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

- 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

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

Tatars in Russia

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

Tatar language

- 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

- ***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
  - morhologically 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

- ***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

- 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

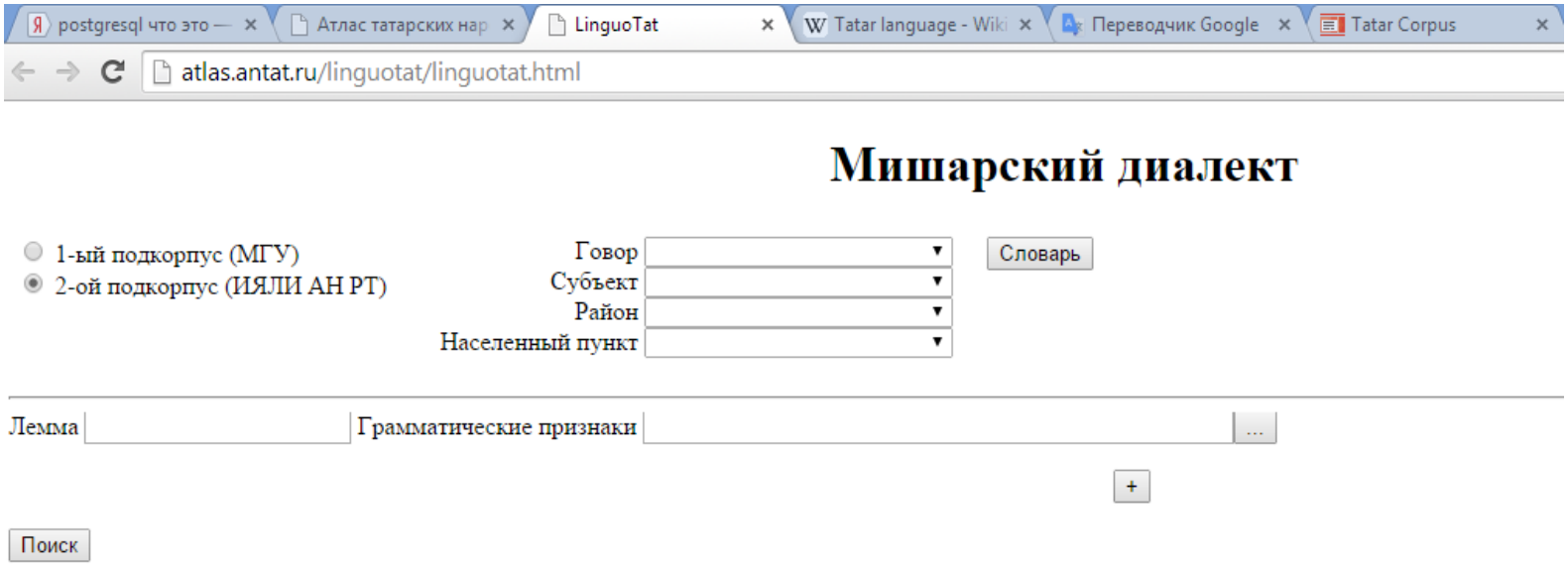
Sources:

- 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

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

Meta-information in database

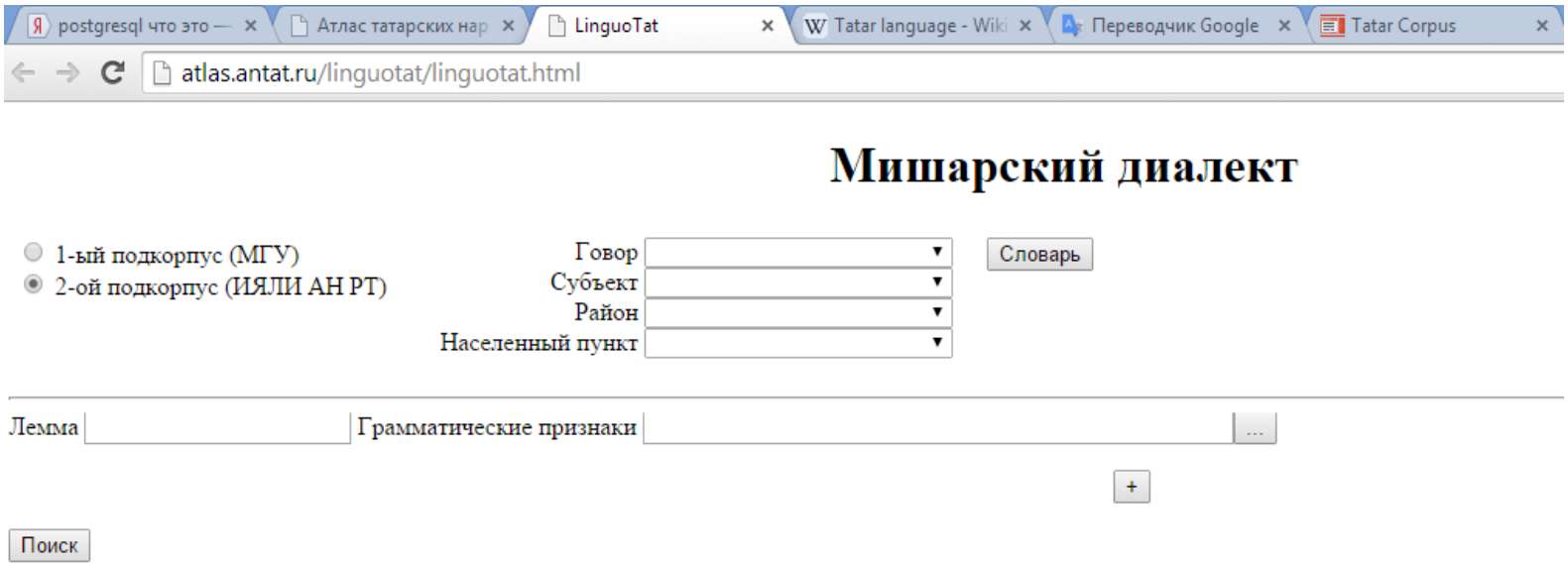


Dialectal Grammatical Annotation

- 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

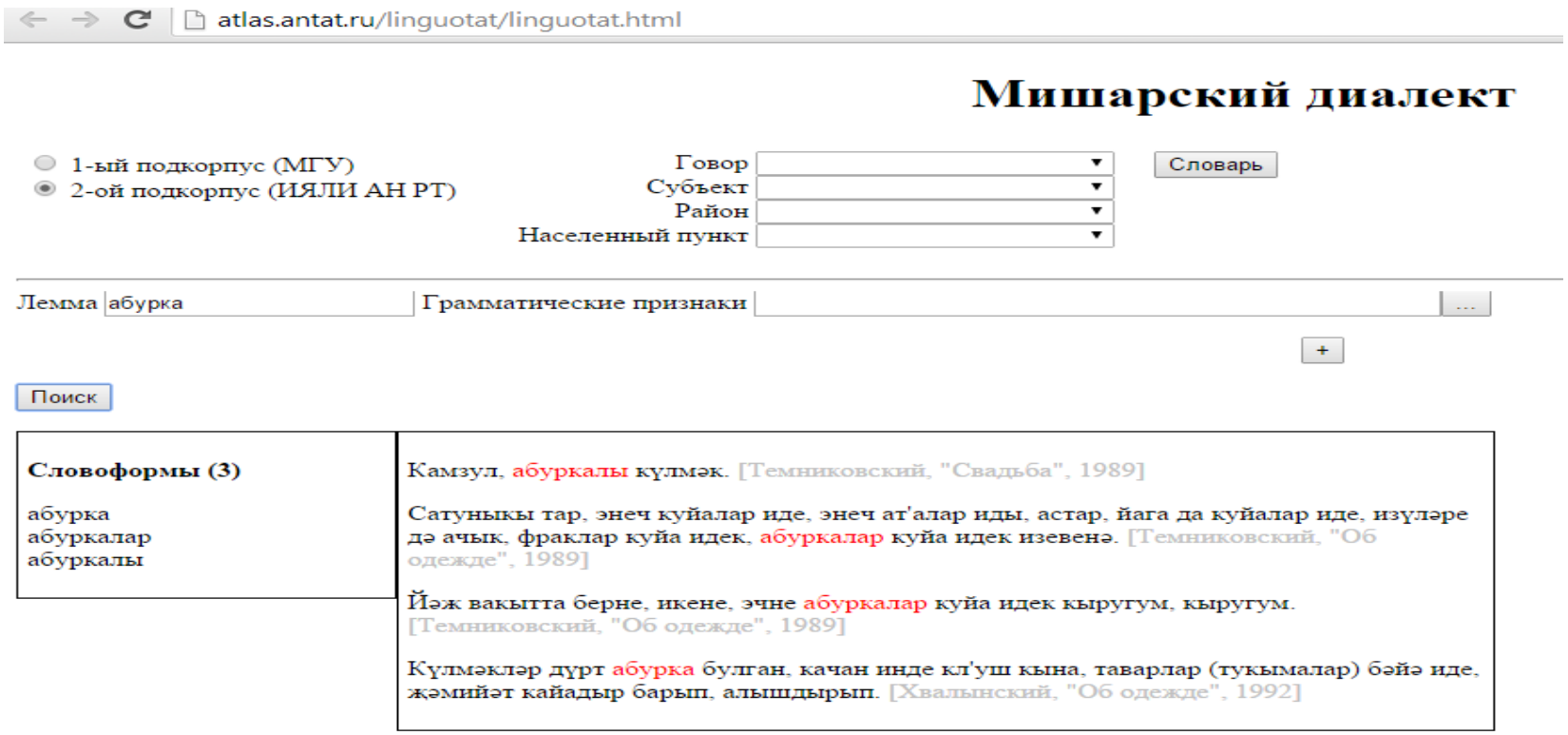
Search quieries

User can specify a query using a special interface:



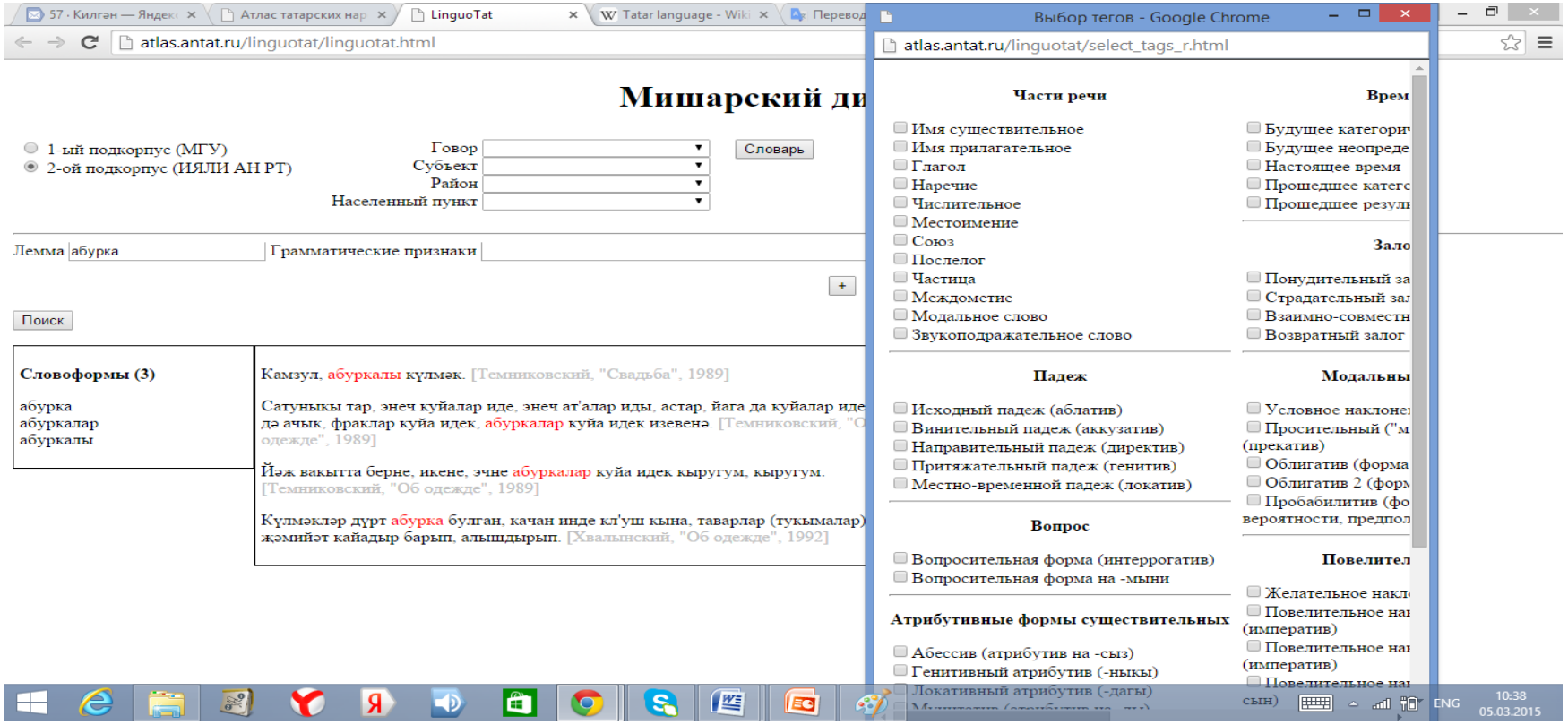
Search quieries

1. Search by lemma:



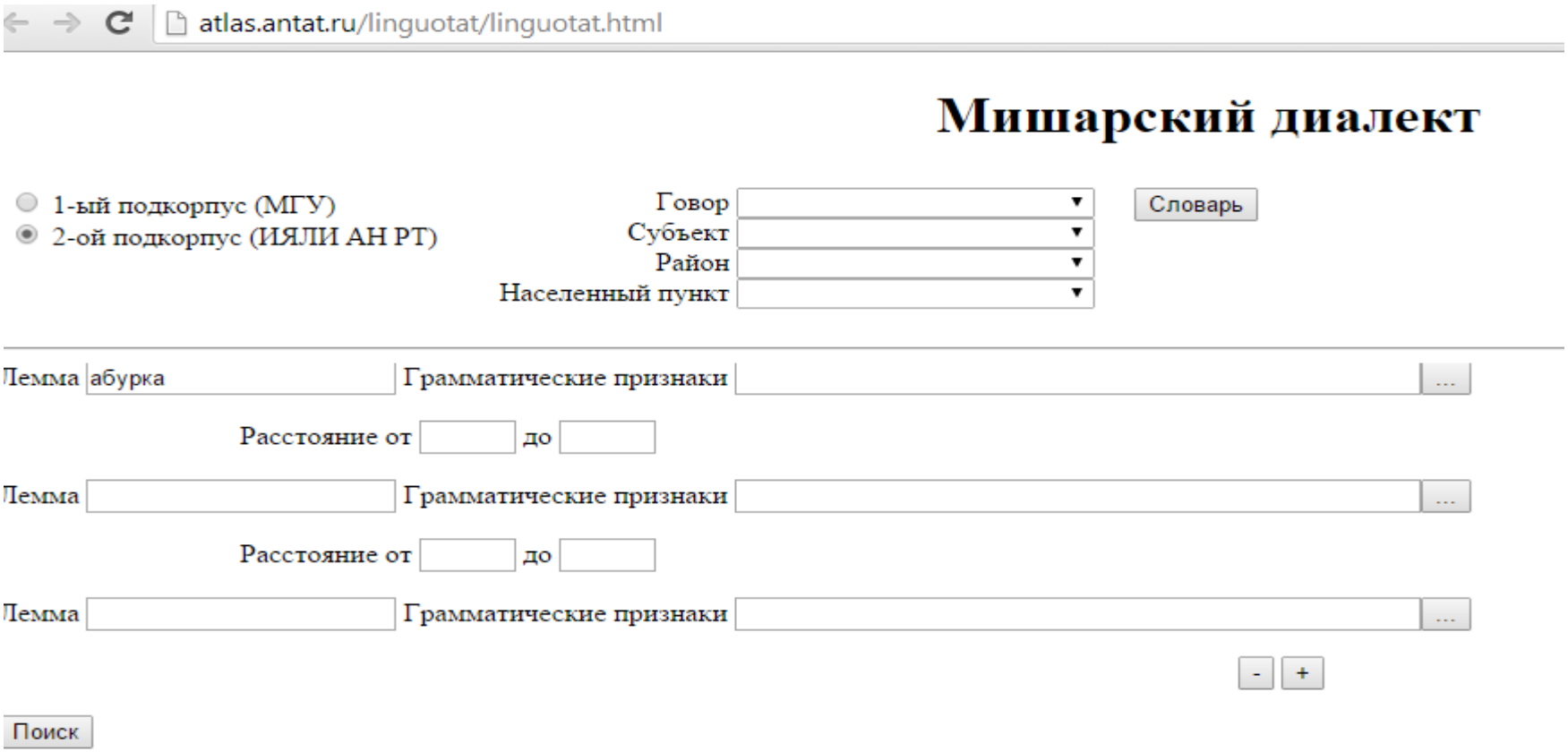
Search quieries

2. Search by grammatical features:



Search quieries

3. Search for collocations:



Integrated resources

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

Corpus-based dictionary of dialectisms

Словарь - Google Chrome

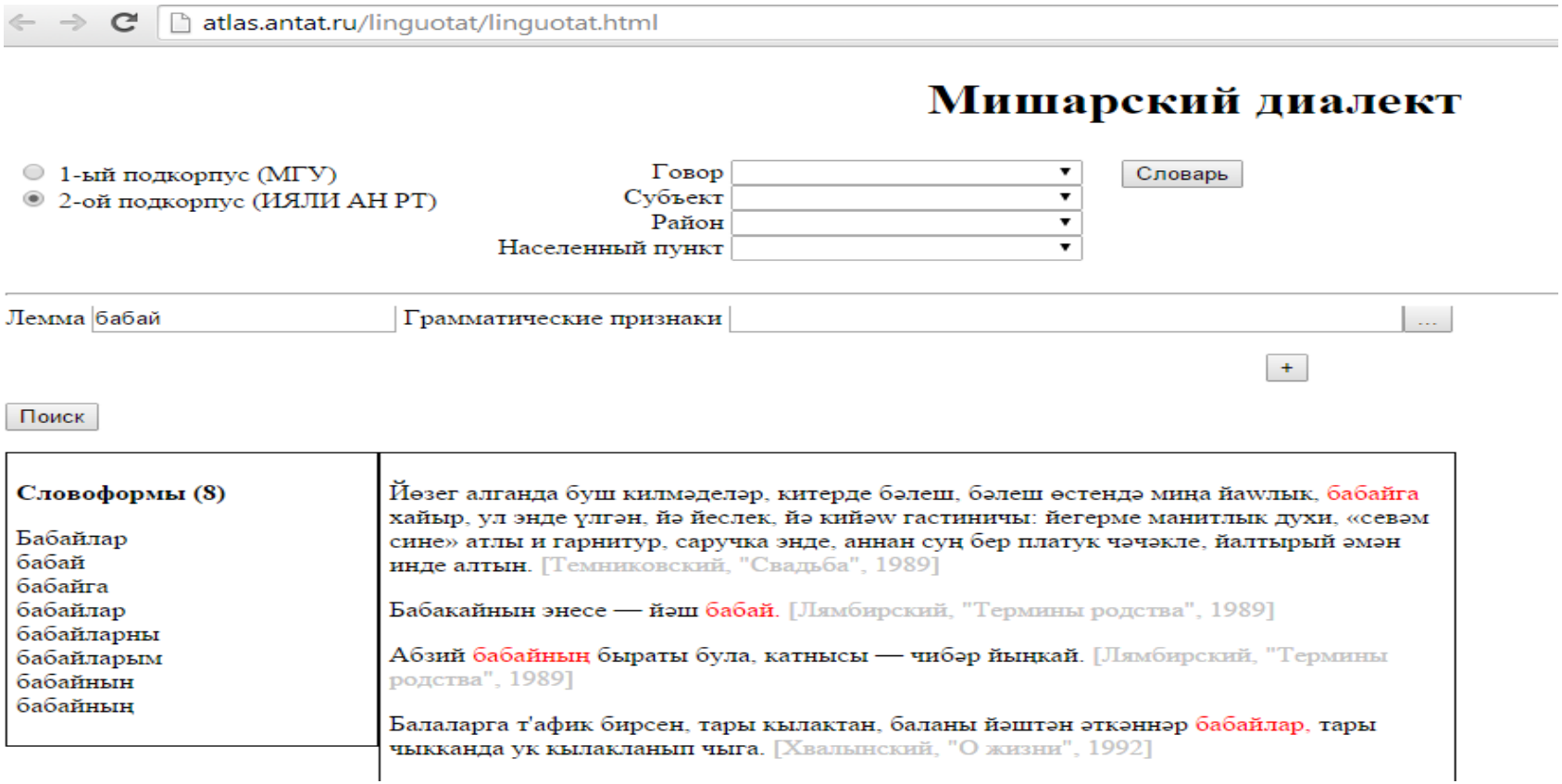
atlas.antat.ru/linguotat/dictionary.html

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 ә 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 сәwтәне йар'енай.
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 киландә.
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 төзәту	Иртә тырып баш күндерергә дип чакыралар

Integrated resources

Corpus-based dictionary of dialectisms



Conclusion and future work

- increasing of the text base amount
- providing more detailed annotation
- implemention 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!