Finite-state morphological transducers for three 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
Aida Sundetova
sun27aida@gmail.com

Kypchak languages

- Turkic languages (SOV, agglutinative, vowel harmony)

	Kazakh	**Tatar**	**Kumyk**
	/qɑzɑq/	/tɤtɑr/	/qumuq/
classif'tion	S Kypchak	N Kypchak	W Kypchak
population of speakers			
number	8M-12M	5.4M	430K
primary	Kazakhstan	Tatarstan	Dagestan
secondary	China, Mongolia	Bashqortostan	—
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

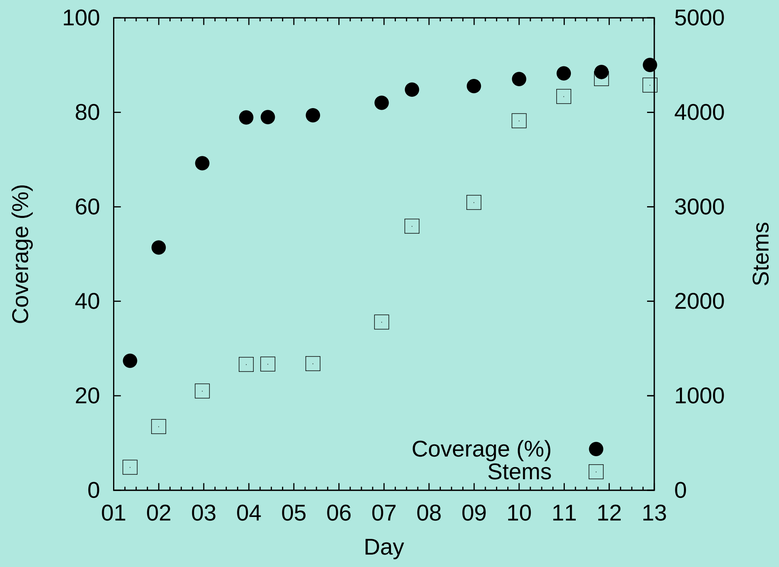
- Turkish (Çöltekin, 2010 & 2014; Öflazer, 1994)
- Crimean Tatar (Altıntaş, 2001)
- Turkmen (Tantuğ et al., 2006)
- Kyrgyz (Washington et al., 2012)
- Kazakh (Бекманова & Махимов, 2013)
- our Kazakh, Tatar, Kumyk: all GPL (=free and open)!

Framework: HFST

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

Development effort

- Kumyk transducer based on Kazakh, Tatar transducers
- ~1 week to reach 80% coverage, +1 week to reach 90%



Categorisation

- Other Turkic transducers: `0`-derivation (overgenerates)
- Our approach: categorization (e.g., adjectives, below)

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

Further information

- 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!

Example output

Gloss

(1)	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.
	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.'

Output

Kazakh (kaz)	**Tatar** (tat)	**Kumyk** (kum)
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>	Аллаһ<n><nom>	Аллагъ<n><nom>
өз<prn><ref><px3sp><gen>	Үз<prn><ref><px3sp><nom>	Оьз<prn><ref><px3sp><nom>
жарат<v><tv><ger_past><pl><px3sp><gen>	яраткан<v><tv><gpr_past>	яратгъан<v><tv><gpr_past>
бәрі<prn><qnt><px3sp><dat>	нәрсә<n><pl><dat>	зат<n><pl><dat>
қара<v><tv><gna_perf>	карап<v><tv><gna_perf>	къарап<v><tv><gna_perf>
,<cm>	,<cm>	,<cm>
—	алар<prn><pers><p3><pl><gen>	олар<prn><pers><p3><pl><nom>
өте<adv>	бик<adv>	бек<adv>
жақсы<adj>	яхшы<adj>	яхшы<adj>
е<cop><ger_past><px3sp><acc>	и<cop><ger_past><px3sp><acc>	э<cop><ger_past><px3sp><acc>
көр<v><tv><ifi><p3><sg>	күр<v><tv><past><p3><sg>	гёр<v><tv><past><p3><sg>
.<sent>	.<sent>	.<sent>

Tagset

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

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/	қиюда '*in the process of chopping down*'
		/əj, əw, jəw/	киюде '*in the process of getting dressed*'
tat	е	э / C _	дәрәсләр '*lessons*'
		/j/+ы	еллар '*years*'
		/j/+э	егетләр '*boys*'
kum	ё, ю	/ø, y/ / C _	гёзлер '*eyes*', гюнлер '*days*'
		/jø, jy/	юреклер '*hearts*', ёнкюлер '*darlings*'
		/jo, ju/	юлдузлар '*stars*', ёллар '*roads*'

- solution: hairy twol rules for majority of cases
- unaccounted-for words marked harmony-forcing char
- adjust rules for harmony-forcing characters

ugly twol example goes here

Loanwords

•

Kazakh - елді / Назарбаевты Tatar - галимнәр, артистлар Kumyk - сёзлер, самолётлар - separate continuation lexicon - result: super messy twol rules

Acronyms and numerals

отыздан, бестен handled by twol 30-дан, 5-тен not handled by twol

```
5{э}{с}>-{D}{A}н
```

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

	Wikipedia	**News**	**Religion**
kaz	Wikipedia	azattyk.org	Quran + Bible
tat	Wikipedia	tat.tatar-inform.ru	Quran + New Testament
kum	—	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** - of a form's analyses, % correct
- **Recall** - % of analyses provided by transducer that are correct for a form, by comparing against a gold standard

Evaluation results

Language	**Corpus**	**Tokens**	**Coverage** (%)	**Amb.**
Kazakh	Wikipedia	25.6M	85.61 ± 1.37	0.00
	News	3.8M	92.12 ± 2.72	0.00
	Religion	851K	92.49 ± 1.66	0.00
(r50547)	Average		90.07 ± 1.91	0.00
Tatar	Wikipedia	159K	86.35 ± 2.17	0.00
	News	5.2M	89.75 ± 0.07	0.00
	Religion	382K	91.25 ± 2.55	0.00
(r50260)	Average		89.12 ± 1.60	0.00
Kumyk	News	286K	91.10 ± 0.86	0.00
	Religion	227K	92.47 ± 1.03	0.00
(r50300)	Average		91.78 ± 0.94	0.00

- selected & proofed unique random surface forms from news corpora

Language	**Forms**	**Precision** (%)	**Recall** (%)
Kazakh	1000	98.61	57.98
Tatar	1000	95.03	85.65
Kumyk	500	96.57	69.11

Future Work

- Disambiguation (already exists for Kazakh)
- More stems (especially Kumyk)
- Machine translation between these languages