A prototype machine translation system for Tatar and Bashkir based on free/open-source components

Francis M. Tyers
Grup Transducens
Dept. Lleng. i Sist. Inform.
Universitat d'Alacant
ftyers@dlsi.ua.es

Jonathan North Washington
Depts. Linguistics &
Central Eurasian Studies
Indiana University
jonwashi@indiana.edu

Ilnar Salimzyanov
Department of Russian Philology
Kazan Federal University
ilnar.salimzyan@gmail.com

Rustam Batalov
Department of Economics
Bashkir State University
taqmaq@mail.ru

Introduction

Layout of the talk

- Introduction
- Description of architecture and development process
- Evaluation
- Discussion and future work

Introduction

Why make a Tatar ↔ Bashkir MT system?

- There are quite a few examples of MT systems made with Apertium
  - but no working examples between “morphologically complex” languages
- Course in Šupaškar (Čeboksary), January 2012
  - Nearly all the languages of Russia are “morphologically complex”, but we had no example system to present.
- The idea was to create a pedagogical example
  - A simple demonstration system
  - that nonetheless showed all modules working
- We chose Tatar and Bashkir because,
  - Native speakers available to help out
  - Good results could be accomplished with little work

Tatar and Bashkir
Geography

Tatar and Bashkir
Demographics

Tatar

- spoken by $>$ 6.5m people (**?**)
- coofficial with Russian in Tatarstan
- minority language even in Tatarstan
- high rate of bilingualism with Russian

Bashkir

- spoken by $>$ 1.3m people (**?**)
- coofficial with Russian in Bashqortostan
- minority language even in Bashqortostan
- high rate of bilingualism with Russian
- classified as “vulnerable” by UNESCO

Tatar and Bashkir
Comparison

The similarities

- very close relatives: same branch of Kypchak group of Turkic
- share many innovations
- high level of mutual intelligibility when spoken
- large percentage of the lexicon are similar

The differences

- Bashkir has quite a few phonological innovations not found Tatar
  - handful of extra historical changes
  - rounding harmony
  - desonorisation of high-sonority suffix-initial consonants
    cf. **?**
- a number of morphological differences
  - e.g., different volitional participles
- many inherent similarities are obscured:
  - phonological and morphological differences
  - different orthographical systems

Tatar and Bashkir
Comparison

A sentence

tat Һава бүген бик әйбәт, җылы гына.
[hɑwɑ bəʏən bijk æjbæt, ʑ́ələ ʁənɑ]

bak Һауа бөгөн бик әйбәт, йылы ғына.
[hɑwɑ bəʏən bik æjbæt, jələ ʁənɑ]

Volitional participles

tat Барасым килә. '*I want to go*'

bak Барғым килә. '*I want to go*'
Бараһым килә.

Phonology: desonorisation

tat башны '*head-ACC*'

bak башты '*head-ACC*'

cf. **?**

Tatar and Bashkir
Comparison

(transcriptions are approximate)

Tatar and Bashkir

tat Һава бүген бик әйбәт, җылы гына.
[hɑwɑ bəʏən bijk æjbæt, ʑələ ʁənɑ]

bak Һауа бөгөн бик әйбәт, йылы ғына.
[hɑwɑ bəʏən bik æjbæt, jələ ʁənɑ]

Comparison with other Turkic languages

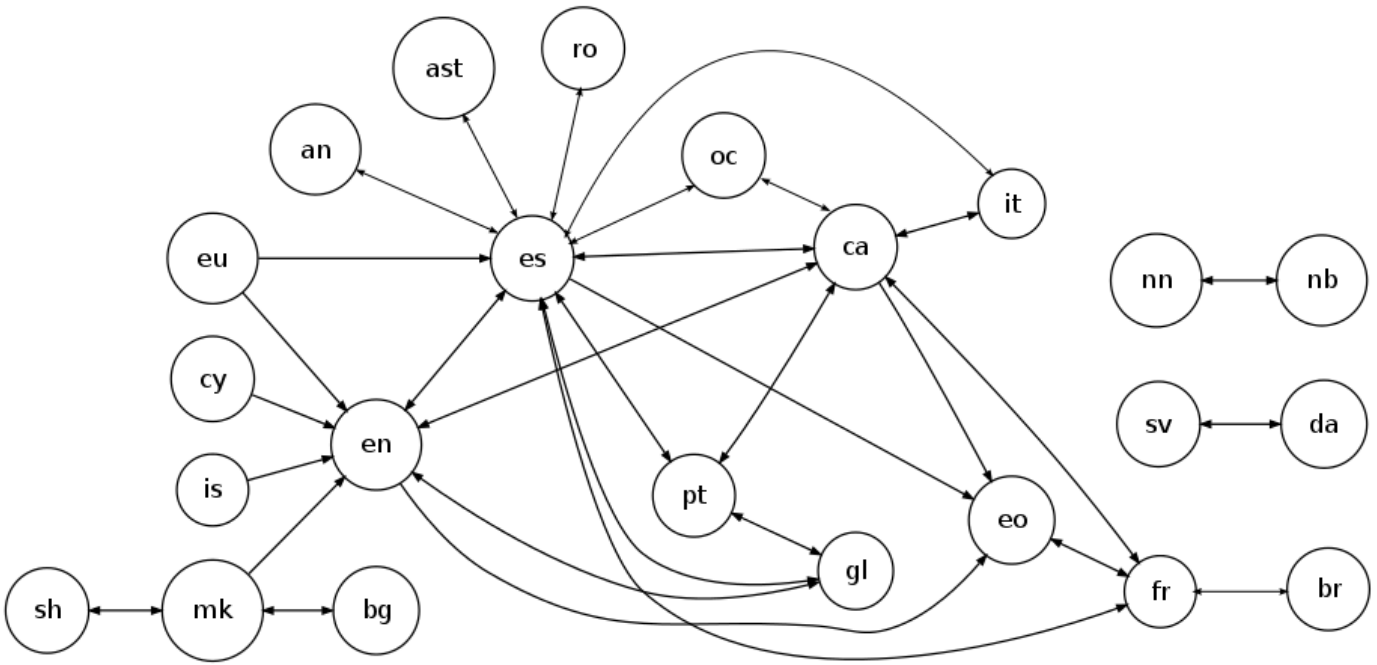
tur Hava bugün çok güzel, yeterince sıcak.
[hɑvɑ buʏyn ʧok gyzæl, jɛtʰɛrɪnʤɛ sɯʤɑkʰ]

chv Çанталăк паян питĕ хитре, самай ăшă.
[ɕanˈtalək pɑˈjɑn ˈpitə xitˈre, saˈmaj ˈɔʂə]

kaz Ауа райы бүгін әбден жақсы, жылы.
[ɑwɑ ɾɑjə bʉʏən æbdi̯ən ʐˌɑχsə, ʐˌələ]

kir Аба ырайы бүгүн аябай жакшы, жылуу.
[ɑβɑ ɾɑjɯ βyʏyn ɑjɑβɑj ʤɑqʃɯ, ʤɯluː]

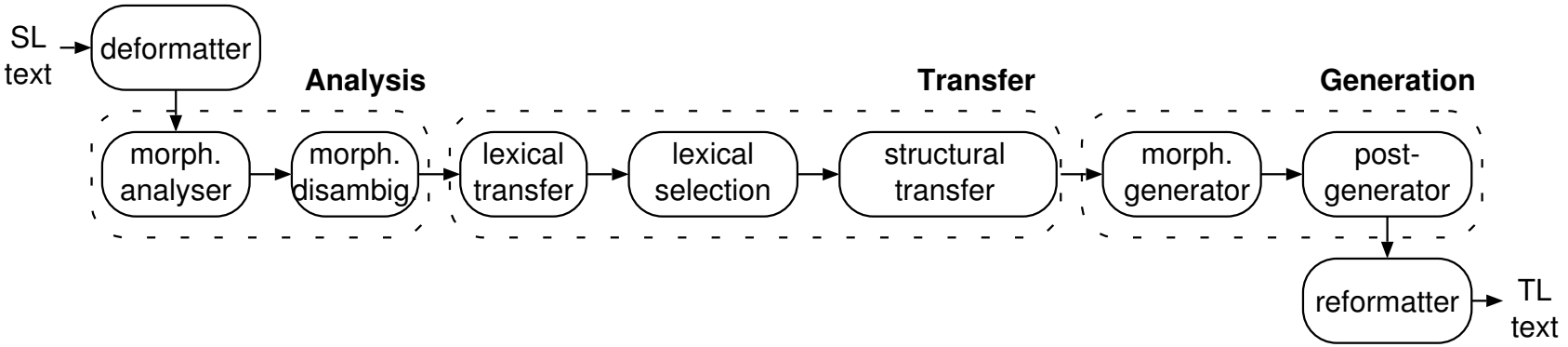
Apertium

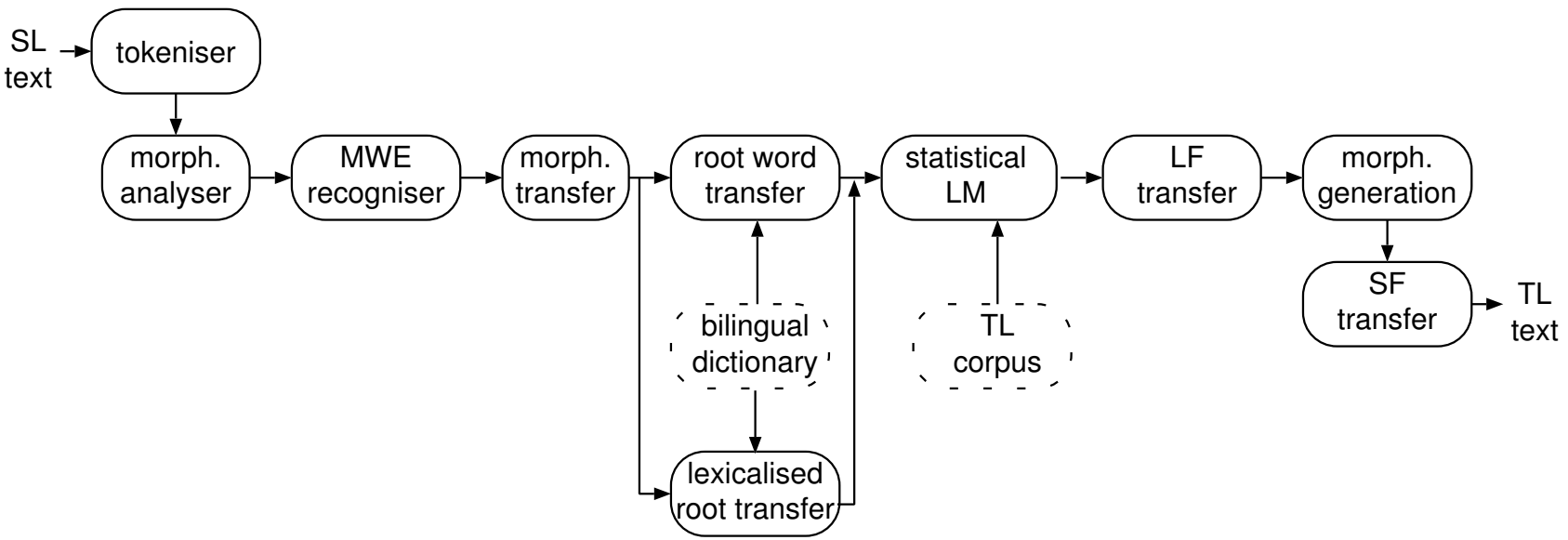


Overview

- Free/open-source software (GPL licence)
  - Not only the programs, but the data too!
- Fairly mature data for over 30 language pairs
- ... but so far no mature data for Turkic languages!

Architecture





Machine translation between Turkic languages (Tantuğ et al., 2007)

Tools

Helsinki Finite-State Toolkit (HFST)

- Free/open-source implementation of the Xerox finite-state formalisms `lexc`/`twol`
- Used for morphological analysis and generation

`http://hfst.sf.net/`

VISL Constraint Grammar

- Rule-based morphological disambiguation

`http://beta.visl.sdu.dk/constraint_grammar.html`

Apertium

- Lexical and structural transfer
- “Platform”: Build system, ancilliary tools, etc.

`http://www.apertium.org/`

Һава бүген бик әйбәт, җылы гына.

Һава бүген бик әйбәт, җылы гына.

Morphological analysis

^Һава/һава<n><attr>/һава<n><nom>$ ^бүген/бүген<adv>$
^бик/бик<adv>/бик<n><attr>/бик<n><nom>$
^әйбәт/әйбәт<adj>/әйбәт<adj><subst><nom>$^,/,<cm>$
^җылы/җылы<n><attr>/җылы<n><nom>/җылы<adj>/җылы<adj><subst><nom>$
^гына/гына<postadv>$^./.<sent>$

Һава бүген бик әйбәт, җылы гына.

Morphological analysis

^Һава/һава<n><attr>/һава<n><nom>$ ^бүген/бүген<adv>$
^бик/бик<adv>/бик<n><attr>/бик<n><nom>$
^әйбәт/әйбәт<adj>/әйбәт<adj><subst><nom>$^,/,<cm>$
^җылы/җылы<n><attr>/җылы<n><nom>/җылы<adj>/җылы<adj><subst><nom>$
^гына/гына<postadv>$^./.<sent>$

Morphological disambiguation

^Һава<n><nom>$ ^бүген<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$
^җылы<adj>$ ^гына<postadv>$^.<sent>$

ħава бүген бик әйбәт, җылы гына.

Morphological analysis

^ħава/ħава<n><attr>/ħава<n><nom>$ ^бүген/бүген<adv>$
^бик/бик<adv>/бик<n><attr>/бик<n><nom>$
^әйбәт/әйбәт<adj>/әйбәт<adj><subst><nom>$^,/,<cm>$
^җылы/җылы<n><attr>/җылы<n><nom>/җылы<adj>/җылы<adj><subst><nom>$
^гына/гына<postadv>$^./.<sent>$

Morphological disambiguation

^ħава<n><nom>$ ^бүген<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$
^җылы<adj>$ ^гына<postadv>$^.<sent>$

Lexical transfer and selection

^ħава<n><nom>/ħaya<n><nom>$ ^бүген<adv>/бөгөн<adv>$ ^бик<adv>/бик<adv>$
^әйбәт<adj>/әйбәт<adj>$^,<cm>/,<cm>$ ^җылы<adj>/йылы<adj>$
^гына<postadv>/ғына<postadv>$^.<sent>/.<sent>$

Һава бүген бик әйбәт, җылы гына.

Morphological analysis

^Һава/Һава<n><attr>/Һава<n><nom>$ ^бүген/бүген<adv>$
^бик/бик<adv>/бик<n><attr>/бик<n><nom>$
^әйбәт/әйбәт<adj>/әйбәт<adj><subst><nom>$^,/,<cm>$
^җылы/җылы<n><attr>/җылы<n><nom>/җылы<adj>/җылы<adj><subst><nom>$
^гына/гына<postadv>$^./.<sent>$

Morphological disambiguation

^Һава<n><nom>$ ^бүген<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$
^җылы<adj>$ ^гына<postadv>$^.<sent>$

Lexical transfer and selection

^Һава<n><nom>/Һауа<n><nom>$ ^бүген<adv>/бөгөн<adv>$ ^бик<adv>/бик<adv>$
^әйбәт<adj>/әйбәт<adj>$^,<cm>/,<cm>$ ^җылы<adj>/йылы<adj>$
^гына<postadv>/ғына<postadv>$^.<sent>/.<sent>$

Structural transfer

^Һауа<n><nom>$ ^бөгөн<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$
^йылы<adj>$ ^ғына<postadv>$^.<sent>$

ҥава бүген бик әйбәт, җылы гына.

Morphological analysis

^ҥава/ҥава<n><attr>/ҥава<n><nom>$ ^бүген/бүген<adv>$
^бик/бик<adv>/бик<n><attr>/бик<n><nom>$
^әйбәт/әйбәт<adj>/әйбәт<adj><subst><nom>$^,/,<cm>$
^җылы/җылы<n><attr>/җылы<n><nom>/җылы<adj>/җылы<adj><subst><nom>$
^гына/гына<postadv>$^./.<sent>$

Morphological disambiguation

^ҥава<n><nom>$ ^бүген<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$
^җылы<adj>$ ^гына<postadv>$^.<sent>$

Lexical transfer and selection

^ҥава<n><nom>/ҥаya<n><nom>$ ^бүген<adv>/бөгөн<adv>$ ^бик<adv>/бик<adv>$
^әйбәт<adj>/әйбәт<adj>$^,<cm>/,<cm>$ ^җылы<adj>/йылы<adj>$
^гына<postadv>/ғына<postadv>$^.<sent>/.<sent>$

Structural transfer

^ҥаya<n><nom>$ ^бөгөн<adv>$ ^бик<adv>$ ^әйбәт<adj>$^,<cm>$
^йылы<adj>$ ^ғына<postadv>$^.<sent>$

Morphological generation

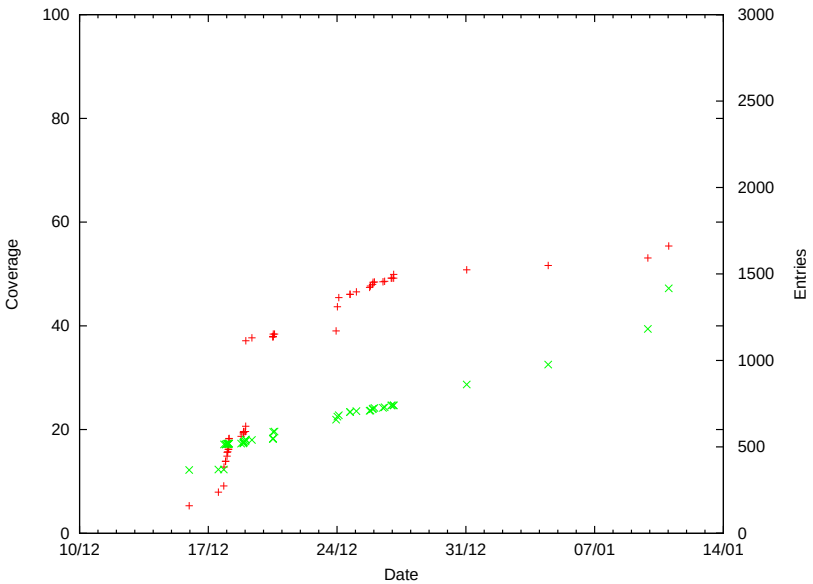
ҥаya бөгөн бик әйбәт, йылы ғына.

Development process

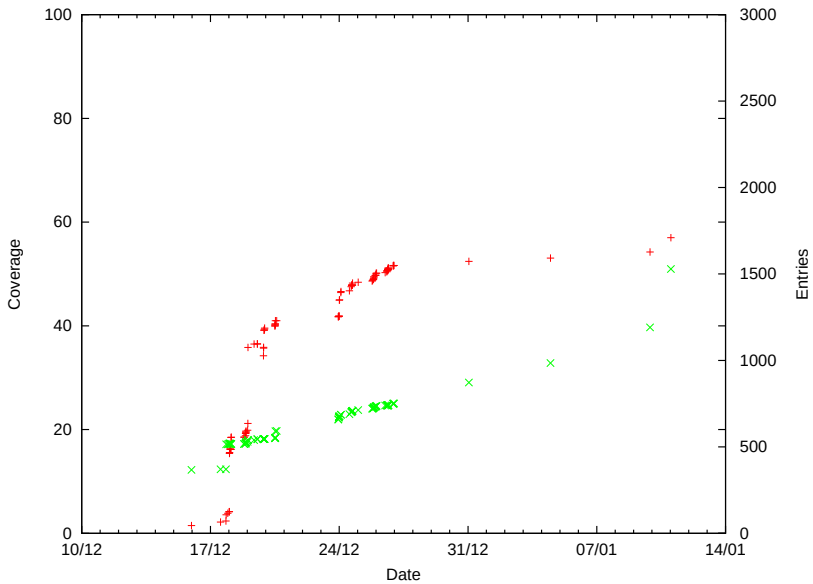
- Stems in both Tatar and Bashkir are added to an online spreadsheet
  - Added according to frequency in the RNC.
  - At the same time as translations, POS categories are added, and some relevant subcategorisations (transitivity, Tatar infinitive)
- Scripts are used to automate the process of converting the spreadsheet to bilingual (`.dix`) and monolingual (`.lexc`) formats.
- At the same time, the morphotactics and phonology are written according to available grammatical descriptions

Development speed

Coverage and number of stems / time



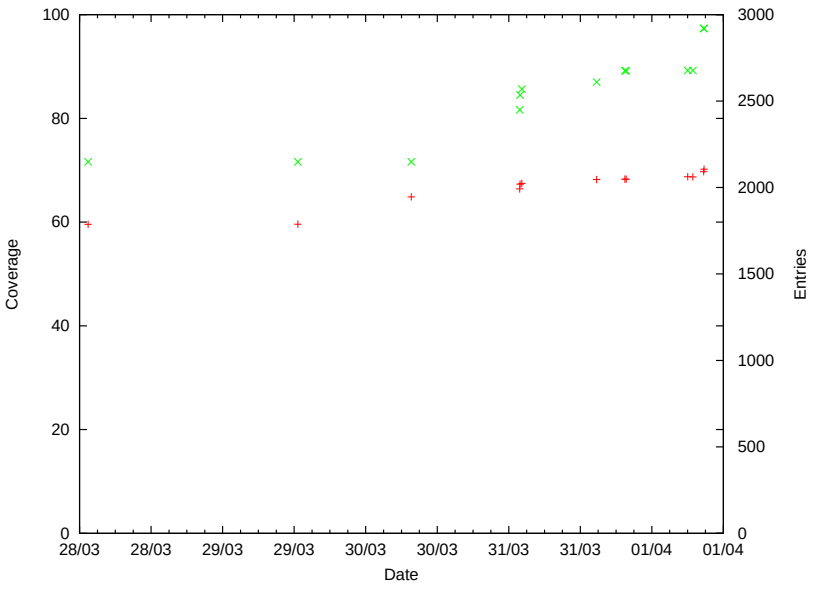
Tatar



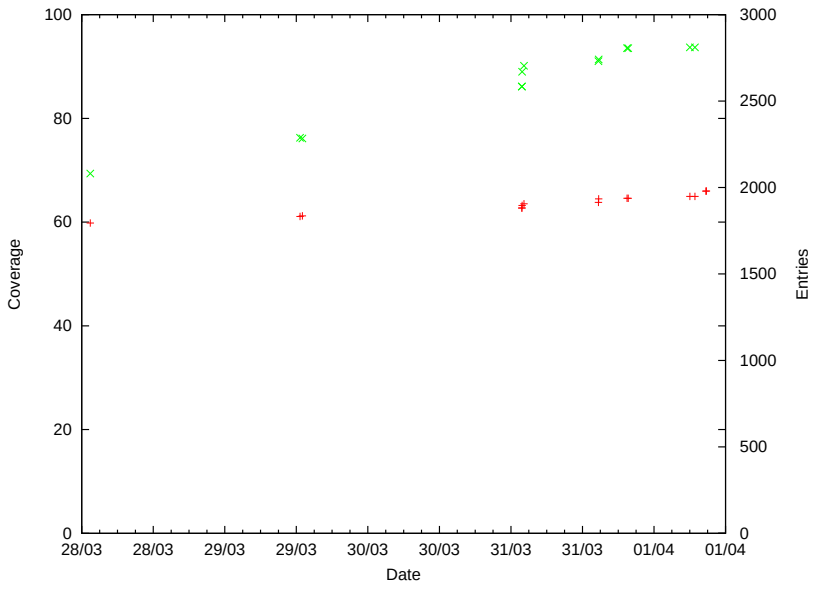
Bashkir

Development speed

Coverage and number of stems / time



Tatar



Bashkir

Evaluation

Rules and lexica

Corpus	Entries
Bilingual dictionary	2,685
Disambiguation (`tt` and `ba`)	6
Transfer (`tt`→`ba`)	3
Transfer (`ba`→`tt`)	3

Coverage

Corpus	Tokens	Coverage
Tatar New Testament (NT)	163,603	72.04%
Tatar Wikipedia	37,123	70.19%
Bashkir Wikipedia	12,267	65.99%

Error rate

Corpus	Direction	Tokens	Unknown	WER
story	`tt`→`ba`	311	2	6.73%
	`ba`→`tt`	312	0	6.43%

Weak points

Linguistic: Majority of errors due to:

- mistakes and gaps in Tatar morphophonology (certain combinations)
- orthographical representations of phonology
  e.g., tat: сәгать /sæʁæt/, сәгате /sæʁætə/, сәгатьтә /sæʁættæ/
- coverage (not enough stems)

Technical

- Vowel harmony processing on clitics (e.g., *да*/*дә* ‘and’) after unknown words.
- Traditional methods of quality control are difficult to apply

Observations

Consistency

- For tagsets: If something is the same, tag it the same
- Otherwise you spend a lot of time on useless ‘transfer’ (e.g. ky, 50 transfer rules, around 4 ‘real’ rules)
- If you solve a problem well in one language, and it comes up in another one, solve it the same way (e.g. epenthesis, vowel harmony)

Parallel development

- Building the transducers and bilingual dictionary at the same time is good

Starting from scratch

- Much easier to work with new code than adapt existing code
  - Not to say that existing code can’t be useful as a model

Future work

For Tatar↔Bashkir:

- Expand lexicons
- Improve morphophonology

Other Turkic language projects:

- We're currently working on:
  - Turkish↔Chuvash
  - Tatar↔Kazakh[1]
  - Turkish↔Turkmen[1]
  - Turkish↔Tatar[1]
- We have worked on:
  - Turkish↔Azerbaijani[2]
  - Turkish↔Kyrgyz[2]

[1] As projects in the 2012 Google Summer of Code
[2] As projects in the 2011 Google Summer of Code (pending reworking of Turkish morphology)

Conclusion

- We've presented
  - a prototype MT system for translating between Tatar and Bashkir,
  - described the tools used to make it,
  - given a preliminary evaluation of the translation quality,
  - and given some observations about the development process

Try it out! `http://elx.dlsi.ua.es/~fran/tt-ba/`
Source on apertium svn.

Pəxmət

Räxmät

Paxmət

Teşekkürler