

BUILDING DIALECTOLOGICAL CORPORA FOR TURKIC LANGUAGES (MISHAR DIALECT OF TATAR)

Bulat Khakimov

Farid I. Salimov

Dariya B. Ramazanova

Kazan Federal University

Tatarstan Academy of Sciences

Kazan, Tatarstan, Russia

Outline

2

- General background information
- Turkic and Tatar corpora
- Corpus-based dialectology
- Project of Mishar Corpus
- Database
- Dialectological annotation
- Search queries
- Integrated resources
- Conclusion and future work

Republic of Tatarstan, Russia

3



- Population – 4 mln. (Tatars – 53%, Russians – 40%)
- Kazan – former capital of Kazan Khanate (conquered in 1552)
- FC Rubin Kazan))

Tatars in Russia

4



- A Turkic and Muslim nation
- Tatarstan – 2 mln. (~36%)
- Other regions – 3,5 mln (5,5 mln total, 2nd place in Russia)

Tatar language

5

- A Turkic language spoken by Volga Tatars
- Should not be confused with Crimean Tatar
- Grammar: agglutinative
- Official language in the Rep. of Tatarstan (along with Russian)
- Writing and alphabet: Arab (since X century), Latin (in 1927-1939), Cyrillic (since 1939)
- Three main dialects: Middle, **Western (Mishar)**, Siberian.
- **Mishár dialect**
 - **actively used in oral communication**
 - **differs from standard Tatar mostly in phonetics and lexicon, grammar is almost the same**

Corpora of Tatar language

6

- **National Tatar Corpus “Tugan Tel”**
 - http://web-corpora.net/TatarCorpus/search/index.php?interface_language=en
 - expected to be 80 mln to the end of 2015
 - morphologically annotated
 - semantic annotation is now developed
 - Includes different domains of texts

- **Written corpus of Tatar**
 - <http://corpus.tatfolk.ru>
 - 116 mln (reported in 2014)
 - not annotated
 - includes texts from the web and literature

Other Turkic Corpora

7

- ***Turkish***
- ***Uyghur (China)***
- ***Kazakh***
- ***Kyrgyz***
- ***Bashkir (Russia)***
- ***Khakass (Russia)***
- ***Tuvan (Russia)***
- ***Crimean Tatar***

Corpora of Dialects

- Corpus-based dialectological studies represent the relatively new field in modern Turkic and Tatar linguistics.
- While written literary corpora of Turkic languages develop actively only the first steps are made in building corpora of dialects.

Corpus of the Mishar Dialect

Started in 2012 as a part of two projects:

- **Tatar National Corpus.**

which is expected to develop as an integrated textual base for Tatar language

- **Electronic Atlas of Tatar Dialects.**

a project of the geoinformational web resource about dialects of Tatar language, consists of more than 200 maps and describes territorial distribution of different phenomena.

Database of the Corpus

10

- ❑ texts recorded since 1950 until nowadays
- ❑ about 50000 words
- ❑ texts are classified according to the special set of metatags
- ❑ part of the texts is accompanied by English translation
- ❑ texts are morphologically annotated and glossed
- ❑ developed on the basis of PostgreSQL

Database of the Corpus

11

Sources:

- ❑ dialectological expeditions of the Tatarstan Academy of Sciences
- ❑ collection of Moscow State University
- ❑ earlier recordings of the Soviet time (printed, adaptive OCR recognition needed)

Meta-information in database

12

- ❑ subdialect
- ❑ region
- ❑ the place and time of recording
- ❑ the informant
- ❑ subject/genre characteristics of the text

Meta-information in database

13

postgresql что это x Атлас татарских нар x LinguoTat x W Tatar language - Wiki x Переводчик Google x Tatar Corpus x

atlas.antat.ru/linguotat/linguotat.html

Мишарский диалект

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

Говор

Субъект

Район

Населенный пункт

Лемма Грамматические признаки

Dialectal Grammatical Annotation

14

- ❑ each token has its grammatical annotation
- ❑ based on the model of the Standard Tatar language (Tatar National Corpus tagset)
- ❑ consistent with commonly used typological terminology and glossing rules (Leipzig, etc.)
- ❑ additional tags are used for specific dialectal grammatical phenomena

Search queries

15

User can specify a query using a special interface:

The screenshot shows a web browser window with the URL `atlas.antat.ru/linguotat/linguotat.html`. The page title is "Мишарский диалект". On the left, there are two radio buttons for selecting a corpus: "1-ый подкорпус (МГУ)" and "2-ой подкорпус (ИЯЛИ АН РТ)". To the right, there are four dropdown menus for "Говор", "Субъект", "Район", and "Населенный пункт". A "Словарь" button is located to the right of these menus. Below the menus, there are two input fields: "Лемма" and "Грамматические признаки". A "+" button is positioned below the "Грамматические признаки" field. At the bottom left, there is a "Поиск" button.

Search queries

16

1. Search by lemma:

← → ↻ atlas.antat.ru/linguotat/linguotat.html

Мишарский диалект

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

Говор
Субъект
Район
Населенный пункт

Лемма Грамматические признаки

Словоформы (3) абурка абуркалар абуркалы	<p>Камзул, абуркалы күлмәк. [Темниковский, "Свадьба", 1989]</p> <p>Сатуныкы 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]</p> <p>Йәж вакытта берне, икене, эчне абуркалар куйа идек кыругум, кыругум. [Темниковский, "Об одежде", 1989]</p> <p>Күлмәкләр дүрт абурка булган, качан инде кл'уш кына, таварлар (тукымалар) бәйә иде, жәмийәт кайадыр барып, алышдырып. [Хвальинский, "Об одежде", 1992]</p>
--	---

Search queries

17

2. Search by grammatical features:

The screenshot displays the LinguoTat search interface. The main search page shows the word 'абурка' and its grammatical features. A detailed view of the word is shown in a separate window, listing various grammatical features and their corresponding forms.

Мишарский диалект

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

Говор: [dropdown]
Субъект: [dropdown]
Район: [dropdown]
Населенный пункт: [dropdown]

Словарь

Лемма: абурка | Грамматические признаки: [dropdown]

Поиск

Словоформы (3)

абурка
абуркалар
абуркалы

Камзул, **абуркалы** күлмэк. [Темниковский, "Свадьба", 1989]

Сатуныкы тар, энеч куйалар иде, энеч аг'алар иды, астар, йага да куйалар иде да ачык, фраклар куйа идек, **абуркалар** куйа идек изевенә. [Темниковский, "Об одежде", 1989]

Йәж вакытта берне, икене, энче **абуркалар** куйа идек кыругум, кыругум. [Темниковский, "Об одежде", 1989]

Күлмәкләр дүрт **абурка** булган, качан инде кл'уш кына, таварлар (тукымалар) жәмийәт 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

Врем

- Будущее категори
- Будущее неопреде
- Настоящее время
- Прошедшее категс
- Прошедшее резул

Зало

- Понудительный за
- Страдательный за
- Взаимно-совместн
- Возвратный залог

Падеж

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

Модальны

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

Вопрос

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

Повелител

- Желательное накл
- Повелительное на (императив)
- Повелительное на (императив)
- Повелительное на (императив)
- Повелительное на (императив)

Атрибутивные формы существительных

- Абессив (атрибутив на -сыз)
- Генитивный атрибутив (-ныкы)
- Локативный атрибутив (-дагы)
- Местно-временной атрибутив (-дагы)

Search queries

18

3. Search for collocations:

← → ↻ atlas.antat.ru/linguotat/linguotat.html

Мишарский диалект

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

Говор

Субъект

Район

Населенный пункт

Словарь

Лемма Грамматические признаки

Расстояние от до

Лемма Грамматические признаки

Расстояние от до

Лемма Грамматические признаки

- +

Поиск

Integrated resources

19

Corpus-based dictionary of dialectisms

- ❑ contains information about the texts and sentences in which the dialectism appears
- ❑ includes the literary equivalents of the dialectisms
- ❑ phonetical variants
- ❑ associated with the corpus

Integrated resources

20

Corpus-based dictionary of dialectisms

Словарь - Google Chrome

atlas.antat.ru/linguotat/dictionary.html

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

бабай	биредә ирем мәгънәсендә	Кайнем бар, бабай белә ике бырат оланнары.
байагырсың	боегырсың	Агидел буйыйның йары да кырута, Йарыладыр кыйаж жарына ; Бик матур булсаң дай байагырсың, Т сәүгәне 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 тыга бермә, эстенә линт 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 төзәтү	Иртә тырып баш күнлеренгә лип чакыраһан

Integrated resources

21

Corpus-based dictionary of dialectisms

← → ↻ atlas.antat.ru/linguotat/linguotat.html

Мишарский диалект

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

Говор
Субъект
Район
Населенный пункт

Словарь

Лемма | Грамматические признаки ...

+

Поиск

Словоформы (8) Бабайлар бабай бабайга бабайлар бабайларны бабайларым бабайнын бабайның	<p>Йөзег алганда буш килмәделәр, китерде бәләш, бәләш өстендә миңа йаулык, бабайга хайыр, ул энде үлгән, йә йөслөк. Йә кийәү гастиничы: йегерме манитлык духи. «севәм сине» атлы и гарнитур, саручка энде, аннан суң бер платук чәчәкле, йалтырый эмән инде алтын. [Темниковский, "Свадьба", 1989]</p> <p>Бабакайнын энесе — йәш бабай. [Лямбирский, "Термины родства", 1989]</p> <p>Абзий бабайның быраты була, катнысы — чибәр йыңкай. [Лямбирский, "Термины родства", 1989]</p> <p>Балаларга г'афик бирсен, тары кылактан, баланы йәштән эткөннәр бабайлар, тары чыкканда ук кылакланып чыга. [Хвальинский, "О жизни", 1992]</p>
---	---

Conclusion and future work

22

- ❑ increasing of the text base amount
- ❑ providing more detailed annotation
- ❑ implementation of additional integrated resources
- ❑ including other dialects
- ❑ further integration with the Atlas of Tatar dialects

¡Muchas gracias por su atención!

Thanks for your attention!

Игътибарыгыз өчен рәхмәт!

Спасибо за внимание!