Finite-state morphological transducers for three Kypchak languages

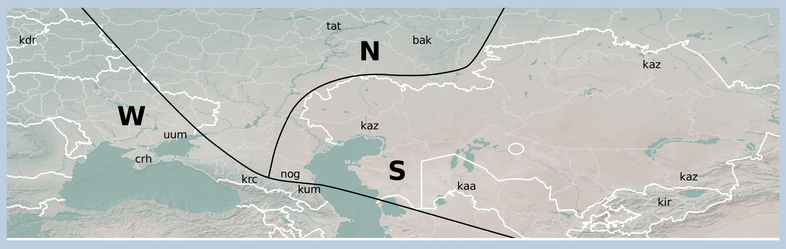
Jonathan North Washington
Indiana University
jonwashi@indiana.edu

Ilnar Salimzyanov
Some University
ilnar.salimzyan@gmail.com

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

Also special thanks to
Aida Sundetova
email@email

Kypchak languages



- Turkic languages (SOV, agglutinative, vowel harmony)

	Kazakh	Tatar	Kumyk
	population of speakers		
number	8M-12M	5.4M	430K
pronunc	/qɑzɑq/	/tɒtɑr/	/qumuq/
primary	Kazakhstan	Tatarstan	Dagestan
secondary	China, Mongolia		
	external influences		
Mongolic	moderate	light	light
Oghuz	—	light	moderate
Persian	heavy	heavy	heavy
Russian	heavy	heavy	heavy

Morphological transducers

Morphological transducers

- 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; Öflazer, 1994)
- Crimean Tatar (Altıntaş, 2001)
- Turkmen (Tantuğ et al., 2006)
- Kyrgyz (Tyers et al., 2012)
- GPL (=free and open)!

Framework: HFST

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

Development effort

Morphotactics

Morphological & orthographical words

- өнүктүрөбүзбү ? *‘will we develop [it]?’*
  `өнүк<v><tv><caus><aor><p1><pl>+бы<qst>`
- келатсаң *‘if you come’*
  `кел<v><iv><prt_impf>+жат<vaux><gna_cnd><p2><sg>`

Irregular [noun + possessive + case] forms

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

case	form	1SG	2SG	3SP
nom	—	-(I)м	-(I)ң	-(S)I
acc	-NI	-(I)мдI	-(I)ңдI	**-(S)Iн**
gen	-NIн	-(I)мдIн	-(I)ңдIн	-(S)IнIн
loc	-DA	-(I)мдA	-(I)ңдA	**-(S)IндA**
abl	-DAн	-(I)мдAн, **-(I)мAн**	-(I)ндAн, **-(I)ңAн**	**-(S)IнAн**
dat	-GA	**-(I)мA**	**-(I)ңA**	**-(S)IнA**

- Trade-off:
  - morphophon. complicateder, morphotactics simpler
  - underlying form used: `{S}{I}{n}`
  - phonological rules delete `{n}`, `{S}` by context

Noun-noun compounds

- one type of N-N compunds: N2 has `<px3>` and related morphology

```
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”
```

Example output

Kazakh	Tatar	Kumyk
Құдай өзінің жаратқандарының бәріне қарап, өте жақсы екенін көрді.	Аллаһ Үзе яраткан нәрсәләргә карап, аларның бик яхшы икәнен күрде.	Аллагъ Оьзю яратгъан затлагъа къарап, олар бек яхшы экенин гёрген.
Құдай<n><nom> өз<prn><ref><px3sp><gen> жарат<v><tv><ger_past><pl><px3sp><gen> бәрі<prn><qnt><px3sp><dat> қара<v><tv><gna_perf> ,<cm> — өте<adv> жақсы<adj> е<cop><ger_past><px3sp><acc> көр<v><tv><ifi><p3><sg> .<sent>	Аллаһ<n><nom> Үз<prn><ref><px3sp><nom> ярат<v><tv><gpr_past> нәрсә<prn><itg><pl><dat> кара<v><tv><gna_perf> ,<cm> алар<prn><pers><p3><pl><gen> бик<adv> яхшы<adj> и<cop><ger_past><px3sp><acc> күр<v><tv><past><p3><sg> .<sent>	Аллагъ<n><nom> Оьз<prn><ref><px3sp><nom> ярат<v><tv><gpr_past> зат<n><pl><dat> къара<v><tv><gna_perf> ,<cm> олар<prn><pers><p3><pl><nom> бек<adv> яхшы<adj> э<cop><ger_past><px3sp><acc> гёр<v><tv><past><p3><sg> .<sent>

Gloss

(1) Аллагъ Оьзю яратгъан затларгъа къарап, олар бек яхшы экенин гёрген.
God own-his made thing-s-to look-having, they very good being saw.

‘God looked at everything he had made and saw that it was very good.’

Output

```
^Үстөл/Үстөл<n><nom>$
^жана/жан<v><iv><prc_impf>/жана<adv>/жана<cnjcoo>$
^отургучтардын/отургуч<n><pl><gen>$
^астын/аст<n><px3pl><acc>/аст<n><px3sg><acc>$
^карап/кара<v><iv><gna_perf>/кара<v><iv><prc_real>/кара<v><tv><gna_perf>/кара<v><tv><prc_real>$
^жатат/жат<vaux><aor><p3><pl>/жат<vaux><aor><p3><sg>/жат<vaux><prc_irre>$ (intransitive verb forms removed)
^,/,<cm>$
^бирок/бирок<cnjadv>$
^Азамат/Азамат<np><ant><m><nom>$
^аякта/ал<det><dem>+жак<n><loc>/аяк<n><loc>/аякта<v><tv><imp><p2><sg>$
^эмес/э<cop><neg><p3><pl>/э<cop><neg><p3><sg>$
^./.<sent>$
```

Tagset

`<n>`	Noun	`<p2>`	Second person	`<px3sg>`	3rd person poss. (Singular)
`<np>`	Proper noun	`<p3>`	Third person	`<px3pl>`	3rd person poss. (Plural)
`<v>`	Verb	`<ant>`	Anthroponym	`<neg>`	Negative
`<det>`	Determiner	`<dem>`	Demonstrative	`<aor>`	Aorist
`<cnjcoo>`	Coord. conjunct.	`<m>`	Masculine	`<imp>`	Imperative
`<cnjadv>`	Adv. conjunct.	`<sg>`	Singular	`<gna_perf>`	Verbal adverb (Perfect)
`<adv>`	Adverb	`<pl>`	Plural	`<prc_impf>`	Participle (Imperfect)
`<vaux>`	Auxiliary verb	`<nom>`	‘Nominative’	`<prc_irre>`	Participle (Irrealis)
`<cop>`	Copula	`<gen>`	Genitive	`<prc_real>`	Participle (Realis)
`<iv>`	Intransitive	`<acc>`	Accusative	`<cm>`	Comma
`<tv>`	Transitive	`<loc>`	Locative		

Morphophonology

Desonorisation

- {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 %>: __ ;
```

Lenition

- 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 ү и и ү ы ы у у ы у у )
LastVowel in ( и ү е э ө я а ё о ы ю у )
             matched ;
```

й+vowel letters

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

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

Evaluation

Number of stems

Part of speech	Number of stems		
	Kazakh	Tatar	Kumyk
Noun	2640	2795	2568
Verb	1470	1143	386
Adjective	754	816	219
Proper noun	5701	5361	1443
Adverb	171	177	63
Numeral	63	63	44
Conjunction	46	45	13
Postposition	50	43	12
Pronoun	32	28	17
Determiner	39	34	9
Total:	11224	10737	4845

Test corpora

Encyclop	kaz	wpdump	20131006
	tat	wpdump	20130225
	kum	—	—
News	kaz	RFE/RL	azattyq.org 2010
	tat	Татар-информ	tat.tatar-inform.ru 200
	kum	Ёлдаш	yoldash.etnosmi.ru
Religion	kaz	quran + bible	kkitap.net, kuran.kz
	tat	quran + nt	ibt.org.ru, tanzil.net
	kum	genesis + nt	ibt.org.ru

- split into 10 equal parts; coverage calculated over each separately; standard deviation of mean calculated

Coverage measures

- **Naïve coverage** - percentage of surface forms in a given corpus receiving ≥ 1 analysis (surface forms may have missing analyses)
- **Mean ambiguity** - average number of analyses for each surface form found in analyed corpus

Coverage results (as of r36739)

Language	Corpus	Tokens	Coverage (%)
Kazakh	Wikipedia	25.6M	85.61 ± 1.37
	News	3.8M	92.12 ± 2.72
	Religion	851K	92.49 ± 1.66
	Average	–	90.07 ± 1.91
	Wikipedia	159K	86.35 ± 2.17

Further information