Available online at www.sciencedirect.com

ScienceDirect

Procedia
Social and Behavioral Sciences

Procedia - Social and Behavioral Sciences 198 (2015) 218 – 225

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Building dialectological corpora for Turkic languages: Mishar dialect of Tatar

Bulat Khakimov[a,b]*, Farid Salimov[a,b], Dariya Ramazanova[c]

[a] *Kazan (Volga region) Federal University, Kremlevskaya str., 18, Kazan, 420008, Russia*
[b] *Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Levo-Bulachnaya str., 36a, Kazan, 420111, Russia*
[c] *Institute of Language, Literature and Arts of the Tatarstan Academy of Sciences, Lobachevskogo str., 2/31, Kazan, 420111, Russia*

Abstract

Corpus-based dialectology of less-resourced and functionally limited native languages is a developing field of linguistics. In this paper we discuss challenges of annotating dialect corpora for Turkic languages of Russia by the example of Mishar dialect of Tatar language. Peculiarities of grammatical variability in Mishar dialect are investigated from the point of view of automatic annotation and the search functionality of the corpus is described. The proposed methodology of annotation can be used when creating multilingual integrated resources and parallel corpora of closely related languages.

Keywords: Turkic languages; dialectology; linguistic variation; corpus design; grammatical annotation; search queries

1. Introduction

Corpus is an effective method to store, preserve and investigate dialect data. Corpus-based dialectological studies represent the relatively new trend in modern Turkic and Tatar linguistics. It is valuable for applied and computational linguistics, comparative and historical turkology, typology, history and other fields of Humanities. Corpus building projects are quite relevant in contemporary Turkic studies. There are a number of Turkish corpora

* Corresponding author. Tel.: +7-917-399-2099; fax: +7-843-292-4274.
E-mail address: bulat.khakeem@gmail.com

doi:10.1016/j.sbspro.2015.07.439

of different types with different aims, one of them is Turkish National Corpus (Aksan et al., 2012). Among the well-known projects we can also mention the corpora of Kazakh (Makhambetov et al., 2013), Tatar (Suleymanov et al.,2013), Uyghur (Aibaidulla and Kim-Teng Lua, 2002), Bashkir (Buskunbaeva and Sirazitdinov, 2011), Khakassian (Sheimovich, 2011), and Tuvan (Salchak, 2012) languages. While written literary corpora of Turkic languages develop actively only the first steps are made in building corpora of dialects. On the other hand, during the last decades a large amount of text samples was accumulated by dialectologists, but most of them are not digitalized. Importance and relevance of creating corpora of Turkic dialects is supported by the fact that such projects help to preserve the authentic languages, which are strongly influenced by globalization and assimilation processes.

Problems of corpus-based dialectology are discussed actively during the last years. Special chapters of books on corpus linguistics investigate relationship between corpus linguistics and dialectology, like in Anderwald and Szmrecsanyi (2009). Corpus-based approach gives more opportunities and methodological development to dialectometry, e.g. refer to Haimerl (2006) and Szmrecsanyi (2011). There are some special projects of dialect corpora. For example, Nordic dialect corpus (Johannessen et al., 2009), Freiburg English dialect corpus (Kortmann and Wagner, 2005), the Crubadan project (Scannell, 2007), and the Syntactic Atlas of the Dutch Dialects (Barbiers et al., 2007). The Pangloss Collection includes documented data in a wide range of endangered languages across the world (Michailovsky et al., 2014).

The corpus-based studies in the Republic of Tatarstan focus on Tatar and Russian languages in comparison with other languages. For example, in Nevzorova and Salimov (2012) the model and methodology of Russian-Tatar lexicographical database is described. Bochkarev, Solovyev and Wichmann (2014) discuss using the Google Books N-Gram Corpus and propose ways of measuring changes in word frequency. Galiullin et al. (2014) describe corpus-based studies of the Kazan regiolect of Russian. Some linguistic researches were carried out using the National Corpus of Tatar language (e.g. Zamaletdinov and Galieva, 2014 and Galieva, Nevzorova and Gatiatullin, 2014). Tatevosov (2013) and Pereltsvaig and Lyutikova (2013) investigate typological structure of Tatar language using the data from their own Moscow State University collection of Mishar texts which were included to the Mishar corpus during our project.

The purpose of this study is to investigate problems of dialectological corpus annotation, dialectological database compilation, and integrated resources for corpus-oriented and computational dialectology in Turkic languages by the example of Mishar dialect of Tatar. The research is based on the Electronic Atlas of Tatar Dialects (http://atlas.antat.ru), which is the geoinformational web resource about dialects of Tatar language spoken mostly in Russian Federation by over 5 mln people. The Atlas consists of more than 200 maps and describes territorial distribution of different phonological, lexical, morphological and syntactical phenomena (Salimov et al., 2011). In 2012 we started to develop its textual extension – the corpus of Mishar dialect.

Mishar (also known as Western dialect) is one of the main Tatar dialects and it is actively used in oral communication. It is spoken in the Republic of Tatarstan and several other regions of Russia and consists of many subdialects, which have differences as from the standard Tatar and from each other (Ramazanova, 2008). We had to analyze and classify these differences in order to find ways of formal representation of dialectological phenomena.

2. Overview of Mishar corpus

The main problems of building corpora for the native languages and dialects are generally related to the following tasks:

- collection of text samples and database compilation;
- providing the representativeness and balance in the corpus collection;
- automation of the tagging process.

Each of these tasks has its own specifics and ways of realization. In our project most effort was spent on the stage of automatic tagging. As for creation of representative collection of texts, it was performed by a team of researchers from Kazan Federal University, Tatarstan Academy of Sciences and Moscow State University in 2012-2013. Mishar corpus is made accessible online as an integrated part of the Electronic Atlas of Tatar dialects supported by the Applied Semiotics Institute of Tatarstan Academy of Sciences.

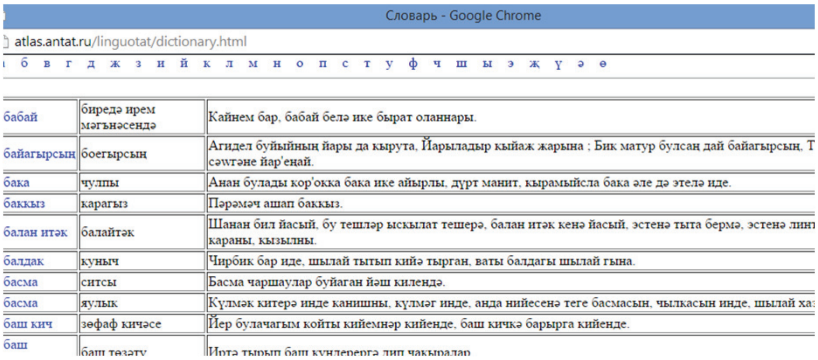
The database of the corpus is developed on the PostgreSQL platform and it includes texts recorded since 1950 until the present time and consists of about 50000 words. Dialect texts are morphologically annotated and classified according to the special set of metatags. The part of the texts is accompanied by English translation.

The structure of dialectological annotation depends on what information we want to include. The necessary meta-information in our corpus is represented by a detailed dialectological tagset, which contains information about the dialect, the place and time of recording, the informant and subject/genre characteristics of the text. The set of subdialects is aligned with classification used in the Atlas of Tatar dialects.

Dialect texts in the Mishar corpus come from different sources. It includes field trip recordings collected by the authors, folklore texts, published researches, etc. Some of them were collected during dialectological expeditions of the Tatarstan Academy of Sciences; another part comes from the collection of Moscow State University. And finally, it includes earlier recordings of the Soviet time, which were published in several compilations. These earlier published texts were scanned and recognized using OCR technology with adaptation to dialectological transcription character set. So we can say that the wide range of sources enables us to build representative corpora of Tatar dialects.

The Mishar corpus also includes a variety of integrated resources, for example, dictionaries containing information about the tags appeared in annotation in each subdialect, providing a comparative view on peculiarities of grammatical inflection in different subdialects within the Mishar dialect.

Another special resource is the corpus-based dictionary of dialectisms. It contains information about the texts and sentences in which the dialectism appears. The dictionary also includes the standard equivalents of the dialectisms, their phonetic variants and more. This dictionary is associated with the corpus, so one can select a word in the dictionary and easily find examples from the corpus.



Словарь - Google Chrome

atlas.antat.ru/linguotat/dictionary.html

б в г д ж з и й к л м н о п с т у ф ч ш ы э җ ү ә ө

бабай	биредә ирем мәгънәсендә	Кайнем бар, бабай белә ике бырат оланнары.
байагырсын	боегырсың	Агидел буйыйның йары да кырута, Йарыладыр кыйаж жарына ; Бик матур булсаң дай байагырсың, Т сәwтәне йар'еңай.
бака	чулпы	Анан булады кор'окка бака ике айырлы, дүрт манит, кырамыйсла бака әле дә этелә иде.
баккыз	карагыз	Пәрәмәч ашап баккыз.
балан итәк	балайтәк	Шанан бил йасый, бу тешләр ыскылат тешерә, балан итәк кенә йасый, эстенә тыта бермә, эстенә лин караны, кызылны.
балдак	куныч	Чирбик бар иде, шылай тытып кийә тырган, ваты балдагы шылай гына.
басма	ситсы	Басма чаршаулар буйаган йәш килендә.
басма	яулык	Күлмәк китерә инде канишны, күлмәг инде, анда нийесенә теге басмасын, чылкасын инде, шылай ха
баш кич	зәфаф кичәсе	Йер булачагым койты кийемнәр кийенде, баш кичкә барырга кийенде.
баш	баш тәзәту	Иртә тырып баш кундерергә дип чакыралар

Fig.1. Fragment of the corpus-based dictionary of dialectisms.

3. Grammatical annotation

The problem of grammatical annotation of dialect texts is particularly topical for the languages with rich inflection like Tatar and other Turkic languages. In addition to morphological features, dialect texts show significant lexical, morphological and syntactic variability. Such variability is a hindrance to the development of unified tagset for a particular dialect. However, along with the variability of grammatical structure a stable invariant can also be found in the Turkic languages and dialects. This makes the universal formal description possible to some extent and it simplifies the creation of multilingual resources and comparative study of closely related languages.

At the present time we can say that certain morphological standard has been developed in the field of the corpus annotation systems for Tatar and other Turkic languages. Generally, both the parts of speech and inflectional categories of a token are indicated. For example, the most developed two-level model of the Tatar morphology by Suleymanov and Gilimullin consists of special phonological and morphotactic rules related to the verbal and noun inflectional paradigms. This model defines the relationship between the stem and the sequence of affixes (Suleymanov et al., 2000).

It is known that in the agglutinative Turkic languages words are inflected by attaching a consequent set of morphemes to the stem. Stems usually do not change during inflection, and the affixes phonetically depend on the stem. And as a rule, every grammatical meaning is expressed by a particular affix, while affixes on the whole are regular and unambiguous. Because of this, grammatical features are easily recognized during the automatic analysis of the morphemic structure of corpus tokens.

We also take into account that according to the Tatar morphotactic rules, grammatical features are divided into the complex and simple, from the one hand, and required and optional, from the other hand. All the complex features are represented by a set of affixes associated to a grammatical category, while the simple ones are represented by a single affix. A required feature, such as noun case or number, is always explicitly or implicitly expressed in a word form. So any tag from the group of tags describing such features is always assigned to tokens of corresponding part of speech or grammatical class, even if it stands in the null form like singular and nominative case. For optional features, the grammatical meaning is not obligatory, e.g. possessives. Those features are annotated only if they are explicitly inflected.

One of the main problems of grammatical annotation for the corpora of Turkic languages and dialects is to identify the core of inflectional categories and to create the optimum meta-language of description, which would be suitable as a standard for the family of languages. In 2014, on the Uniturk workshop which was held in Kazan, this problem was discussed for the first time for Turkic languages and the special declaration was adopted (Khakimov, Galieva and Gatiatullin, 2014). In our Mishar dialect corpus, we follow this declaration and create dialectal grammatical tags taking into consideration the general Turkic background.

The completeness of the description and reasonable balance between the reality of the language and traditions of grammatical theory is also very important. For dialects of Tatar, there are different approaches to the description of grammatical phenomena and transcription. First of all, these are the authentic traditions of native Tatar dialectology (Ramazanova, 2008). According to them, texts are transcribed using the standard Tatar Cyrillic orthography with the addition of some special characters. In Tatar dialectology the grammatical structure of dialects is described based on traditional approaches in the Tatar linguistics and in comparison with the standard Tatar language and other Turkic languages. On the other hand, there are certain works based on general typological approach, where linguistic phenomena are described and explained using general theory and terminology. In those studies universal and commonly accepted Latin transcription is used (for example, Misharskij dialekt, 2007). We found it reasonable to include text samples to our corpus in their original transcription according to what source they come from, and we created two parallel tagsets reflecting different approaches mentioned above. We believe that corpus annotation and transcription should provide the convenience of search, regardless of the users' theoretical preferences. Certain decisions which allow mapping between the two schemes were also implemented.

Grammatical annotation in our dialect corpus is based on the model of the standard Tatar language and it is consistent with commonly used typological terminology and glossing rules. The core of the tagset is used for the annotation of the Tatar National Corpus (Suleymanov et al., 2013). It's full variant can be viewed on the website of

Tatar national corpus (see the list of web resources). In order to annotate specific dialectal grammatical phenomena, additional tags were developed.

As for automation of the annotation process, we investigated the opportunities of using the parser which was developed for the standard Tatar language (Suleymanov et al., 2000). It contains the related lemma list and the affix list. All items in the lemma list are distributed in 4 morphological types which are divided into number of morphonological types. Our aim was to adapt the parsing tool to the dialect variability. It was possible because the Mishar dialect is almost identical with the Tatar standard language morphology. There is no need to rewrite most of the morphotactic rules and models of inflectional paradigms. The lemma list also should be reviewed and extended with dialect words.

4. Variability in the dialectological tagset

There are certain affixes in Mishar dialect, which are not found in modern standard Tatar language. Most of such phenomena consist of dialectal equivalents of standard Tatar affixes with similar semantics. Usually the specific affixes do not replace the standard ones and they are used in parallel. We use the principle of similarity and include some groups of related tags with the core tag from the standard tagset. It should be mentioned that along with such variability some affixes are ambiguous and express different grammatical meanings in standard language and dialects. Below some examples are shown.

1) Affixes *-ıqla/-eklä, -qla/-klä, -ğalaqla/-gäläklä*, which derive raritive – a special modal verb form with the meaning of doing something rarely, from time to time (*uqıqla* – to read from time to time, *süläklä* – to tell sometimes, *kergäläklä* – to visit from time to time, etc.). The standard variant is *-ğala/-gälä, -ıştır/-eşter*. In this case we annotate the dialectal affixes accordingly to the standard tagset by adding a number to the basic tag. For example, standard affix *-ğala/-gälä* has the tag RAR, and the dialectal variant is annotated as RAR1.

2) *-ğı keli/-ge keli, -qı keli /-ke keli, -k keli* is the dialectal equivalent of standard construction with the meaning ‘want/wish to do something’ (*-ası kılä/-äse kılä)*. It consists of obligative affix (*-ğı/-ge/-qı/-ke/-k* vs. *ası/-äse*) and auxiliary verb *kelä* (dialect) or *kil* (standard). The affixes in this construction are annotated as OBL and OBL1 respectively.

3) *-dır/-der, -tır/-ter* and other allomorphs are equivalent to null form in standard Tatar and expresses 3rd person in verbs. This affix is annotated as 3SG.

4) There is a variety of infinitive forms in different Mishar subdialects. All of them are annotated in the similar way. Standard variant *-(ı)rğa/-(e)rgä* corresponds to the commonly used INF tag. Other tags for dialectal infinitives are given in Table 1.

5) The affix *-ın/-en* is ambiguous in Mishar dialect. Similar to standard Tatar, it derives “seasonal” adverbs: *yazın* (in spring), *qışın* (in winter), *kıçen* (in the evening), *irtän* (in the morning). As an inflectional category it also forms genitive case (GEN1 tag) along with the standard genitive (*-nıñ/-neñ*).

6) Past indefinite (resultative) tense is usually expressed by *-ğan/-gän* in standard Tatar. In certain Mishar subdialects we can find also the archaic Turkic *-ıp/-ep*, which is homonymical with converb inflection in modern Tatar. Such a complicated ambiguity is a challenge for annotators.

7) Another kind of dialectal variability in word structure is related with the manner of performing folklore songs. It is typical that additional vowels or syllables appear within the word form to fill the rhythmical gaps or express emotions. Although those elements are not exactly morphological, we find it reasonable to reflect them in corpus annotation using special PHON tag.

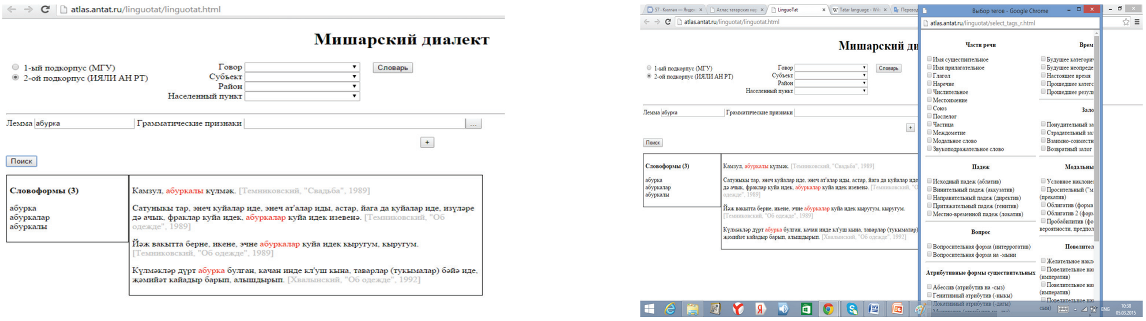
Table 1. Additional grammatical tags for Mishar dialect.

Tag	Explaination	Affixes	Standard equivalent	Parallel usage	Ambiguity
RAR1	raritive	*-ıqla/-eklä, -qla/-klä, -ğalaqla/-gäläklä*	*-ğala/-gälä, -ıştır/-eşter*	yes	no
OBL1	obligative	*-ğı keli/-ge keli, -qı keli /-ke keli, -k keli*	*-ası kılä/-äse kılä*	yes	no
3SG	3rd person	*-dır/-der, -tır/-ter*	null form	yes	yes
INF1	infinitive	*-mağa/-mägä*	*-(ı)rğa/-(e)rgä*	yes	no
INF2	infinitive	*-ğalı/-gäle*	*-(ı)rğa/-(e)rgä*	yes	no
INF3	infinitive	*-malı/-mäle*	*-(ı)rğa/-(e)rgä*	yes	yes
INF4	infinitive	*-ma/-mä*	*-(ı)rğa/-(e)rgä*	yes	yes
GEN1	genitive	*-ın/-en*	*-nıñ/-neñ*	yes	yes
PST.INDF1	past indefinite	*-ıp/-ep*	*-ğan/-gän*	yes	yes
PHON	phonological element	*ay, way, la,* etc.	-	no	no

5. Search functionality

From the point of view of implementation, our Mishar corpus is an indexed set of word forms. Indexes determine to which sentence and text a particular token belongs, and each token has its grammatical annotation. Two tables are used to represent morphemic structure: table of lexemes and table of affixes with the corresponding tags. These tables are related, and this allows users to search for lemma and for a set of grammatical features using a special interface.

All search requests in Mishar corpus are represented by two options. The main option is searching by word (lemma) with (or without) its grammatical features (Fig.2a). As a result, examples of contexts that meet the search terms are displayed. As the second option, user can specify a set of grammatical features using special checkboxes (Fig.2b). Then the search query is generated and executed. According to parameters specified by user, the list of relevant contexts is shown.



○ 1-ый подкорпус (МГУ)
● 2-ой подкорпус (ИЯЛИ АН РТ)

Говор
Субъект
Район
Населенный пункт

Лемма абурка Грамматические признаки

Словоформы (3)

абурка
абуркалар
абуркалы

Камзул, абуркалы кулмәк. [Темниковский, "Свадьба", 1989]

Сатунык̧ы тар, энеч куйалар иде, энеч ат'алар иды, астар, йага да куйалар иде, изүләре дә ачык, фраклар куйа идек, абуркалар куйа идек изевенә. [Темниковский, "Об одежде", 1989]

Йәж вакытта берне, икене, эчне абуркалар куйа идек кырутум, кырутум. [Темниковский, "Об одежде", 1989]

Күлмәкләр дүрт абурка булган, качан инде кл'уш кына, таварлар (тукымалар) бәйә иде, җәмийәт кайадыр барып, алышдырып. [Хвалынский, "Об одежде", 1992]

Части речи

□ Имя существительное
□ Имя прилагательное
□ Глагол
□ Наречие
□ Числительное
□ Местоимение
□ Союз
□ Послелог
□ Частица
□ Междометие
□ Модальное слово
□ Звукоподражательное слово

Падеж

□ Исходный падеж (аблатив)
□ Винительный падеж (аккузатив)
□ Направительный падеж (директив)
□ Притяжательный падеж (генитив)
□ Местно-временной падеж (локатив)

Вопрос

□ Вопросительная форма (интеррогатив)
□ Вопросительная форма на -мыни

Атрибутивные формы существительных

□ Абессив (атрибутив на -сыз)
□ Генитивный атрибутив (-ныкы)
□ Локативный атрибутив (-дагы)

Врем

□ Будущее категори
□ Будущее неопреде
□ Настоящее время
□ Прошедшее катего
□ Прошедшее резул

Зало

□ Понудительный за
□ Страдательный за
□ Взаимно-совмести
□ Возвратный залог

Модальны

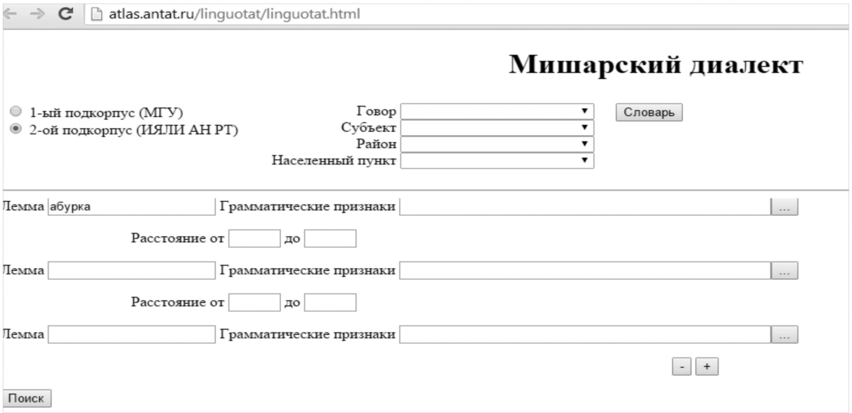
□ Условное наклоне
□ Просительный ("м (прекатив)
□ Облигатив (форма
□ Облигатив 2 (фор
□ Пробабилитив (фо вероятности, предпол

Повелител

□ Желательное накл
□ Повелительное на (императив)
□ Повелительное на (императив)
□ Повелительное на сын)

Fig. 2. Search results (a) for lemma and (b) for grammatical features with the tag selection window.

In order to perform comparative analysis of dialect data, user can create a custom subcorpus of texts related to a certain subdialect, genre, theme, or geographic region using corresponding filters in the metadata table. Also it is possible to search for collocations (see Fig.3).The appropriate search procedures were developed and tested. The interface of the corpus also includes tools for viewing integrated dictionaries and accessing corpus examples from within the dictionaries.



atlas.antat.ru/linguotat/linguotat.html

Мишарский диалект

○ 1-ый подкорпус (МГУ)
● 2-ой подкорпус (ИЯЛИ АН РТ)

Говор
Субъект
Район
Населенный пункт

Лемма: абурка
Грамматические признаки:
Расстояние от до
Лемма:
Грамматические признаки:
Расстояние от до
Лемма:
Грамматические признаки:

Fig. 3. Search form for collocations by lemma and grammatical features.

6. Conclusion and future work

As a further development of the Mishar dialect corpus we plan to increase the amount of text collection, provide more detailed annotation, improve tagging process and search functionality, and implement additional integrated resources. Also we aim to extend our dialectological corpus to other Tatar dialects.

We hope that results and methods of our corpus-based study of Tatar dialects can be used to create annotated corpora of other Turkic languages, including parallel corpora and multilingual comparative resources.

Acknowledgements

This work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

Aibaidulla, Y., and Kim-Teng, L. (2003). The Development of Tagged Uyghur Corpus. *Proceedings of PACLIC17, 1-3 October 2003, Sentosa, Singapore* (pp. 228-234). Singapore: COLIPS publications.

Aksan, Y. et al. (2012). Construction of the Turkish National Corpus (TNC). In Nicoletta Calzolari et al. (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012).* Istanbul: European Language Resources Association. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/papers.html

Anderwald, L., and Szmrecsanyi, B. (2009). Corpus linguistics and dialectology. *Corpus Linguistics. An International Handbook. Handbücher zur Sprache und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science*. Berlin/New York: Mouton de Gruyter.

Barbiers, S., Cornips, L., and Kunst, J. P. (2007). The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas. *Creating and digitizing language corpora: Synchronic databases,* 54 - 90.

Bochkarev, V., Solovyev, V., and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of the Royal Society Interface*, *11/101*.

Buskunbaeva, L., and Sirazitdinov Z. (2011). The System of Annotation in the National Corpus of Bashkir Language (Sistema razmetok v natsional'nom korpuse bashkirskogo jazyka). *Proceedings of the International Conference "Languages of Minorities in Computer Technologies: Experience, Tasks and Perspectives"* (pp. 46-51). Yoshkar-Ola: Mari El Ministry of Culture. In Russian.

Electronic atlas of Tatar dialects: http://atlas.antat.ru.

Galieva, A., Nevzorova, O., and Gatiatullin, A. (2014) Towards Building Wordnet for the Tatar Language. *Communications in Computer and information Science*, *468*, 57 - 66.

Galiullin, K., Gizatullina, A., Gorobets, E., Karimullina, G., Karimullina R., and Martyanov, D. (2014). Corpus-based regiolect studies: Kazan region. *Lecture Notes in Computer Science*, *8773*, 169 - 175.

Haimerl, E. (2006) Database Design and Technical Solutions for the Management, Calculation, and Visualization of Dialect Mass Data. *Literary and Linguistic Computing*, *21/4*, 437 - 444.

Johannessen, J. B., Priestley, J., Hagen, K., Afarli, T. A., and Vangsnes, A. (2009). The Nordic dialect corpus-An advanced research tool. In *Proceedings of the 17th Nordic conference of computational linguistics NODALIDA 2009. NEALT proceedings series*, 4.

Khakimov, B., Galieva, A., and Gatiatullin, A. (2014) To the problem of unification of the annotation systems of grammatical categories in the corpora of Turkic languages. *Proceedings of the International Conference on Turkic Language Processing (TURKLANG-2014)* (pp. 131-135). Istanbul: Özkaracan Matbaacılık-Bağcılar.

Kortmann, B., and Wagner, S (2005). The Freiburg English dialect project and corpus. In: Kortmann, B., Herrmann, T., Pietsch, L., and Wagner, S. (eds). *The Comparative Grammar of British English Dialects. Agreement, Gender, Relative Clauses.* Berlin/New York : Mouton de Gruyter (pp. 1-20).

Lyutikova, E., Kazenin, K., Solovyev, V., and Tatevosov, S., eds. (2007). *Mishar Dialect of Tatar Language: Essays on Syntax and Semantics (Misharskij dialekt tatarskogo jazyka: ocherki po sintaksisu i semantike).* Kazan: Magarif. In Russian.

Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov B., Sabyrgaliyev, I., and Sharafudinov, A. (2013) Assembling the Kazakh Language Corpus. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, pp. 1022-1031.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., Adamou, E. (2014) Documenting and Researching Endangered Languages: The Pangloss Collection. *Language Documentation and Conservation*, Vol. 8, pp. 119-135.

Nevzorova, O., and Salimov, F. (2012) Model of lexicographical database: structure, basic functionality, implementation. *Information Models and Analyzes,* Vol. 1, pp. 21-27.

Pereltsvaig A., and Lyutikova, E. (2013) Elucidating Nominal Structure in Articleless Languages: A Case Study of Tatar, *Proceedings of 38th Berkeley Linguistic Society Meeting*, Berkeley.

Ramazanova, D. (2008). *Problems of Tatar dialectology (Voprosy tatarskoj dialektologii).* Kazan: Alma-Lit. In Russian.

Salchak, A. (2012). Electronic Corpus of the Tuvan Language. *New Researches in Tuva,* Vol. 3. Retrieved from URL: http://www.tuva.asia/journal.

Salimov, F., Ramazanova, D., Pilyugin, A., and Salimov, R. (2011) Elektronnaja versija atlasa tatarskikh narodnykh govorov. *Vestnik TGGPU,* Vol.4(26), pp. 205-210. In Russian.

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.

Sheimovich, A. (2011). Morphological Annotation of the Corpus of the Hakass Language. *Russian Turkology,* 2(5), 48-61. In Russian.

Suleymanov, D., Guilmoulline, R., and Guilmoulline, A. (2000) Tatar phonological rules as a base of two-level morphological analyzer, *Proceedings of LP'2000*, B.Palek and O.Fujimura (eds.), Prague: The Karolinum Press, pp. 495-504.

Suleymanov, D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., and Khakimov, B. (2013) National corpus of the Tatar language "Tugan Tel": Grammatical Annotation and Implementation. *Procedia Social and Behavioral Sciences*, Vol. 95, pp. 68-74.

Szmrecsanyi, B. (2011) Corpus-based dialectometry: a methodological sketch. *Corpora*, Vol.6, Issue 1, pp. 45-76.

Tatar national corpus "Tugan tel": http://web-corpora.net/TatarCorpus/search/index.php?interface_language=en

Tatar grammatical tagset: http://webcorpora.net/TatarCorpus/search/frame_parts/gramsel.php?interface_language=enandsearch_language=tatarandcontexts_output_language=tatar

Tatevosov, S. (2013) Decomposing event structure: Evidence from denominal verbs in Tatar, *MIT Working Papers in Linguistics: Ozge, Umut (ed.) Proceedings of the 8th Workshop on Altaic Formal Linguistics,* MITWPL, Cambridge, pp. 349-360.

Zamaletdinov, R., and Galieva, A. (2014). Tatar Possessive Verbs with the Meaning Component "Part of Plant". *Middle East Journal of Scientific Research,* Vol.21(1), pp.229-233.