



# FINITE-STATE MORPHOLOGICAL TRANSDUCERS FOR THREE KYPCHAK LANGUAGES

Jonathan North Washington Ilnar Salimzyanov

Indiana University  
jonwashi@indiana.edu

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

Francis M. Tyers

UiT Norgga Árktaš Universitehta  
francis.tyers@uit.no

Special thanks to  
Aida Sundetova  
sun27aida@gmail.com



- Turkic languages (SOV, agglutinative, vowel harmony)

	Kazakh	Tatar	Kumyk
population of speakers			
number	8M-12M	5.4M	430K
pronunc	/qazaq/	/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

- 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

## 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)ңдAн,	-(S)IнAн
		-(I)мAн	-(I)ңAн	
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 compounds: 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"

## Gloss

- (1) Аллагы Обзю яратгъан затларгъа къарап, олар бек яхшы экенин гөрген.  
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

Kazakh	Tatar	Kumyk
Кудай Өзінің жаратқандарының бәріне қарап, өте жақсы екенін көрді.	Аллаһ Үзе яраткан нәрсәләргә карап, аларның бик яхшы икәнен күрдә.	Аллагы Обзю яратгъан затлагъа къарап, олар бек яхшы экенин гөрген.
Кудай<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>	Аллагы<n><nom> Обз<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> гөрг<v><tv><past><p3><sg> .<sent>

## Tagset

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

## 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 ≤ /I/  
сыр-{L}{A}p → сырлар 'secret-PL'
- кыз-{L}{A}p → 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 у у )  
LastVowel in ( и ү е э ө я а ё о ы 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 yoticed vowels"

й:0 <=> \_\_ [ :YotVow ]/[[:0] | %>: ] ;

- Part of Apertium Turkic project:  
http://wiki.apertium.org/wiki/Apertium\_Turkic
- Transducers available live at turkic.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 paper in the LREC 2014 proceedings
- And feel free to contact the authors any time!

## 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

type	lang	contents	origin
Encyclop	kaz tat kum	wpdump wpdump —	20131006 20130225 —
News	kaz tat kum	RFE/RL Татар-информ Ёлдаш	azattyq.org 2010 tat.tatar-inform.ru 2005-2011 yoldash.etnosmi.ru
Religion	kaz tat kum	quran + bible quran + nt genesis + nt	kkitap.net, kuran.kz ibt.org.ru, tanzil.net 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 analysed 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
Tatar	Wikipedia	159K	86.35 ± 2.17
	News	5.2M	89.75 ± 0.07
	Religion	382K	91.25 ± 2.55
	Average	—	89.12 ± 1.60
Kumyk	Wikipedia	—	—
	News	286K	91.10 ± 0.86
	Religion	227K	92.47 ± 1.03
	Average	—	91.78 ± 0.94

## Precision & recall

- selected 1000 surface forms at random from RFE/RL corpus, proof read analyses
- Precision** (of a form's analyses % correct): **97.32%**
- Recall** (percentage of analyses provided by the transducer that are correct for a form, by comparing against a gold standard): **94.56%**