# DESIGNING FINITE-STATE MORPHOLOGICAL TRANSDUCERS FOR KYPCHAK LANGUAGES

**Jonathan North Washington**
Indiana University
jonwashi@indiana.edu

**Ilnar Salimzyanov**
Казан (Идел буе) федераль университеты
ilnar.salimzyan@gmail.com

**Francis M. Tyers**
UiT Norgga Árktalaš Universitehta
francis.tyers@uit.no

Special thanks to:
Tolgonay Kubatova
Aida Sundetova
Ağarahim Sultanmuradov

## Kypchak languages



- Turkic languages (SOV, agglutinative, vowel harmony)

| | **Kyrgyz** /qɯrɣɯz/ | **Kazakh** /qazaq/ | **Tatar** /tɤtɤr/ | **Kumyk** /qumuq/ |
|---|---|---|---|---|
| classification | Eastern | Southern | Northern | Western |
| **population of speakers** | | | | |
| number | 3M | 8M–12M | 5.4M | 430K |
| primary | Kyrgyzstan | Kazakhstan | Tatarstan | Dagestan |
| secondary | China, etc. | China, Mongolia | Bashqortostan | — |
| **external influences** | | | | |
| Mongolic | moderate | moderate | light | light |
| Oghuz | — | — | light | moderate |
| Persian | heavy | heavy | heavy | heavy |
| Russian | heavy | heavy | heavy | heavy |

## Morphological transducers

**........ Morphological transducers ..........**

- Efficient (in speed & size) models of a language's morphology
- Take a surface form, and produce valid lexical form(s)
- Take a lexical form, and produce valid surface form(s)
  алдым ↔ ал<v><tv><ifi><p1><sg>, алд<n><px1sg><nom>

**........ Transducers for Turkic languages ........**

- Turkish (Çöltekin, 2010 & 2014; Oflazer, 1994)
- Crimean Tatar (Altıntaş, 2001)
- Turkmen (Tantuğ et al., 2006)
- Kazakh (Бекманова & Махимов, 2013)
- our Kyrgyz, Kazakh, Tatar, Kumyk: all GPL (=free and open)!

**........ Framework: HFST ..................**

- Reimplements Xerox FST formalisms (lexc & twol)
- Also provides a wrapper around popular free/open-source FST toolkits: SFST, OpenFST, and Foma

**........ Approach ....................**

- morphotactics implemented in lexc
- morphophonology implemented in twol
- compiled separately; compose-intersected to single transducer
  алдым ↔ ал>{D}{I}>м ↔ ал<v><tv><ifi><p1><sg>
  алдым ↔ алд>{I}м ↔ алд<n><px1sg><nom>

## Further information

- Part of **Apertium Turkic** project:
  http://wiki.apertium.org/wiki/Apertium_Turkic
- Transducers available **live** at apertium.org
- **Source code** available from Apertium's svn repo
- Turkic RBMT **mailing list** (>25 subscribers):
  apertium-turkic@lists.sourceforge.net
  Feel free to post in any language!
- See our papers in LREC proceedings
  (2012: Kyrgyz, 2014: Kazakh, Tatar, Kumyk)
- And feel free to contact the authors any time!

## Example output

**............... Gloss ..................**

(1) Кудай   Өзу      жаратканынын   баарына     карап,    өтө  жакшы экенин көрду.
    Кудай   Өзінің   жаратқандарының бәріне     қарап,    өте  жақсы екенін көрді.
    Аллаһ   Үзе      яраткан        нәрсәләргә  карап,    аларның бик яхшы иканен күрде.
    Аллагъ  Оьзю     яратгъан       затлагъа   къарап,   олар  бек яхшы экенин гёрген.
    God     own-his  created        [everything/thing-s]-to looked.at, they/their very good being saw.

*'God looked at everything he had created and saw that it was very good.'* (Bible, Genesis 1:31)

**............... Output ..................**

**Kyrgyz (kir)**
Кудай Өзу жаратканынын баарына карап, өтө жакшы экенин көрду.

```
Кудай<n><nom>
Өз<n><ref><px3sp><nom>
жарат<v><tv><ger_past><p3><gen>
баары<prn><qnt><px3sp><dat>
кара<v><iv><gna_perf>
,<cm>
өтө<adv>
жакшы<adj>
э<cop><ger_past><px3sp><acc>
көр<v><tv><ifi><p3><sg>
.<sent>
```

**Kazakh (kaz)**
Кудай Өзінің жаратқандарының бәріне қарап, өте жақсы екенін көрді.

```
Құдай<n><nom>
Өз<n><ref><px3sp><nom>
жарат<v><tv><ger_past><pl><px3sp><gen>
бәрі<prn><qnt><px3sp><dat>
қара<v><iv><gna_perf>
,<cm>
өте<adv>
жақсы<adj>
е<cop><ger_past><px3sp><acc>
көр<v><tv><ifi><p3><sg>
.<sent>
```

**Tatar (tat)**
Аллаһ Үзе яраткан нәрсәләргә карап, аларның бик яхшы иканен курде.

```
Аллаһ<n><nom>
Үз<prn><ref><px3sp><nom>
ярат<v><tv><gpr_past>
нәрсә<n><pl><dat>
,<cm>
алар<prn><pers><p3><pl><gen>
яхшы<adj>
и<cop><ger_past><px3sp><acc>
күр<v><tv><ifi><p3><sg>
.<sent>
```

**Kumyk (kum)**
Аллагъ Оьзю яратгъан затлагъа къарап, олар бек яхшы экенин гёрген.

```
Аллагъ<n><nom>
Оьз<prn><ref><px3sp><nom>
ярат<v><tv><gpr_past>
зат<n><pl><dat>
къара<v><iv><gna_perf>
,<cm>
олар<prn><pers><p3><pl><nom>
яхшы<adj>
э<cop><ger_past><px3sp><acc>
гёр<v><tv><past><p3><sg>
.<sent>
```

**............... Tagset ..................**

| | | | | | | |
|---|---|---|---|---|---|---|
| <n> | Noun | <iv> | Intransitive | <nom> | 'Nominative' | <sent> Sentence | <gna_perf> Verbal adverb (Perfect) |
| <v> | Verb | <tv> | Transitive | <gen> | Genitive | <past> Past (General) | <ger_past> Verbal adjective (Past) |
| <prn> | Pronoun | <p3> | Third person | <acc> | Accusative | <ifi> Past (Eyewitness/Recent) | |
| <det> | Determiner | <pl> | Plural | <dat> | Dative | <px3sp> 3rd person poss. | <ger_past> Verbal noun (Past) |
| <adj> | Adjective | <ref> | Reflexive | <qnt> | Quantifier | | |
| <adv> | Adverb | <pers> | Personal | <cm> | Comma | | |

## Morphophonology

**........... Desonorisation (kaz & kir) ............**

- {N} desonorises to д after a consonant
  алма-{N}{I} → алманы *'apple'*–ACC
  сыр-{N}{I} → сырды *'secret'*–ACC
- {L} desonorises to д after cons. of sonority ≤ /l/
  сыр-{L}{A}р → сырлар *'secret'*–PL
  кыз-{L}{A}р → кыздар *'girl'*–PL

```
"L Desonorisation"
%{L%}:д <=> :VoicedLowSonCns %>: __ ;

"N Desonorisation"
%{N%}:д <=> :VoicedCns %>: __ ;
```

**........... Epenthesis ..............**

- Turn {y} into a harmonised high vowel when a vowel doesn't follow the following consonant:
  мур{y}н → мурун *'nose'*
  мур{y}н+{I}м → мурдум *'my nose'*

```
%{y%}:Vy <=> [ :LastVowel :Cns* :Cns ]/[:0] __
              [ :Cns [ .#. | :Cns ] ]/[ :0 | %>:] ;
  where Vy in ( и у и ү ы ы у у у у )
LastVowel in ( и у е э ө я а ё о ю у )
        matched ;
```

## Morphotactics

**...... Morphological & orthographical words ......**

- өнүктүрөбузбу ?  *'will we develop [it]?'*
  өнүк<v><tv><caus><aor><p1><pl>+бы<qst>
- келатсан *'if you come'*
  кел<v><iv><prc_impf>+жат<vaux><gna_cnd><p2><sg>

**.... Irregular [noun + possessive + case] forms ....**

- Some combinations of possessive + case morphemes are unpredicted (i.e., not formed simply by concatenation and application of phonology):

| case | form | 1SG | 2SG | 3SP |
|---|---|---|---|---|
| nominative | -NI | -(I)м | -(I)ң | -(с)I |
| accusative | -NI | -(I)мдI | -(I)ңдI | -(с)Iн |
| genitive | -DA | -(I)мдIң | -(I)ңдIң | -(с)IндA |
| locative | -DA | -(I)мдA | -(I)ңдA | -(с)IндA |
| ablative | -DAн | -(I)мдAн | -(I)ңдAн | -(с)IндAн |
| | | -(I)мАн | -(I)нАн | |
| dative | -GA | -(I)мА | -(I)нА | -(с)IнА |

A,I,N,D,G have various allophones; (I) null after vowels; (c) null after cons.

- underlying <px3sp> form used: {s}{I}{n}
- {s} and {n} default to с and н; rules map to null by context
- morphophonology more complicated, morphotactics simpler

**........... Noun-noun compounds ..............**

- a N-N compound type: N2 has <px3> and related morphology
  e.g., аба ырайы<n><loc>: аба ырайында, *аба ырайыда

```
LEXICON N-INFL-3PX-COMPOUND
%<n%>:>%>%{S%}%{I%}%{n%} GEN-POS ;

LEXICON Nouns
аба% ырайы:аба% ырай N-INFL-3PX-COMPOUND ;
                                        ! "weather"
чакыруу% кагазы:чакыруу% кагаз N-INFL-3PX-COMPOUND ;
                                        ! "invitation"
```

## Other uses

- HFST transducers are trivially converted to **spell checkers**
- Segmenter, e.g. көргөзгөндөрдөнсүнбу :
  көр>{G}{A}э>{G}{A}н>{L}{A}р>{D}{A}н>с{I}н>{B}{I}

## Orthography-phonology mapping issues

**........... Ambiguous characters ............**

- Have front- and back-vowel readings in native words

| | letters | values | examples |
|---|---|---|---|
| **kaz** | и, у, ю | /əj, əw, jəw/ | киюда *'chopping down'* |
| | | /əj, əw, jəw/ | киюде *'getting dressed'* |
| **tat** | e | /ə / C _ | дәресләр *'lessons'* |
| | | /j/+ы | елллар *'years'* |
| | | /j/+ә | егетләр *'boys'* |
| **kum** | ё, ю | /ø, y/ / C _ | гёзлер *'eyes'* |
| | | /jø, jy/ | юреклер *'hearts'* |
| | | | ёнкюлер *'darlings'* |
| | | /jo, ju/ | юлдузлар *'stars'* |
| | | | ёллар *'roads'* |

- solution: hairy twol rules cover majority of examples
- unaccounted-for words get a harmony-forcing character
- adjust rules for harmony-forcing characters

**........... Loanwords ..............**

- Letters that represent front vowels in native words may represent "back" vowels in Russian words

| | native word example | Russian word example |
|---|---|---|
| kaz | елдің *'country's'* | Назарбаевтың *'Nazarbayev's'* |
| tat | галимнәр *'scientists'* | артистлар *'artists'* |
| kum | сёзлер *'words'* | самолётлар *'airplanes'* |

- solution: separate continuation lexicon (messy rules)

```
LEXICON N1-RUS
:%{;a%} N1 ;

LEXICON Nouns
артист:артист N1-RUS ; ! "artist"
галим:галим N1 ; ! "scientist"
```

**........... Acronyms and numerals ..............**

- twol rules handle phonology for spelt-out words
  отыздан *'from thirty'*, бестен *'from five'*
- no phonological triggers available in numerals
  (incorrect phonological triggers in acronyms)
  30-дан *'from 30'*, 5-тен *'from 5'*
- solution: phonology-triggering characters
- simplified: e.g., {c} for all voiceless obstruents

```
4:4%{э%}%{c%} NUM-DIGIT ; ! "төрт"
5:5%{э%}%{c%} NUM-DIGIT ; ! "бес"
3%0:3%0%{a%}%{э%} NUM-DIGIT ; ! "отыз"
```

**........... й + vowel letters ..............**

- [ а о у ] become [ я ё ю ] after й and й deletes
- й incorporated into the context of many rules
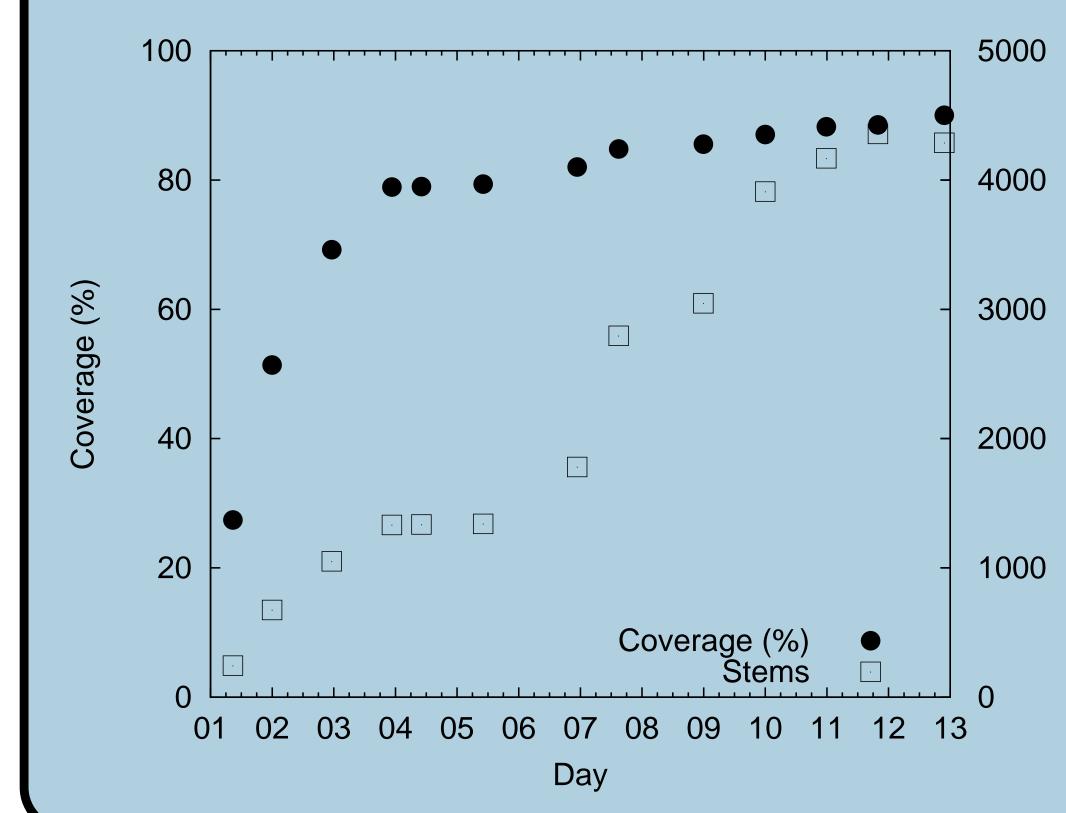- additional rules to change the characters and delete original й

```
"Deletion of й before yoticised vowels"
й:0 <=> __ [ :YotVow ]/[ :0 | %>: ] ;
```

**........... A resulting messy twol rule ..............**

```
RdYotVow:е ё Ю ;
AbstractVow = %{a%}%{э%} %{y%} %{o%} ;

"A front unrounded harmony"
%{A%}:e <=>
        [ [:FrontVow | [ :Vow :ь ] ] :Cns Cns* ]/:0 __ ;
        [ :RdVow :ь ] :Cns :Cns Cns* ]/:0 __ ;
        [ [ \| .#. | :Vow ] :RdYotVow ] :Cns Cns* ]/:0 __ ;
        [ :RdYotVow й:0 :RdYotVow :Cns :Cns* ]/[ :0 - й:0 ] __ ;
      [%{э%}:0|%{y%}:0] :Cns* ]/[ :0 - AbstractVow: ] __ ;
      [ :RdYotVow :Cns* %{A%}:0 :Cns* ]/[ :0 - %{A%}:0 ] __ ;
        [ :Cns :p %{A%}: ь>: :Cns* ]/[ :0 - й:0 ] __ ;
      [ [ :Vow - :RdYotVow ] :RdYotVow :Cns Cns* ]/:0 __ ;
  except
        [ :RdYotVow :Cns* %{A%}:0 :Cns* ]/[ :0 - %{A%}:0 ] __ ;
          [ :Cns :p %{A%}: ь>: :Cns* ]/[ :0 - й:0 ] __ ;
      [ [ :Vow - :RdYotVow ] :RdYotVow :Cns Cns* ]/:0 __ ;
        [ :Vow ]/[ [ :0 - й:0 ] %>: ] __ ;
```

## Development effort

| | **Kyrgyz** | **Kazakh** | **Tatar** | **Kumyk** |
|---|---|---|---|---|
| begun | Apr. 2011 | Dec. 2010 | Dec. 2011 | Oct. 2013 |
| 80% cov. | Aug.? 2011 | Aug. 2012 | Aug. 2012 | Oct. 2013 |
| time | 4 months | 19 months | 7 months | 1 week |

- (various periods of intermission, various rewrites)
- Kazakh transducer based on Kyrgyz transducer
- Kyrgyz transducer currently being rewritten based on insights gained while writing other Turkic transducers
- Kumyk transducer based on Kazakh, Tatar transducers:
  ~1 week to reach 80% coverage, +1 week to reach 90%



## Categorisation

**.................. Adjectives ..................**

- morphologically distinct adjective classes
- most sources claim: adjectives can be used substantively and adverbially
- Other Turkic transducers: θ-derivation (overgenerates)
- but not all adjectives have all of the following:
  comparative forms, substantive readings, adverbial readings
- Our approach: categorisation
- if properly categorised, only correct forms are analysed and generated

| Type | Gloss | <adj><comp> | <adj><comp><subst> | <adj><comp><advl> |
|---|---|---|---|---|
| A1 | *'good'* | яхшы (яхшырак) | яхшы (яхшырак) | яхшы (яхшырак) |
| A2 | *'old'* | иске (искерәк) | иске (искерәк) | иске (искерәк) |
| A3 | *'dead'* | үле (—) | үле (—) | — (—) |
| A4 | *'basic'* | төп (—) | — (—) | — (—) |

**.................. Adverbs ..................**

- Certain adverbs have special attributive and ablative forms
- Mostly time adverbs
- Some also have noun readings: regular ablative, other cases:

| word | бүгүн | быйыл | кечээ | жана |
|---|---|---|---|---|
| gloss | *'today'* | *'this year'* | *'yesterday'* | *'just now'* |
| <attr> form | бүгүнкү | быйылкы | кечээги | жанагы |
| <adv><abl> form | бүгүнтөн | быйылдан | кечээден | жанатан |
| <n><abl> form | бүгүнден | | | |

```
LEXICON ADV-WITH-KI-ABL

ADV ;
ADV-KI ;
ADV-ABL ;
```

## Evaluation

**............... Number of stems ...............**

| Part of speech | | Number of stems | | |
|---|---|---|---|---|
| | **Kyrgyz** | **Kazakh** | **Tatar** | **Kumyk** |
| Noun | 4582 | 2640 | 2795 | 2568 |
| Verb | 1193 | 1470 | 1143 | 386 |
| Adjective | 1211 | 754 | 816 | 219 |
| Proper noun | 5887 | 5701 | 5361 | 1443 |
| Adverb | 312 | 171 | 177 | 63 |
| Numeral | 66 | 63 | 63 | 44 |
| Conjunction | 77 | 46 | 45 | 13 |
| Postposition | 50 | 50 | 43 | 12 |
| Pronoun | 51 | 32 | 28 | 17 |
| Determiner | 64 | 39 | 34 | 9 |
| Total: | 13749 | 11224 | 10737 | 4845 |

**............... Test corpora ...............**

| | News | Religion |
|---|---|---|
| **Kyrgyz** | Wikipedia  azattyk.org | Bible |
| **Kazakh** | Wikipedia  azattyq.org | Quran + Bible |
| **Tatar** | Wikipedia  tat.tatar-inform.ru | Quran + New Testament |
| **Kumyk** | — yoldash.etnosmi.ru | Genesis + New Testament |

**............... Evaluation measures ...............**

- **Naïve coverage** - percentage of surface forms in a given corpus receiving ≥ 1 analysis
- **Mean ambiguity** - average number of analyses for each surface form found in analysed corpus
- **Precision** - probability that a provided analysis is valid
- **Recall** - probability that a certain valid analysis is among those provided by the transducer

**............... Evaluation results ...............**

| | Corpus | Tokens | Coverage (%) | Amb. |
|---|---|---|---|---|
| **Kyrgyz** (r54474) | Wikipedia | 5.3M | 84.51 ± 2.27 | 3.56 |
| | News | 4.1M | 91.43 ± 0.51 | 4.19 |
| | Religion | 215K | 91.66 ± 1.81 | 3.99 |
| | Average | | 89.20 ± 3.48 | 3.91 |
| **Kazakh** (r50547) | Wikipedia | 25.6M | 85.61 ± 1.37 | 2.43 |
| | News | 3.8M | 92.12 ± 2.72 | 2.88 |
| | Religion | 851K | 92.49 ± 1.66 | 2.63 |
| | Average | | 90.07 ± 1.91 | 2.64 |
| **Tatar** (r50260) | Wikipedia | 159K | 86.35 ± 2.17 | 2.24 |
| | News | 5.2M | 89.75 ± 0.07 | 2.30 |
| | Religion | 382K | 91.25 ± 2.55 | 2.24 |
| | Average | | 89.12 ± 1.60 | 2.26 |
| **Kumyk** (r50300) | News | 286K | 91.10 ± 0.86 | 1.53 |
| | Religion | 227K | 92.47 ± 1.03 | 1.53 |
| | Average | | 91.78 ± 0.94 | 1.53 |

- selected & proofed unique random surface forms from news corpora

| Language | Forms | Precision (%) | Recall (%) |
|---|---|---|---|
| Kyrgyz | 200 | 90.77 | 69.15 |
| Kazakh | 1000 | 98.61 | 57.98 |
| Tatar | 1000 | 95.03 | 85.65 |
| Kumyk | 500 | 96.57 | 69.11 |

## Ongoing and future work

- Disambiguation, more stems, clean up transducers
- Machine translation between these languages
- Bring other Kypchak transducers to comparable performance:
  Qaraqalpaq, Bashqort, Nogay, Crimean Tatar
- Other Turkic lgs: Uzbek, Uyghur, Chuvash, Yakut, Tuvan, etc.