TurkLang'2017, V-th International Conference,
Kazan, 2017, October, 18

The activities of the Institute of Applied Semiotics on the maintenance of the Tatar Language in IT

Dzhavdet Suleymanov

Academy of Science of the Republic of Tatarstan
Kazan state university, Russia

ips.antat.ru
dvdt.slt@gmail.com

Outlook

- Introduction
- Tatar localization of computer systems
- Linguistic resources and Soft tools for Tatar
- Conclusion

Introduction

TurkLang 2013 - Astana, Kazakhstan, 3/10/2013 - 4/10/2013

"Artificial intelligence« Scientific-Research Institute
L.N. Gumilyov Eurasian National University

TurkLang 2014 - Istanbul, Turkey, 6/11/2014 - 7/11/2014

Technical University of Istanbul (ITU)
Faculty of Computer Engineering and Informatics

TurkLang 2015 - Kazan, Tatarstan, Russia, 17/09/2015 -19/09/2015

Tatarstan Academy of Sciences

Research Institute of Applied Semiotics

Kazan Federal University

Institute of Philology and Intercultural Communication
Institute of Computational Mathematics and Information Technologies
The Higher Institute for Information Technology and Information Systems

TurkLang 2016 - Bishkek, Kyrgyzstan, 24/08/2016 - 26/08/2016

Kyrgyz State Technical University after I. Razzakov

TurkLang 2017 - Kazan, Tatarstan, Russia, 17/09/2015 -19/09/2015

Tatarstan Academy of Sciences
Research Institute of Applied Semiotics
Kazan Federal University
The Higher Institute for Information Technology and Information Systems
Institute of Computational Mathematics and Information Technologies

Introduction

Tatars dispersed people

About 2 mln of Tatars live in Tatarstan (30%)
About 5.3 mln – in Russia
About 1 mln - abroad

Introduction

- The supporting the Tatar language in computer technologies has begun for over 25 years ago
- In 1994, the state program on “Preservation, study and development of languages of the peoples of the Republic of Tatarstan” was adopted
- Today its forth version is in active progress (2014-2020: 1 mlrd. rub.)

Introduction

The Institute of Applied Semiotics of TAS (2009) (Joint scientific Lab of AI KSU and TAS, 1993-2009) ips.antat.ru

Research and development are carried out in two main directions:

- **Tatar localization of the information and communication technologies ("Tatar language within IT")**
- **The development and adaptation of information technologies for Tatar language (Linguistic resources and Soft tools)**

Tatar localization

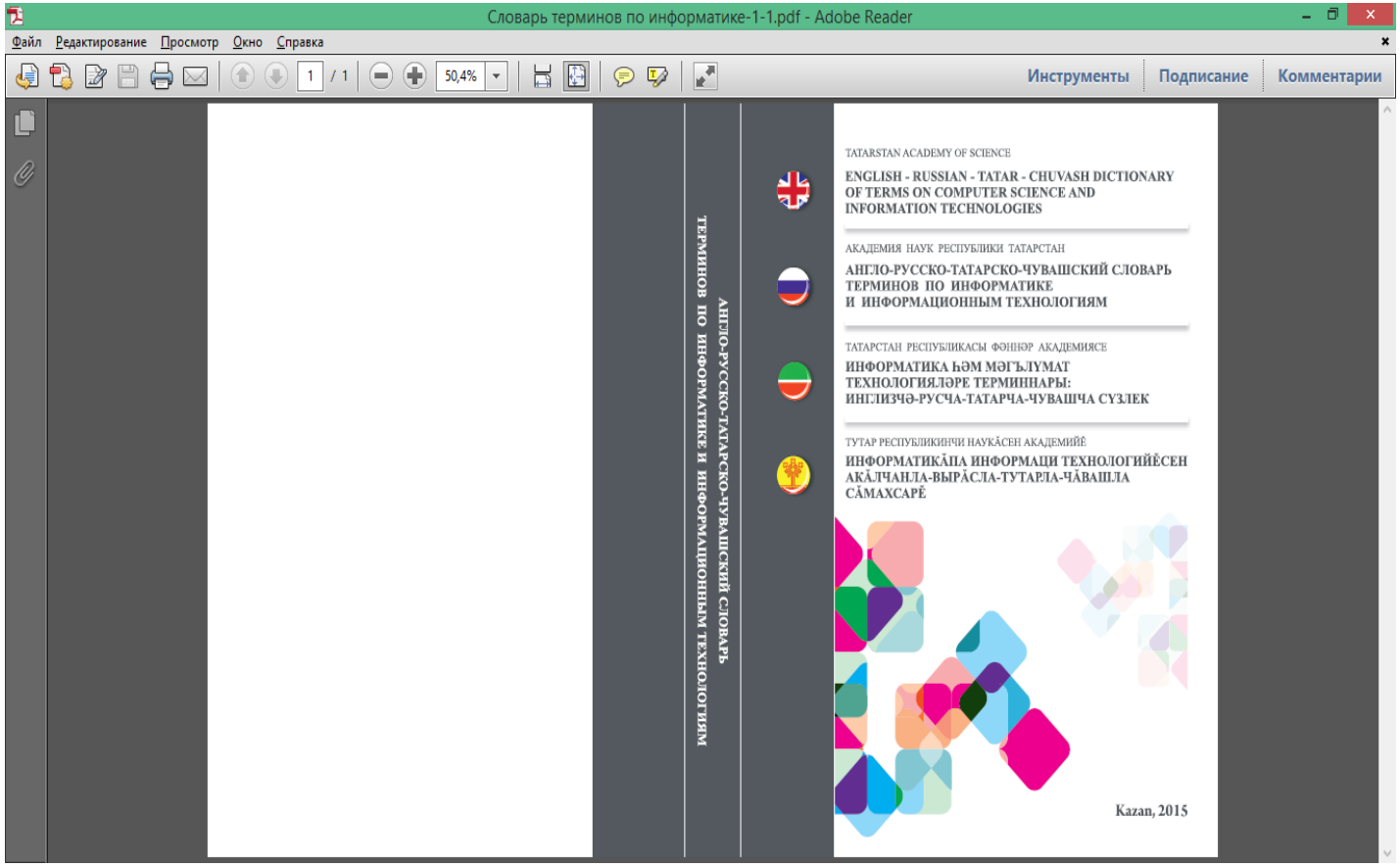
- **Tatar localization includes a providing next possibilities:**
  - to read, edit and print Tatar texts
  - to accumulate and exchange of information in the Tatar language
  - to communicate with the computer systems in the Tatar language (Man-machine Interface)

- **The most important aspects:**
  - the development of standards for the use of the Tatar language in cyberspace
  - developing and implementing a system of Tatar terms and concepts in Computer science and Information technology

- **The Resolution of the Cabinet of Ministers of the Republic of Tatarstan** "On the standards for encoding of the Tatar alphabet characters for computer applications" *№ 1026, December 9, 1996*.

Tatar localization

The explanatory dictionary of Tatar Terms and Concepts

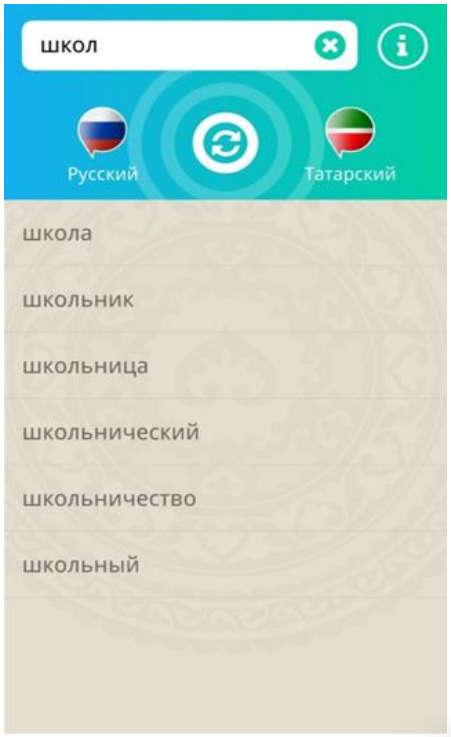


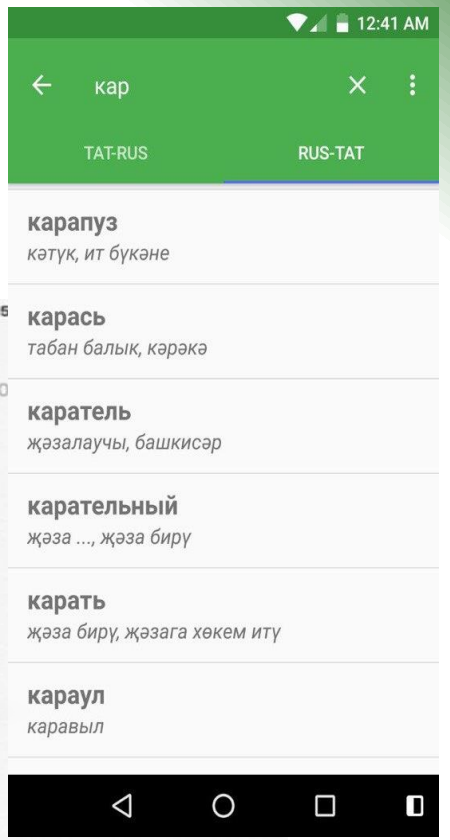
Mobile Systems

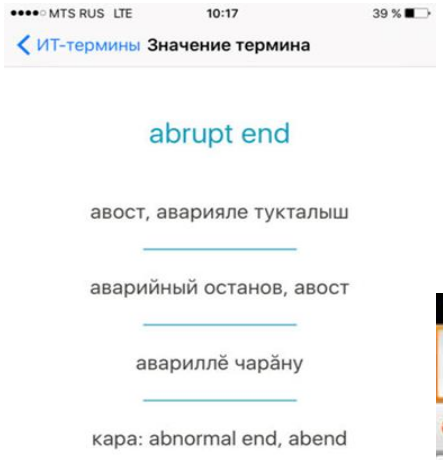
- Tatar keyboard for iOS
  - ✓ 12,000 downloading
  - ✓ 3 different keyboard layout
- A refreshed dictionary of used vocabulary
- Russian-Tatar-Russian dictionary for Android
- Site with demo for speech products
- Dictionary of IT terms for iOS
  - ✓ 2,700 downloading
  - ✓ description of English terms in Russian, Tatar and Chuvash languages
  - ✓ more 5,500 terms

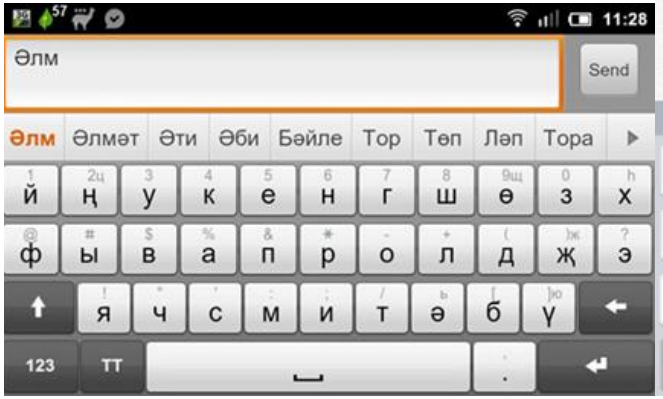
Tatar localization

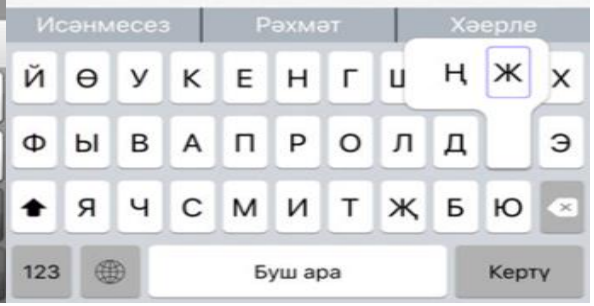
TatDict (Ru)
more 60,000 words
Predictive text input
Modern mobile systems
(iOS 7, Android 4+,
WR 8.0 - 8.1)
Appstore, Google Play,
Windows Phone Marketplace

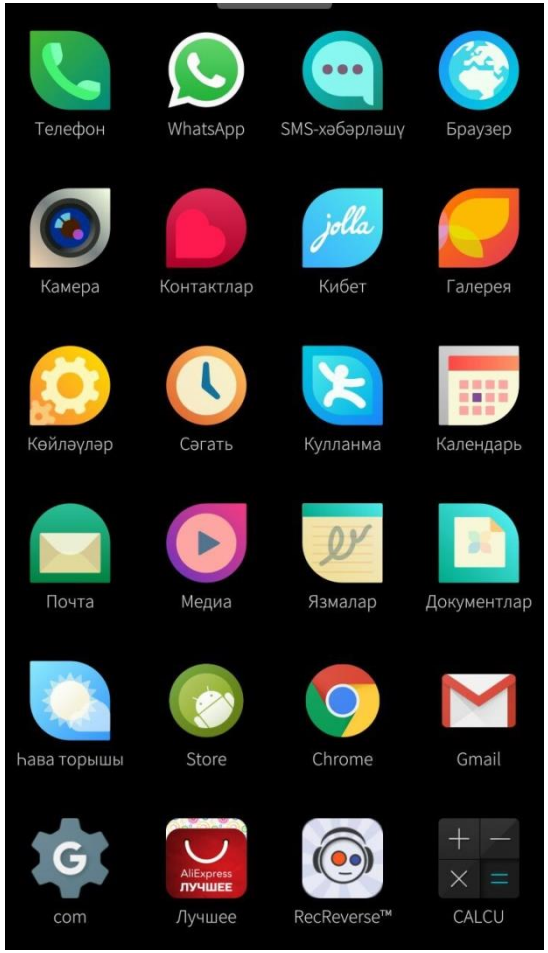


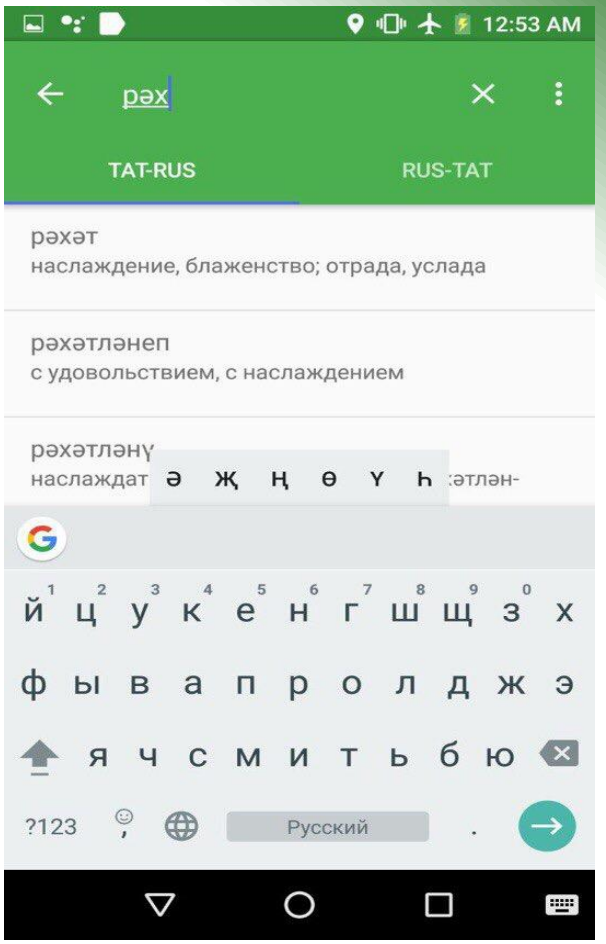






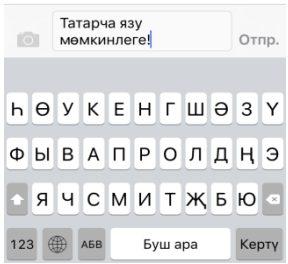


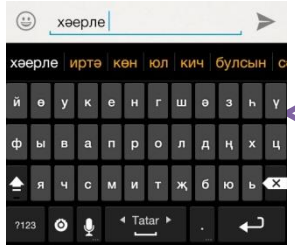




Download on the App Store

более 10 000 скачиваний





ANDROID APP ON Google play

более 40 000 скачиваний

Tatar localization

1. Tatar localization of the **Sailfish OS** http://mic.tatarstan.ru/rus/index.htm/news/948912.htm

Jointly with the company ***"Open mobile platform"***

2. Maintenance of Tatar version of the mobile application **“Portal of State and Municipal Services of the Tatarstan Republic”** https://uslugi.tatarstan.ru/start/mobile-apps

Jointly with the ***Ministry of Informatization and Communication of the TR***

3. Tatar keyboard layout for **macOS**

http://tatsoft.tatar/ru/portfolio-item/tatarskaya-klaviatura-dlya-macos/

4. Maintenance of the **portal of localized versions** of software products http://tatsoft.tatar

5. Tatar keyboard layout (Latin) for **OS Android**

https://play.google.com/store/apps/details?id=ru.antat.latin_tatar_keyboard&hl=ru

Tatar localization

Microsoft OS Windows XP, Windows Vista,
Windows7, Windows 8, Windows10
and Office applications

Developed and translated:
Terminology: 5000 new Terms and Concepts
Interface and Help-files: ~ 800 000 wordforms

Linguistic resources and Soft tools

- **Tatar National Corpus "Tugan Tel" (TNC)** http://tugantel.tatar/

The total volume of the annotated corpus **- 128 million word forms**.

TNC includes also **a search platform, a software package for the linguistic statistical study** of the Language Corpus and the DataBase.

- **Socio-Political Subcorpus** http://tugantel.tatar/corpus/op/

The volume of the subcorpus - more than **10 million words forms.**

- **Russian-Tatar parallel text collection in socio-political domain**

http://tugantel.tatar/corpus/op/parallel/

The volume of the subcorpus - more than **1 million words forms.**

- **Collection of Tatar scientific texts**

The volume of the subcorpus - more than **6 million words forms.**

Linguistic resources and Soft tools

- **The corpus of Russian-Tatar parallel texts**

  The volume of the Corpus - more than **45 million word forms.**

- **Corpus of aligned Russian-Tatar texts**

  The volume of the Corpus - more than **30 million word forms.**

- **Bilingual (Russian-Tatar) specialized Dictionaries** *of surnames, names, patronymics, states, subjects of the Russian Federation, regions of the Tatarstan Republic, populated areas, citizenship, nationalities*

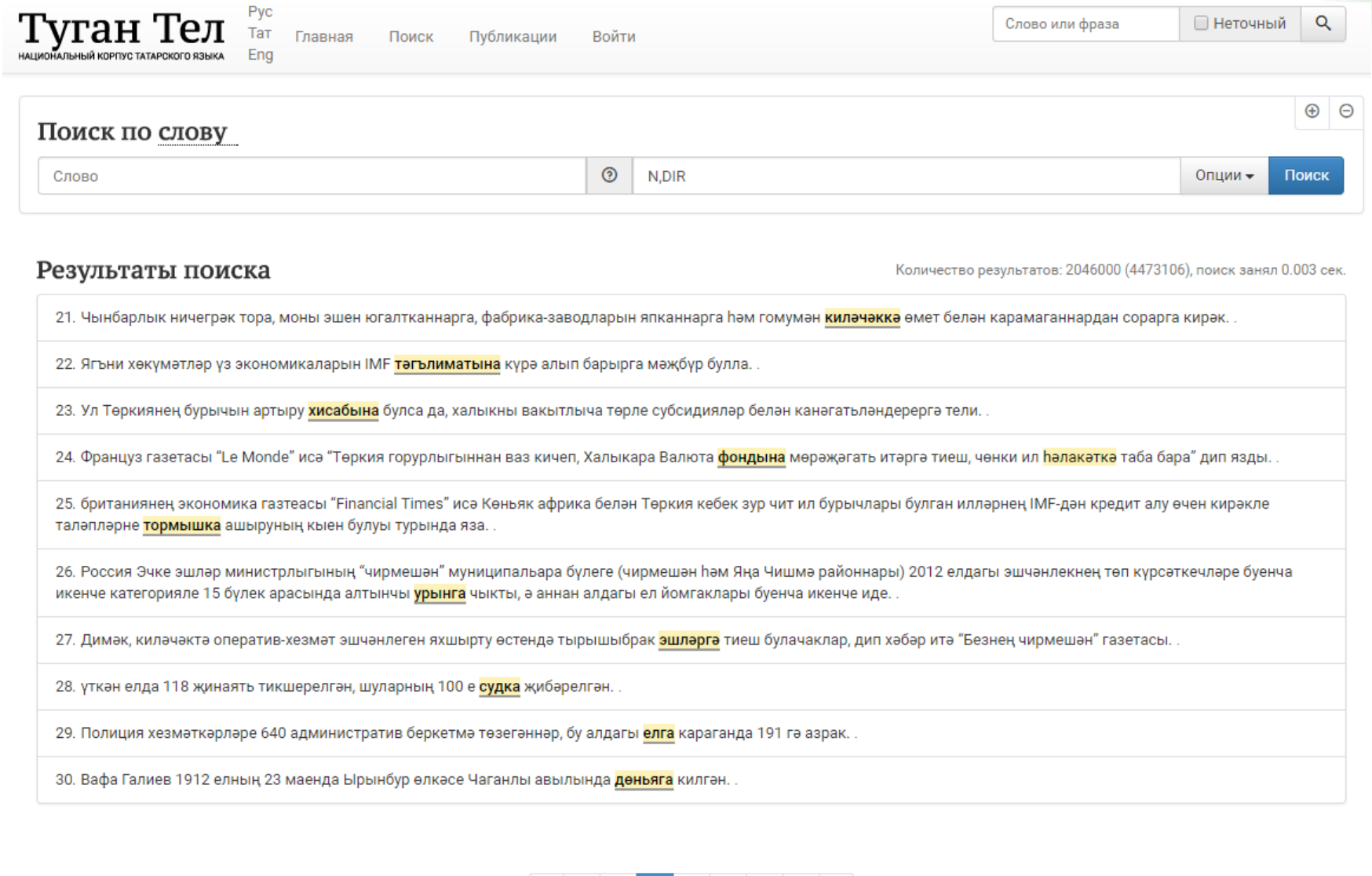
  The total volume of the Dictionaries - more than **70,000 vocabulary pairs.**

- **Bilingual (Russian-Tatar) dictionaries by parts of speech**

  The total volume of the Dictionaries - more than **25,000 vocabulary pairs.**

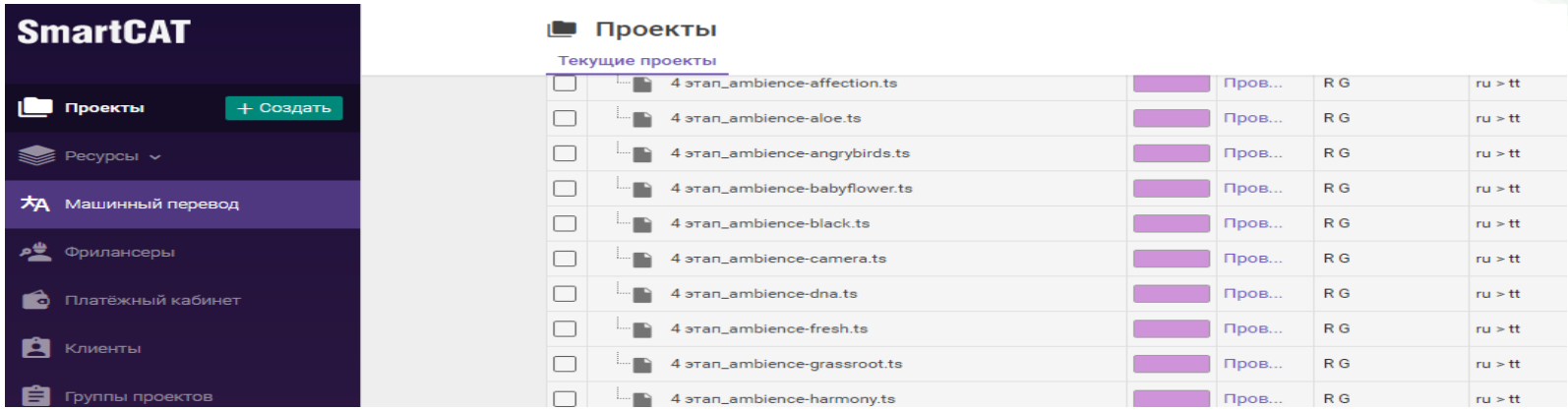
Linguistic resources and Soft tools

http://tugantel.tatar

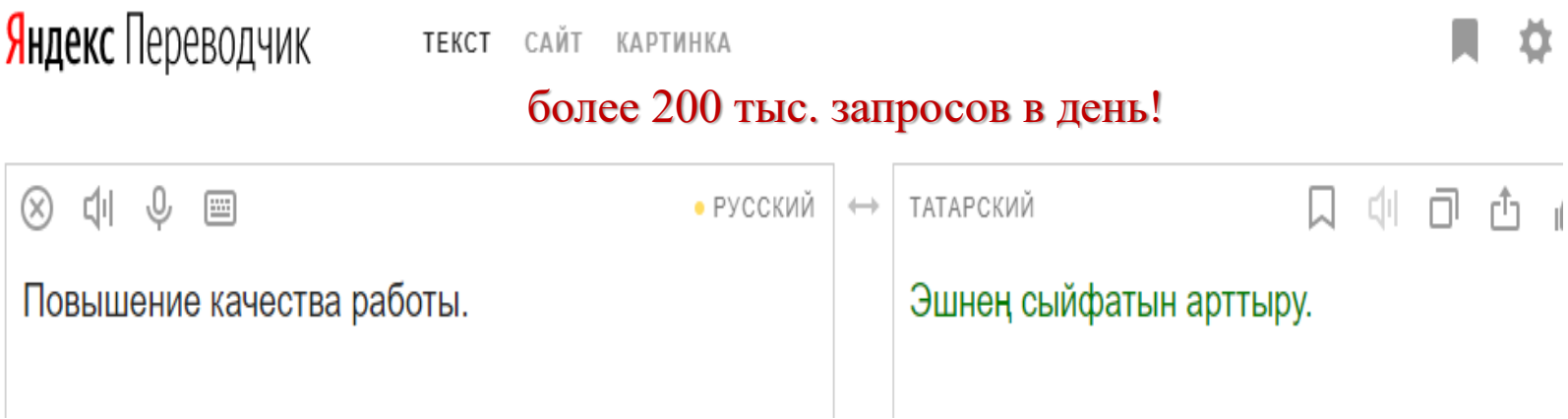


Tatar localization

The support system of professional translation SmartCAT https://www.smartcat.ai/



Open Internet service with the Russian-Tatar machine translation system: https://translate.yandex.ru/



Cooperation

- Cooperation with **Yandex Company** on the development of the system of Tatar-Russian machine translation: https://translate.yandex.ru/
- The following resources are prepared and transferred to Yandex:
  - ✓ **Russian-Tatar parallel corpus - the volume of 17,6 million word forms**
  - ✓ **Tatar texts - the volume of 40 million word forms**
  - ✓ **Module of the Tatar morphological analysis**
- Cooperation with the **ABBYY LS Company** on
  - ✓ **maintenance of the system for professional translation SmartCAT**
  - ✓ **development of the Linguistic resources for MT**

Linguistic resources and Soft tools

FineReader

FR 4.0 Tatar …

Software for the Tatar language

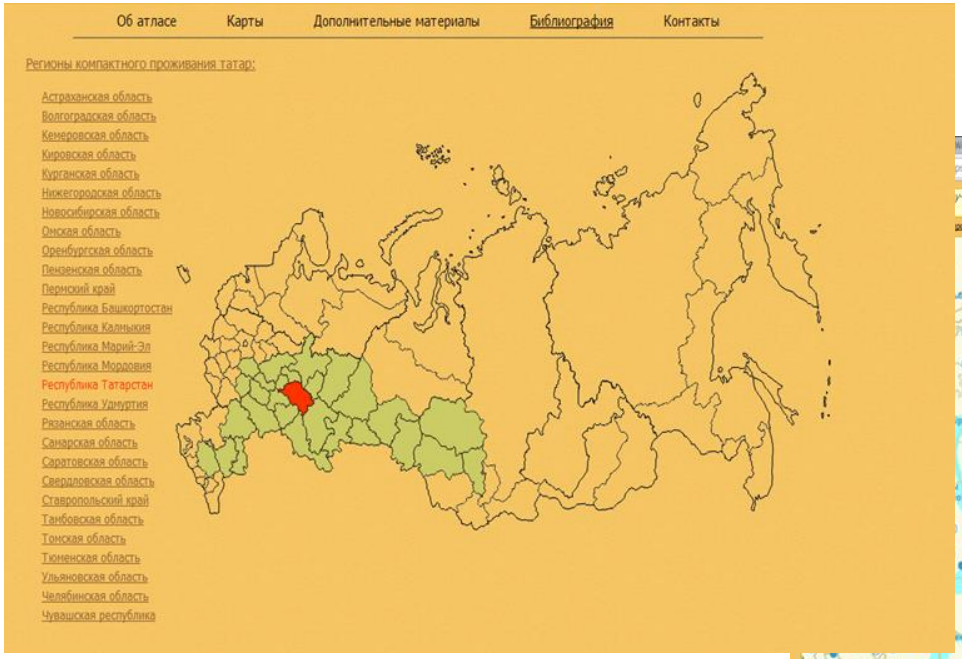
Lingvo 12 Turkish …

Lingvo x3 Tatar …

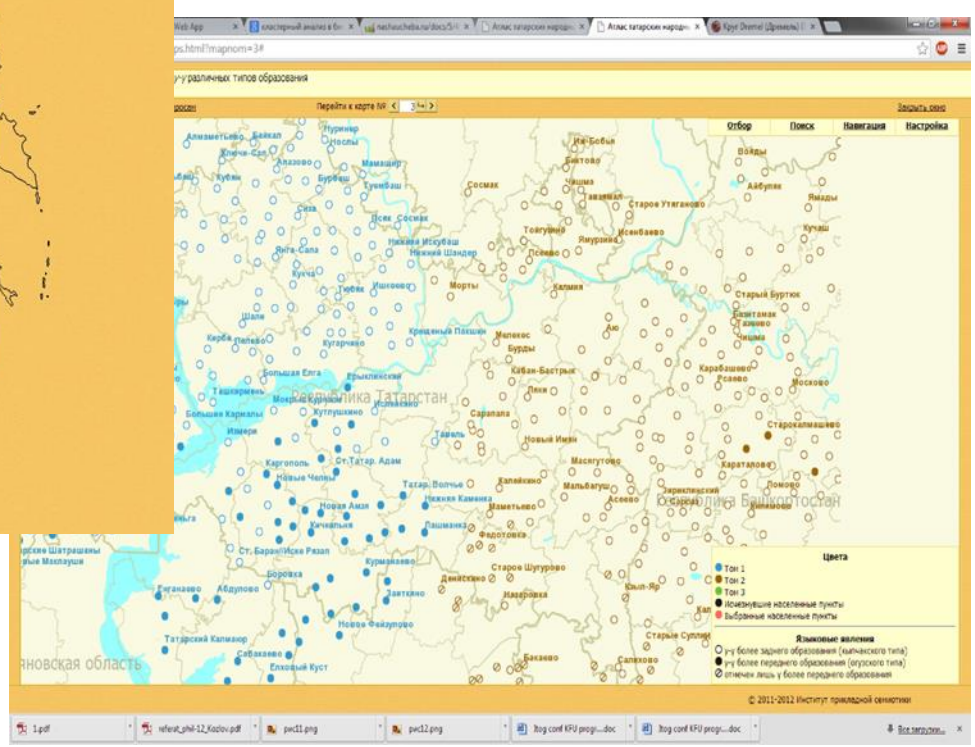
- more **20 languages**
- more **200 dictionaries**

Linguistic resources and Soft tools

The electronic version of the Atlas of the Tatar dialects of the Volga, Ural regions and Siberia (joint project: IPSAN, IALI and KFU) /atlas.antat.ru.



Features of Tatar dialects on phonetics, morphology, vocabulary and syntax



215 language phenomena in 1047 settlements

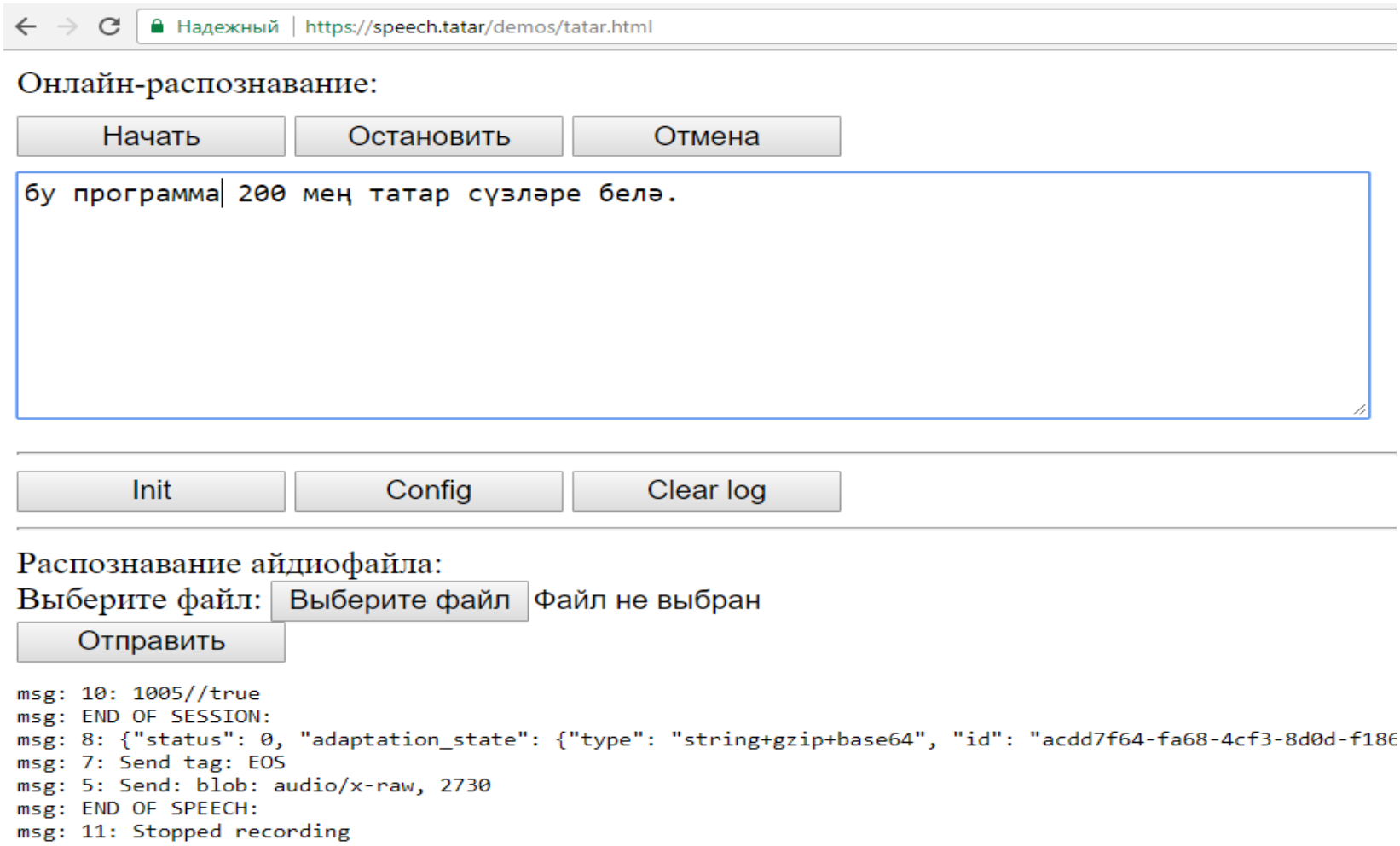
Speech technologies

Speech Synthesizer

Speech technologies

Speech Recognizer

Web version of the Tatar Speech Recognizer (based on deep neural networks). http://speech.tatar/demos/tatar.html.

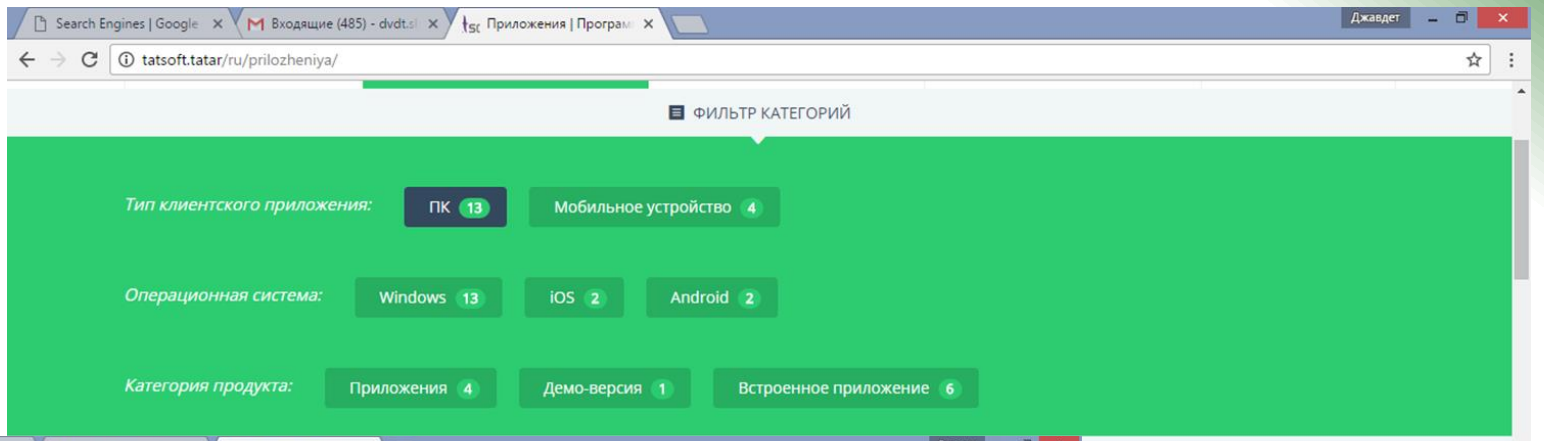


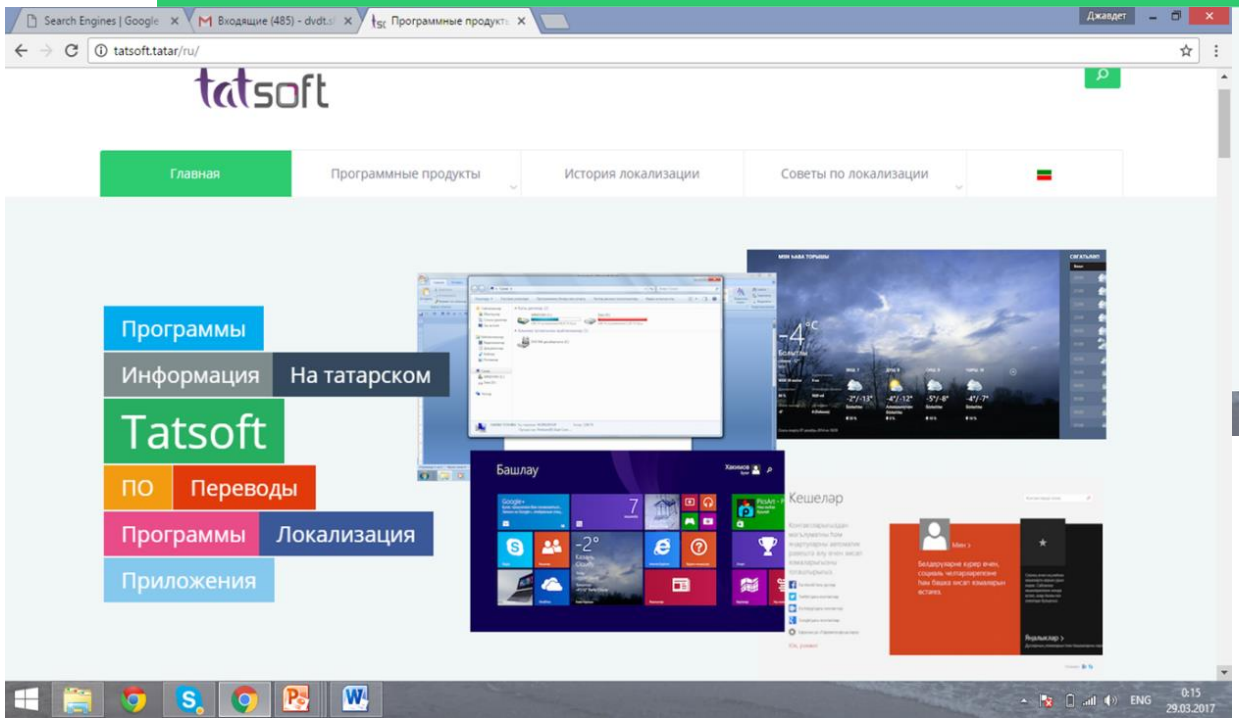
Speech technologies

Speech corpus	
#speakers	**441**
Duration	**78:51:52**
*Spontaneous speech**	*5:19:33*

*Cooperation with the **Youth Fond "Selet" of Tatarstan** on the development of the speech base for Tatar speech synthesizer and recognizer*

Tatar localization





Tatar Morphological Analyzer

Technical parameters	PC-KIMMO (1997)	HFST(2014)
Morphotactical Rules	49	49
Phonological Rules	42	55
Vocabulary	26500	28000
Processing Speed	~500-1000 (words per second)	~5000-20000 (words per second)
Coverage (Recall)	80-85%	93-95%
Alphabet	Latin	Cyrillic
Programming Language	C/C++ and Delphi	C/C++ and Python
Has API	No	Yes
Interface as:	Desktop Application	Web Application

- Completeness of the system – **93-95%**
- Used in «Tugan Tel» Tatar National corpus annotation
- Used in Yandex Russian-Tatar machine translation system
- Demo is available: tatmorphan.pythonanywhere.com

Multifunctional model of Turkic Morphems

Language independent part

Relational-situational model-

RSM

Language dependent part

Linguistic aspects of the model

1. Identificational aspect
2. Morphonological aspect
3. Morphological aspect
4. Syntactical aspect
5. Semantic aspect

The model of language morpheme L1 (Tatar) - MML1:	The model of language morpheme L2 (Kazakh) – MML2:	The model of language morpheme LN (Kyrgyz) - MMLN:
The model of root morpheme.	The model of root morpheme.	The model of root morpheme.
The model of affixal morpheme.	The model of affixal morpheme.	The model of affixal morpheme.

Conclusion

- **A software toolkit for the Tatar language has been developed. The Tatar SoftToolkit provides:**

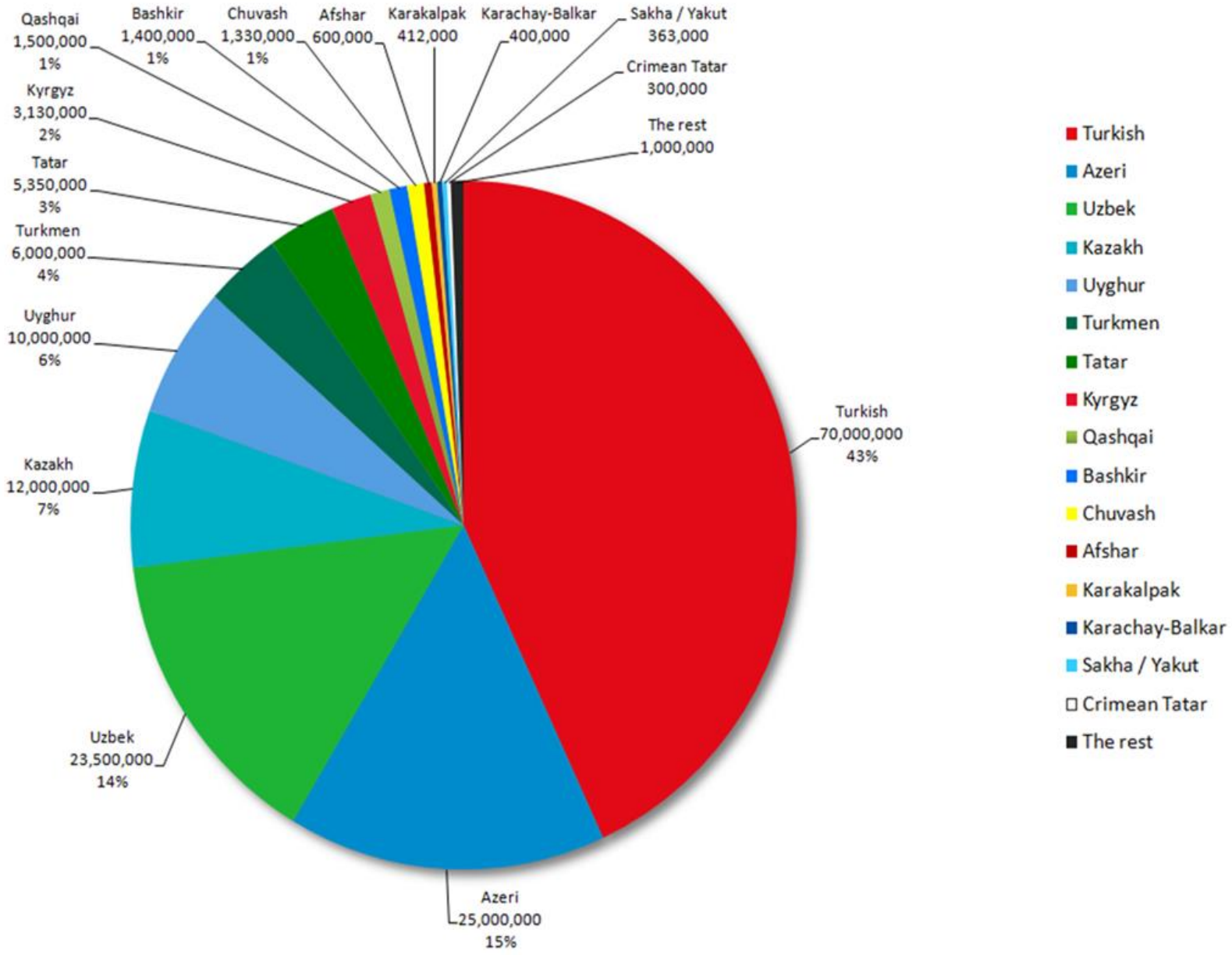
- accumulation, storing up, processing, receiving and sending of information in the Tatar language

- for ordinary users - communication in social networks and solving problems in the Internet space; using mobile devices, computers in the Tatar-speaking environment

- for scientists and specialists - the use of information technology for the pre servation, scientific study and development of the Tatar language

- teaching the Tatar language and teaching subjects in the Tatar language

Number of Native Speakers in the Turkic Language Family



Conclusion: Integration is necessary

- **Development of the general TurkLang portal**

- access to common linguistic resources - dictionaries, thesauruses, ontologies, electronic Corpuses, multilingual electronic databases
- exchange of software tools, technologies and applied programs
- exchange of information on research, development and implementation of computer processing of Turkic languages
- exchange of publications - scientific articles, collections of proceedings, monographs
- development and implementation of joint software products (MT, search syst em, unified electronic Corpus of Turkic languages)

Autumn in Kazan
Enjoy your stay in Kazan