Nevertheless they were not based on corpus data. Oxford Collocations Dictionary for Students of English [8] became the first corpus-based English collocations dictionary which provides reliable and pragmatically selected data on the most frequently used word combinations in British and American English. Another corpus-based monolingual English collocations dictionary - Macmillan Collocations Dictionary for Learners of English - was published in 2010 [9]. These dictionaries focus on students' productive needs and provide a wide repertoire of collocations.

Similar dictionaries were published for other languages: Spanish [10, 11], French [12], Russian [13], etc.

In Russia, approaches to study collocations aim at combining presently used statistical and corpus linguistics methods with the traditions of the Russian semantic and lexicographic schools. E. Enikeeva and O. Mitrofanova apply the theory of lexical functions to extracting collocations; in their study distributed word vector models are used as a state-of-the-art computational basis for the tested method [14]. Using different statistical measures, such as MI and t-score, researchers evaluate the nature, structure, and collocability of collocations retrieved from corpora [15,16].

Maria Khohkhlova focuses on setting a gold standard for Russian collocations that includes data from Russian dictionaries and corpora [17].

The fact that collocations are highly language-specific poses a problem of compiling bilingual collocation dictionaries. Such dictionaries [18-21] provide information about interlingual lexical correspondences and are aimed at encouraging learners and translators to more actively use collocations and incorporate them into their mental lexicon.

Due to the historical destiny of the Tatar people and the influence of geopolitical factors, the main stream of bilingual Tatar lexicography was Tatar-Russian and Russian-Tatar. A brief review of Tatar-Russian lexicography is presented in [22].

The available Tatar-Russian combinatory dictionaries are compiled for education purposes [23] or contain mainly phraseological data [24]. Having limited volume, all of them were compiled before developing the Tatar corpora. So Tatar lexicography needs corpus based dictionaries providing relevant, statistically verified information about word meanings, distributions and contextual environments.

III. Available dictionaries and linguistic corpora

The socio-political domain is a broad sphere of contemporary social relations; awareness of those composes the competence of non-specialist educated people [25]. This domain comprises the following main topics:

- politics and state administration;
- international relations;
- economy and financial issues;
- industry;
- army and military sphere;
- social sphere;
- culture and art;
- religion;
- sports, etc.

The socio-political domain concepts are clearly manifested in news chronicles provided by mass media. These texts include discussions related to a large number of specific domains; they contain a lot of domain-specific terms (criminal law, investor, inflation, legal act, etc.) but at the same time are intended to be understandable by a wide circle of non-professionals [25].

The study of socio-political lexicon is of great linguistic interest. This domain is one of the most dynamically developing spheres of present-day life, with the socio-political vocabulary rapidly developing and being enriched with new lexical items reflecting the realities of the time. Linguistic data from constantly updated corpora are very important for a comprehensive and objective study of the current Tatar language.

As we mentioned above, the last decades saw publication of new Tatar dictionaries on many knowledge domains; nevertheless, recently launched bilingual Russian-Tat

dictionaries are merely extended versions of earlier dictionaries, and new special bilingual Tatar-Russian socio-political dictionaries have not been developed. Available dictionaries do not meet the requirements of the present time. Being compiled on data of manually collected card files, these dictionaries barely reflect current socio-political vocabulary and lexical co-occurrences. The obsoleteness of these dictionaries becomes apparent in three main aspects:

- the dictionaries are incomplete and do not contain entries reflecting present-day socio-political items widely used in Russia (such as *political correctness, lines of communication, framework agreement, taxable* and many others);
- the dictionaries do not contain target language (Tatar) translation units in actual use, containing instead those that had been actively used in the Soviet era;
- the dictionaries pay little attention to lexical co-occurrences.

So the Tatar-Russian Socio-Political Dictionary of collocations aims at a true mapping of the words meanings and actual trends in use, and is based on data of the main available corpora of the Tatar language:

- the Corpus of Written Tatar (CWT) [26];
- the Tatar National Corpus (TNC) [27].

These corpora include texts of various genres, from official documents and scientific publications to fiction, media texts, and school books. They are being hourly replenished, with the update of textual collections mainly carried out through processing media texts, which provides a constant flow of fresh linguistic material. The corpora have comparable volumes and are supplied with morphological description, i.e. information about the part of speech of the word stem and the set of its grammar features. Basic information on the Tatar corpora is displayed in Table 1.

Besides, data from a special Socio-Political Subcorpus of the Tatar language [28] is employed. This subcorpus is composed of texts of electronic media on social and political topics, as well as texts of legal documents.

Corpus technology greatly facilitates obtaining empirical data and their processing for compiling a dictionary; in particular, it allows obtaining objective data about the frequency, distribution and compatibility of lexemes. A large volume of corpora guarantees representing natural environments of linguistic items and ensures completeness of representation of the whole range of linguistic phenomena, which is crucial in compiling dictionaries. Besides, corpora fix new vocabulary and changes in words' environments.

Table 1. Available Tatar corpora and their functionality

Criteria of comparison	**Tatar National Corpus**	**Corpus of Written Tatar**	**Russian-Tatar socio-political subcorpus**
Launch date	2012	2012	2016
Corpus volume	100,000,000 tokens	116,000,000 tokens	15,000,000 tokens
Morphological annotation	+	+	+
Search for individual morphological categories	+	+	+
Search for set of given morphological categories	+	+	+
The main source for morphological tags	Leipzig Glossing Rules (a set of tags specific for Turkic languages is added)	Apertium project tags for Turkic languages	Leipzig Glossing Rules (a set of tags specific for Turkic languages is added)
Grammatical disambiguation	-	-	-
Stylistic features of texts presented	non-fiction texts – 28% fiction- 72%	non-fiction texts – 65% fiction- 35%	non-fiction texts – 100%
Search for a phrase	+	+	+
Search for a part of word	+ for any given part of word	+ for any given part of word	+ for any given part of word
Displayed processing time	+	-	+
Inbuilt speech synthesizer	-	+	-
Texts metadata	+	+	+
Output per page	10 contexts	50 contexts	10 contexts

The search system of corpora makes it possible to conduct search by lemma (lexeme), by word form, as well as by a set of morphological parameters specified by the user. Each corpus has its own set of options that may be better suited for various educational tasks. To illustrate this, the corpus management system of the Tatar National Corpus supports search of stop words, search by any given part of word, and search based on the use of logical formulae. Thus the user can make a sophisticated inquiry – for instance, in order to come up with various types of grammatical phenomena or collocations.

Linguistic data from constantly updated corpora are of great importance for a comprehensive and objective research into the processes taking place in the modern socio-political discourse, and also for fixing the actual state of lexicon in lexicographic resources.

IV. Methodology of compiling the Dictionary

In this section, we describe how we collected and analyzed the data for the Dictionary. We discuss the main challenges and the solutions obtained.

The methodology of compiling the Dictionary included

the following main stages:

- selecting header words using corpus data;
- retrieving collocations in Tatar corpora and linguistic dictionaries;
- translating collocations into Russian.

The initial stage implied compiling the frequency list of actual terms (the list of one-word terms as potential header words) using the Socio-Political Subcorpus. First we automatically generated a list of the most frequently used noun word forms which were lemmatized, and then the list of potential header words was manually compiled. The main criteria for selecting vocabulary are based on objective (frequency in the Socio-Political Subcorpus) and subjective evaluation (considering the words' thematic, stylistic and collocational value and their use in texts of social, political and cultural topics). The current list of potential header words is composed of 1000 items.

Then, using bi-gram models (a sequence of two adjacent elements) obtained from the Corpus of Written Tatar and the Tatar National Corpus, we have built a frequency list of collocations for each frequent term.

The limitations for cutting elements from the collocations lists were based on the frequency of using linguistic items in the corpora, and these limitations were determined empirically; in the current version of the dictionary the lower threshold for including a collocation is its occurrence in at least 50 corpus contexts (actually for the overwhelming majority of collocations this threshold is significantly higher because corpus collocations are given for word forms, not for lemmas).

The Dictionary of collocations contains meaningful common word combinations of different structures such as *дәүләт сәясәте* 'policy of the state' (N + N), *табигый байлыклар* 'natural resources' (ADJ + N), *музей ачу* 'to open the museum' (N + V). Collocations refer to how words go together or form fixed relationships; they are regarded as essential building blocks of the natural language. The capability of linguistic items to be combined with each other when forming higher level units is a fundamental feature of the human language.

Collocations in the Dictionary are represented in basic forms (issues concerning lemmatization are discussed in a special section below) and currently the bulk of the Dictionary is composed of collocated pairs of words.

The main unit in the Dictionary is a noun phrase formed by filling one of possible semantic-syntactic positions of the word and meeting the criteria of semantic completeness. Quantitatively such an item may consist of two or more notional words. As an exception, we also included certain combinations of header words with postpositions derived from nouns, as long as the corresponding collocations are typical for socio-political discourse, for example

карар нигезендә 'on the basis of a resolution' (header word КАРАР 'resolution, decree');

закон каршында 'before the law' (header word ЗАКОН 'law');

хәйрия максатларында 'for charity purposes' (header word ХӘЙРИЯ 'charity').

In the current version of the Dictionary most of the collocations are composed of two notional components. When selecting the collocations, we considered the syntactic structure of each of them and the morphological parameters of their constituents. We also took into account regularities of grammatical (non-inflectional) variants of word combinations, in particular, regular grammatical variants of collocations are considered as the same nominative item and are represented in the same entry line.

The entries of the new Dictionary are currently limited by nouns and relative adjectives arranged alphabetically. The structure of an entry is built as follows:

- header word (capitalized) – the Tatar word, frequently used in socio-political domain;
- Russian translation of the header word (only senses relevant for socio-political discourse are provided);
- lexical compatibility to the right, Russian translation of collocations;
- lexical compatibility to the left, Russian translation of collocations [29].

Table 2 represents typical collocations of the entry ДӘҮЛӘТ 'state' and their frequency in Tatar corpora.

Table 2. Typical collocations of the entry ДӘҮЛӘТ 'state' and their frequency in Tatar corpora

Collocation type	Collocation	Russian translation	English translation	Number in TNC	Number in CWT
To the right	Дәүләт Думасы	Государственная Дума	The State Duma	11441	14108
	Дәүләт Советы	Государственный Совет	The State Council	23697	35367
	дәүләт киңәшчесе	Государственный советник	State Counselor	3492	4100
	дәүләт имтиханнары	Государственные экзамены	State exams	2914	1612
	дәүләт хакимияте	Государственная власть	State power	2206	3157
	дәүләт органнары	Органы государства	State bodies	1512	1657
To the left	мөстәкыйль дәүләт	Самостоятельное государство	Independent state	164	374
	федератив дәүләт	Федеративное государство	Federal state	58	208
	хокукый дәүләт	Правовое государство	Constitutional state	73	228
	күпмилләтле дәүләт	Многонациональное государство	Multinational state	80	196
	дөньяви дәүләт	Светское государство	Secular state	39	172
	суверен дәүләт	Суверенное государство	Sovereign state	57	126

Collocations given in Table 2 designate universal and Russian political realities. Two designations express the peculiarities of state administration in Tatarstan: The State Council is the unicameral Tatarstan parlament; the State Counselor is the current official position of the Tatarstan ex-president Mintimer Shaimiev.

Fig. 1 represents collocations of the entry ДӘҮЛӘТ ‘state’ in the Dictionary.

compatibility to the right is given. Fig. 2 represents the entry of the word ИҖТИМАГЫЙ ‘social’.

If the header word has widespread spelling variants, these are also mentioned:

вазифа, вазыйфа ‘official duties, official position’;

икътисади, икътисадый ‘economic’.

Such spelling variants are produced by violations of vowel harmony in loan words (mainly of Arabic origin) and

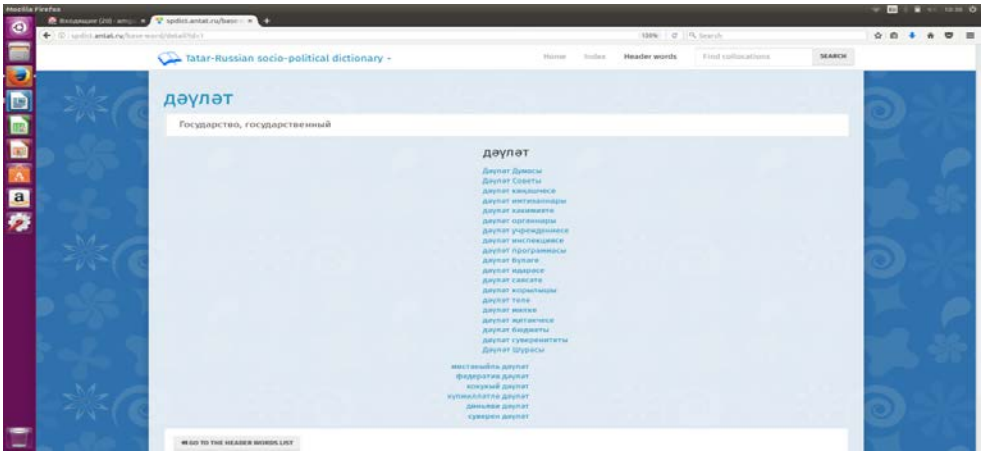

Fig. 3. List of collocations with the word *ДӘҮЛӘТ* (Tat) / *STATE* (En)

The Tatar language has left branching syntax (left-branching structures are head-final), so for adjectives only

are differently represented in Tatar lexicons and dictionaries. While compiling the Dictionary of collocations we selected

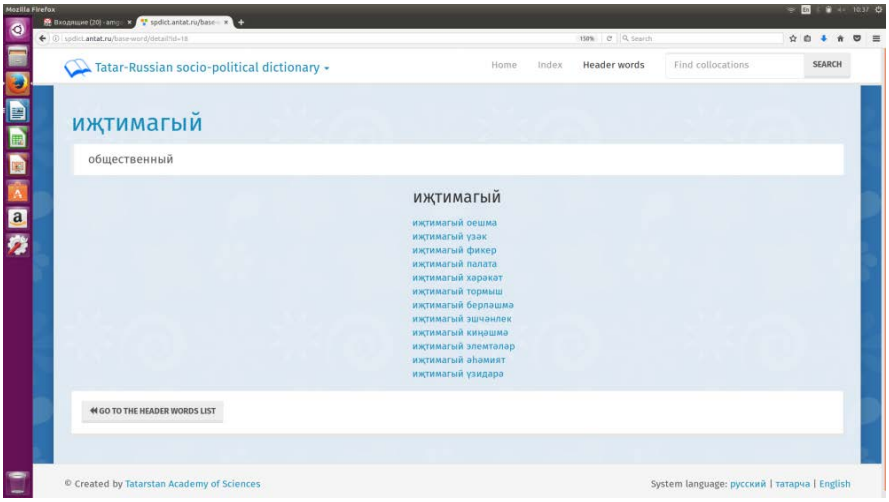

Fig. 1. List of collocations with the word *ИҖТИМАГЫЙ* (Tat) / *SOCIAL* (En)

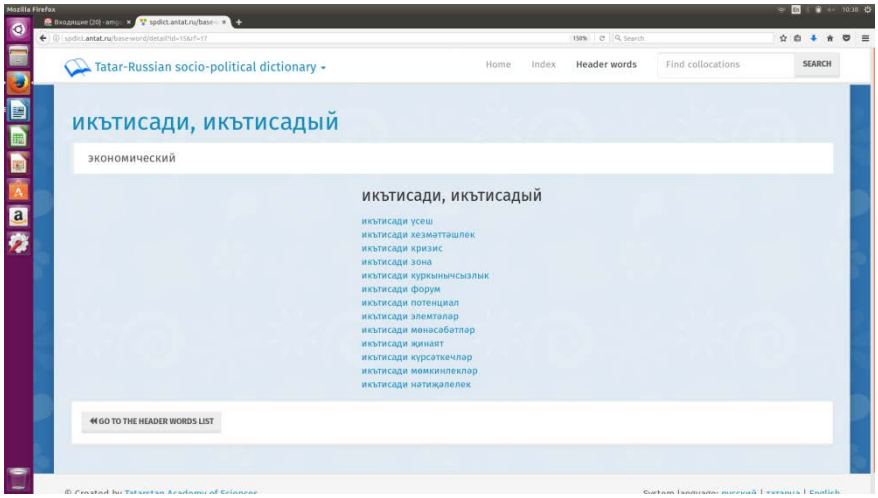

Fig. 2. List of collocations with the word *ИКЪТИСАДИ* or its orthographic variant *ИКЪТИСАДЫЙ* 'economic'

the most widespread item, according to corpus collections, and fixed it as the basic one; the other variant is represented as the header word after the main variant. Collocations merely contain the most widespread spelling variants of the word. Fig. 3 represents the entry of the word ИКЪТИСАДИ / ИКЪТИСАДЫЙ ‘economic’.

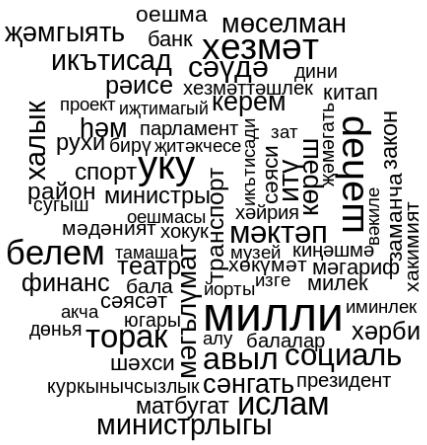

Fig. 4. Word cloud of the 50 most frequently used items in the Dictionary

When collecting collocations we fix them in their basic form, preserving the basic grammatical structure of the collocation as a whole. The total automatic lemmatization of linguistic data is of little avail, because it destructs grammatically conditioned word combinations. The issue of lemmatizing collocations in morphologically rich languages is not trivial in many cases, and the solutions applied are discussed in a special paper [29].

Fig. 4 represents a word cloud of the 50 words that are most frequently used in the Dictionary of collocations. Almost all of them are notional words – nouns and adjectives, besides the two functional words; the most frequently used item is the adjective *милли* 'national, ethnic' which is used in 1.83 % of collocations of the Dictionary.

V. Synonymy in Tatar socio-political domain reflected by the Collocations Dictionary

Searching for frequently used Tatar collocations among corpus data raises the problem of actually used Tatar vocabulary. The Dictionary represents numerous synonymous collocations (semantic doubles) which are used to designate topical realities of modern socio-political life. These synonymous collocations are engendered by a set of factors of different nature. This issue is quite interesting and deserves a special attention.

Synonymy is one of fundamental concepts in linguistics that manifests itself at different language levels. A prevailing point of view is that synonymy shows a high degree of language development, and that absolute (total) synonymy occurs infrequently because “yet even if near-synonyms do name one and the same thing, they name it in different ways: they present different perspectives on a situation" [30, p.1]. Available attitudes range from totally rejecting the idea of absolute (total) synonymy in language to claiming extreme rareness of them [31, 32]. The situation with Tatar socio-political vocabulary in many respects refutes this thesis about rareness of absolute synonymy.

The Tatar culture lies at the crossroads of Oriental and Occidental worlds, which left an imprint on the history of the Tatar people and the language. Permanent cultural contacts led to lexical exchanges. The original Turkic vocabulary is enriched with a huge amount of loan words coming both from Arabic-Persian Muslim cultural area and from Europe, and the mediating role of the Russian language for the latter should not be underestimated. Besides the Russian language itself is the main source for borrowing terms designating realities of Russian political structure and economic, cultural and social issues. In addition, a significant part of synonyms are built on the basis of Turkic and Tatar lexical material.

Table 3. Distribution of synonyms coming from different languages (according to corpus data)

Lexeme	English translation, PoS	Origin of the word	Number in Corpus of Written Tatar	Number in Tatar National Corpus
экономик	Economic (adj.)	Greek	621	676
икътисади		Arabic	27,351	22,274
политик	Political (adj.)	Greek	1,005	1,536
сәяси		Arabic	25,011	24,489
республика	Republic (noun)	Latin	316,667	258,433
җөмһүрият		Arabic	2,479	1,631
информацион	Informational (adj.)	Latin	858	807
мәгълүмати		Arabic	7,076	689
суд	Court of law (noun)	Russian	15,725	14,243
мәхкәмә		Arabic	8,718	8,101

Presently the bulk of Tatar socio-political terms is coined by component-by-component translation of corresponding Russian items. The Dictionary represents numerous semantic doubles (synonymous collocations) which are used to designate topical realities of modern socio-political life.

As we mentioned above, the most of appearing Tatar terms are built by translating the correspondent Russian words. Tatar terminology is created by diverse translators and teams of specialists, and the course of translation is influenced by diverse extralinguistic factors, for example by the competence, different worldviews and ideological orientations of terminology developers. Dissimilar preferences of terminology developers, like Turkocentrism, orientation to Oriental – Arabic or Persian, Russian or European languages, often lead to using different

Table 4. Synonymous collocations composed by different grammatical patterns

Noun phrase	Structure of the noun phrase	Translation	Number in Corpus of Written Tatar	Number in Tatar National Corpus
Җәмгыять палатасы	N + N, POSS_3	Civic Chamber	10	2
иҗтимагый палата	ADJ +N	Civic Chamber	2 637	1 664
икътисад үсеше	N + N, POSS_3	Economic development	489	128
икътисади үсеш	ADJ +N	Economic development	5 663	3 242
Җәмгыять тормышы	N + N, POSS_3	Social life	379	113
иҗтимагый тормыш	ADJ +N	Social life	1,336	1,319
табигать байлыклары	N + N, POSS_3	Natural resources	336	99
Табигый байлыклар	ADJ +N	Natural resources	1,068	878
Хокук акты	N + N, POSS_3	Legal act	25	5
Хокукый акт	ADJ +N	Legal act	1,082	1,271
мәдәният үзәге	N + N, POSS_3	Culture centre	2,942	2,218
мәдәни үзәк	ADJ +N	Culture centre	4,403	2,048
икътисад кризисы	N + N, POSS_3	Economic crisis	75	33
икътисадый (икътисади) кризис	ADJ +N	Economic crisis	1,572	1,484

denominations when referring to the same object. Therefore, words of different origin (Turkic and Tatar, Arabic and Persian, Greek, Latin, English and Russian) related to the same referent coexist in Tatar, which engenders a great number of synonyms at the single word level.

Table 3 illustrates redistribution of absolute synonyms of European and Arabic origin in corpus collections (for nouns the number of the Nominative case forms, and not the number of lemmas, is given).

It is noteworthy that all words of Arabic origin presented in Table 3 entered active Tatar lexicon relatively recently, in the late 80's; nevertheless, they succeeded to push out the words of European origin with the same meaning which had been actively used in the language of the Soviet era.

Some peculiarities of lexical, derivational and grammatical systems of the Tatar language also lead to originating of a great number of synonyms. On the level of multiword terms and phrases lexical synonymy is complicated by the factor of differing grammatical structures in use.

For example, in Turkic languages the following grammatical patterns of noun phrases are regularly corresponding: ADJ +N and N + N, POSS_3. Such regular correspondences multiply the number of grammatical variants of multiword terms.

Table 4 represents collocations with the same meaning built according to dissimilar grammatical models.

Dissimilar preferences of translators when calquing corresponding Russian terms lead to the use of different designations for the same entity. For example, in the media and official texts of corpus collections we found 4 different noun phrases to designate the term Constitutional Court - the high court of Russia that deals primarily with constitutional law; these are formed by 2 core nouns meaning 'court of law' and are based on two different grammatical patterns.

Table 5. Tatar names of Constitutional Court based on different nouns and according to different grammatical patterns.

Noun phrase	Structure of the noun phrase	Number in Corpus of Written Tatar	Number in Tatar National Corpus
Конституция суды	N + N, POSS_	868	119
Конституцион суд	ADJ +N	169	66
Конституция мәхкәмәсе	N + N, POSS_3	261	126
Конституцион мәхкәмә	ADJ +N	50	25

All these synonyms satisfy the demanded frequency criterion (50 occurrences in corpus) and are included in the entries of the two different header words: *суд* (the word of Russian origin) and *мәхкәмә* (the word of Arabic origin), both denoting court of law (see Table 5).

Depending on the header word, such synonymous items fall into the same or different entries.

Another characteristic example is the notion of joint-stock company denoted by 9 compounds that can be found in corpora; these are formed by 3 core nouns meaning 'company' and are based on different grammatical patterns; each of them is characterized by different degree of frequency of use (see Table 6), and only the most frequently used are included into the Dictionary.

Table 6. Tatar compounds designating the term j*oint-stock company*

Grammatical model of the term	Term variants	Number in Corpus of Written Tatar	Number in Tatar National Corpus
N,NMZL + N.POSS_3	акционерлык җәмгыяте	4,320	1,640
	акционерлык ширкәте	13	4
	акционерлык оешмасы	13	6
N, PL + N.POSS_3	акционерлар җәмгыяте	274	76
	акционерлар ширкәте	2	-
	акционерлар оешмасы	1	1
ADJ + N	акционер җәмгыять	97	9
	акционер ширкәт	4	-
	акционер оешма	14	1

So a great number of emerging total synonyms in Tatar occurred under the influence of a combination of intralinguistic and extralinguistic factors.

Since the linguistic situation in the Republic of Tatarstan is unstable, parallel denominations can be used for a wide range of phenomena, including some official names of departments and state structures. So when processing collocations we also trace synonymous terms and fix the most frequently used in the Dictionary of Collocations. Providing corpora with reliable metadata will in future enable us to trace the dynamics of lexical changes.

VI. Conclusion and Future Work

Tatar culture is located at the intersection of Occidental and Oriental civilizations, which leads to active lexical borrowing both from Arab-Muslim cultural area and that of Europe; vocabulary is borrowed from European languages via the Russian language, and a huge amount of words and constructions are certain to be taken from Russian. Currently most Tatar socio-political terms are formed by calquing corresponding Russian terms. Dissimilar preferences of terminology creators lead to the use of different designations for the same entity.

Compiling a dictionary of collocations for low-resource languages is a topical yet rather challenging task due to the numerous details that a lexicographer is to take into consideration, including criteria for selecting header words and collocations, lemmatizing linguistic items, and distinguishing variants of collocations from synonyms. Besides such dictionary should become a user-friendly and practically valuable new resource for its target audience, and its structure should adequately represent linguistic data and be convenient for practical use.

The compiled Tatar-Russian Socio-Political Dictionary of Collocations makes it possible:

1) to fix the real use of words of Tatar, including items which are actively used in a large number of Tatar official and media texts; nevertheless those are absent in available special and bilingual Russian-Tatar dictionaries;
2) to detect and fix typical grammatical models and contexts of using items denoting socio-political realities;
3) to trace in the Tatar language words of new mintage functioning in Russian geopolitical space;
4) to keep numerous synonymous nominations used in Tatar media texts and official documents;
5) to offer Russian translations of words and collocations.

In its current state the Dictionary contains 250 header words and more than 4,000 collocations with their Russian translations. In future it is planned to extend the linguistic database (adding new entries and related collocations), providing information about grammatical structure of collocations and corpus contexts illustrating their use.

VII. List of Abbreviations

ADJ – adjective, N – noun, NMLZ – nominilizer, PL – Plural, POSS_3 – possessive, 3d person, PRES – Present, VN – verbal noun.

Acknowledgment

The reported study was funded by Russian Science Foundation, research project № 16-18-02074.

References

[1] Tatar-Russian Socio-Political Dictionary of Collocations (2018): </http://spdict.turklang.tatar/> [Accessed 29/09/2017].

[2] Jantunen, Jarmo Harri (2016). “Corpora, Phraseology and Dictionaries: How Does Corpus Research Intersect Language Teaching and Learning?”. *In* Vilas, *Begoña Sanromán* (ed.) *Collocations cross-linguistically: Corpora, dictionaries and language teaching.* Helsinki: Société Néophilologique, 97-119.

[3] Stubbs, Michael (2002): *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford: Blackwell.

[4] Partington, Alan. (1998): *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam / Philadelphia: John Benjamins.

[5] Nation, Paul. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

[6] Benson, Morton, Benson, Evelyn & Ilson, Robert (1986): *The BBI Combinatory Dictionary of English.* Amsterdam: John Benjamins.

[7] Hill, Jimmie, Lewis, Michael (1999). *LTP Dictionary of Selected Collocations*. Hove, London: Language Teaching Publications.

[8] Mcintosh Colin, Francis Ben. & Poole Richard (eds.) (2002). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

[9] Rundell, Michael. (2010) *Macmillan Collocations Dictionary for Learners of English.* Oxford: Macmillan Publishers Ltd.

[10] Bosque, Ignacio (ed.) (2004): *Redes. Diccionario combinatorio del español contemporáneo*. Madrid: SM.

[11] Bosque, Ignacio (ed.) (2006). *Diccionario combinatorio práctico del español contemporáneo*. Madrid: SM.

[12] Beauchesne, Jacques (2001). *Dictionnaire des cooccurrences*. Montréal: Guerin.

[13] Borisova, Yelena (1995): *Slovo v tekste. Slovar’ kollokatsiy (ustoychivyih slovosochetaniy) russkogo yazyika s anglo-russkim slovarem klyuchevyih slov* [A Word in the Text. Dictionary of Collocations of the Russian Language with English-Russian Dictionary of Key Words]. Moscow: Filologia.

[14] Enikeeva E.V., Mitrofanova, O.A. Russian Collocation Extraction Based on Word Embeddings // http://www.dialog-21.ru/media/3908/enikeevaevmitrofanovaoa.pdf.

[15] Yagunova, E. V., and L. M. Pivovarova (2010). The nature of collocations in the Russian language. The experience of automatic extraction and classification of the material of news texts // *Automatic Documentation and Mathematical Linguistics* 44, no. 3: 164-175.

[16] Zakharov V.P., Khokhlova M. V. (2010) Study of effectiveness of statistical measures for collocation extraction on Russian texts // Computational linguistics and intellectual technologies. No. 9, Pp. 137-143.

[17] Khokhlova M. (2018) Building a Gold Standard for a Russian Collocations Database // Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts / Jaka

Čibej, Vojko Gorjanc, Iztok Kosem and Simon Krek (Eds.) - Ljubljana. - Pp. 863-869.
[18] Benson, Morton & Benson, Evelyn (1993). *Russian-English Dictionary of Verbal Collocations.* Amsterdam: John Benjamins.
[19] Klégr Aleš, Key Petra & Hronková Norah (2005): *Česko-anglický slovník spojení: podstatné jméno a sloveso* [Czech-English combinatory dictionary: noun and verb]. Praha: Karolinum.
[20] Konecny, Christine & Autelli Erica, E. (2014). Kollokationen Italienisch-Deutsch. Vydavateľ: Helmut Buske Verlag.
[21] Orenha-Ottaiano, Adriane. (2016)." The Compilation of a Printed and Online Corpus-Based Bilingual Collocations Dictionary". *In* Margalitadze Tinatin, Meladze George (eds.). *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity*, 6 – 10 September, 2016. Tbilisi: Ivane Javakhishvili Tbilisi State University, 735-745.
[22] Safiullina, Gulshat (2016)."Bilingual Lexicography in the Republic of Tatarstan in 1990-2010". *In* Margalitadze Tinatin, Meladze George (eds.). *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity*, 6 – 10 September, 2016. Tbilisi: Ivane Javakhishvili Tbilisi State University, 475 – 479.
[23] Agishev, Khanif (2001): *Russko-tatarskiy slovar' slovosochetaniy dlya uchashchikhsia* [Russian-Tatar Dictionary of Word Combinations for Students]. Kazan: Fan.
[24] Safiullina, Flera (2002): *Tatarsko-russkiy slovar' sostavnykh clov* [Tatar-Russian Dictionary of Compound Words]. Kazan: Tatar Publishing House.
[25] Loukachevitch Natalia & Dobrov Boris: (2015) The Sociopolitical Thesaurus as a resource for automatic document processing in Russian. *Terminology*. Vol. 21, 2: 237-262.
[26] Corpus of Written Tatar (2018): <http:/corpus.tatar/> [Accessed 28/09/2018].
[27] Tatar National Corpus (2018): <http:/tugantel.tatar> [Accessed 28/09/2018].
[28] Socio-Political Subcorpus of the Tatar language (2018): <http://tugantel.tatar/corpus/op?lang=en> [Accessed 28/09/2017].
[29] Galieva A.M., Yelezarova Yu. N. Development of Tatar Russian Socio-Political Dictionary // I.L.Kopylov (ed.) Vocabulum et vocabularium. - Minsk: Chetyre chetverti, 2017. - Pp. 348-353.
[30] Divjak, Dagmar. (2010). *Structuring the lexicon: A clustered model for near-synonymy*. Walter de Gruyter.
[31] Ullmann, Stephen. (1967): *Semantics, An Introduction to the Science of Meaning.* Oxford: Basil Blackwell.
[32] Lyons, John. (1995): *Linguistic Semantics.* Cambridge: Cambridge University Press.
[33] Kjellmer, Goran (1994): *A Dictionary of English Collocations: Based on the Brown Corpus:* in three volumes. Oxford, Fairlawn, New Jersey: Clarendon Press: Press, 1994.