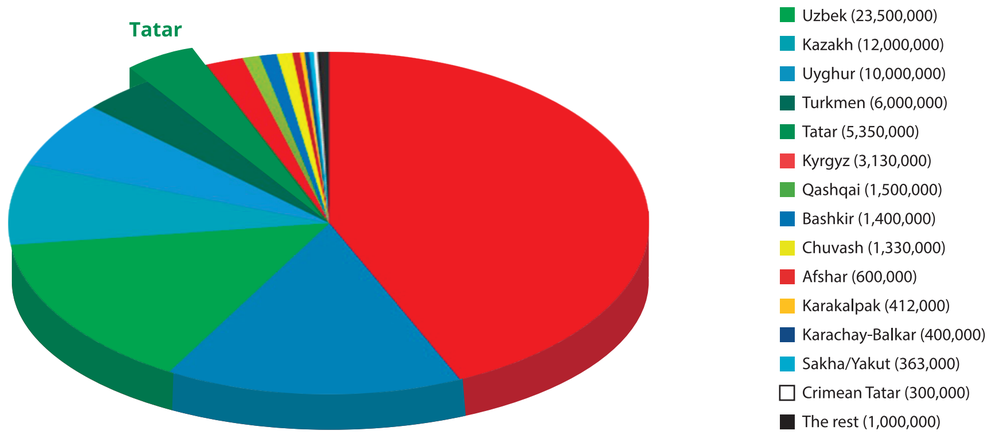
Software and Linguistic Resources for Tatar language preservation and development: Regional Experience

Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan
D. Suleymanov, R.Gilmullin, A.Khusainov
ipsanrt@gmail.com

1.The Tatar language

Native speakers: ~ 5,35 M

The Tatar language belongs to the Turkic group that forms the subfamily of Altaic languages. The Tatar language is spoken in West-central Russia (in the Volga region) and southern parts of Siberia. Different dialects of Tatar can be identified: Western, Kazan (Middle) and Eastern. In 2013, the existing language classifications described the Tatar language as an under-resourced language.



2.Tatar localization of the Information Technology

The development of terms, concepts and standards for Tatar Language in Cyberspace

Support of Tatar Language in Windows and its applications (Joint with Microsoft)

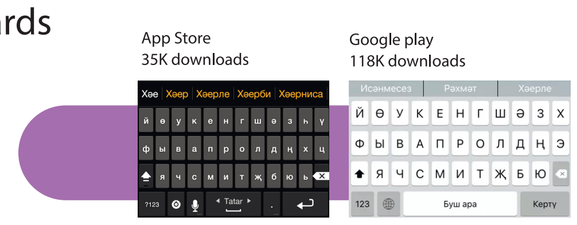
- Language interface Pack for OS Windows and Office
- Localization Volume: Texts of the Interface ~ 700K words; Terms ~ 5K.

Mobile applications

- Tatar localization of the Sailfish OS
- Tatar localization of the «Mobile app «Uslugi RT»
  - more 20K Terms and Phrases Localized.
- Russian-Tatar Dictionary TatDict
  - more 60,000 words; App Store, Google Play.
- English-Russian-Tatar-Chuvash Dictionary of IT-Terms
  - interpretation of Concepts in Tatar
  - more 6K terms
- Tatar keyboards

App Store
35K downloads

Google play
118K downloads



3.The Development Software, Linguistic resources

The Electronic version of the Atlas of Tatar dialects; the e-learning programs; a national corpus “Tugan Tel”; a synthesizer of Tatar speech; a recognizer of Tatar speech; a Tatar-Russian machine translator.

The Electronic version of the Atlas of Tatar dialects

The electronic version of the Atlas of the Tatar dialects of the Volga, Ural regions and Siberia
- 215 language phenomena in 1047 settlements
- Features of Tatar dialects on phonetics, morphology, vocabulary and syntax
- web-site: atlas.antat.ru

Online Tatar language learning platforms “Ana tele”, “Tatar Telle Zaman”, “Tatar-Online”, “Tatar tele 5”

- linguistic games
- 123 interactive exercises
- automatic pronunciation check module
- web-site: tol.edu.tatar.ru

Tatar Speech Synthesis

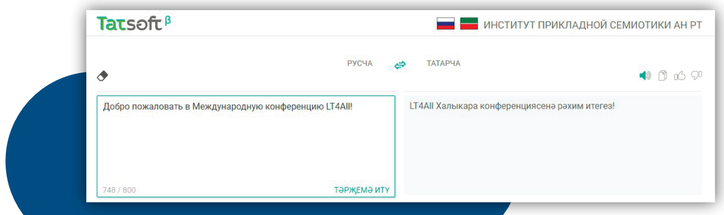
-Neural synthesis
-Natural sounding real-time speech synthesis
-17 hours studio recording training corpus
-Sentence splitting
-Text normalization (numbers, English words…)
-Basic audio player possibilities. Download audio.
Two language interface (Tatar, Russian)
-Web-site: speech.tatar

Tatar National Corpus “Tugan Tel”

TNC “Tugan Tel” includes a search platform and a software package for the linguistic statistical study of the Language Corpus and the DataBase 200M wordforms
- web-site: tugantel.tatar

Russian-Tatar Neural Machine Translation “Tatsoft”

- The Transformer model architecture
- 983,319 pairs sentences
- web-site: translate.tatar



Tatar Speech Recognition

- 88% word accuracy
- Real-time speech recognition
- Unique 100h multispeaker Tatar corpus of reading speech
- Web-site: speech.tatar