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

Apertium

- Turkic languages (SOV, agglutinative, vowel harmony)

	Kazakh	**Tatar**	**Kumyk**
	/qɑzɑq/	/tɒtɑr/	/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 ....

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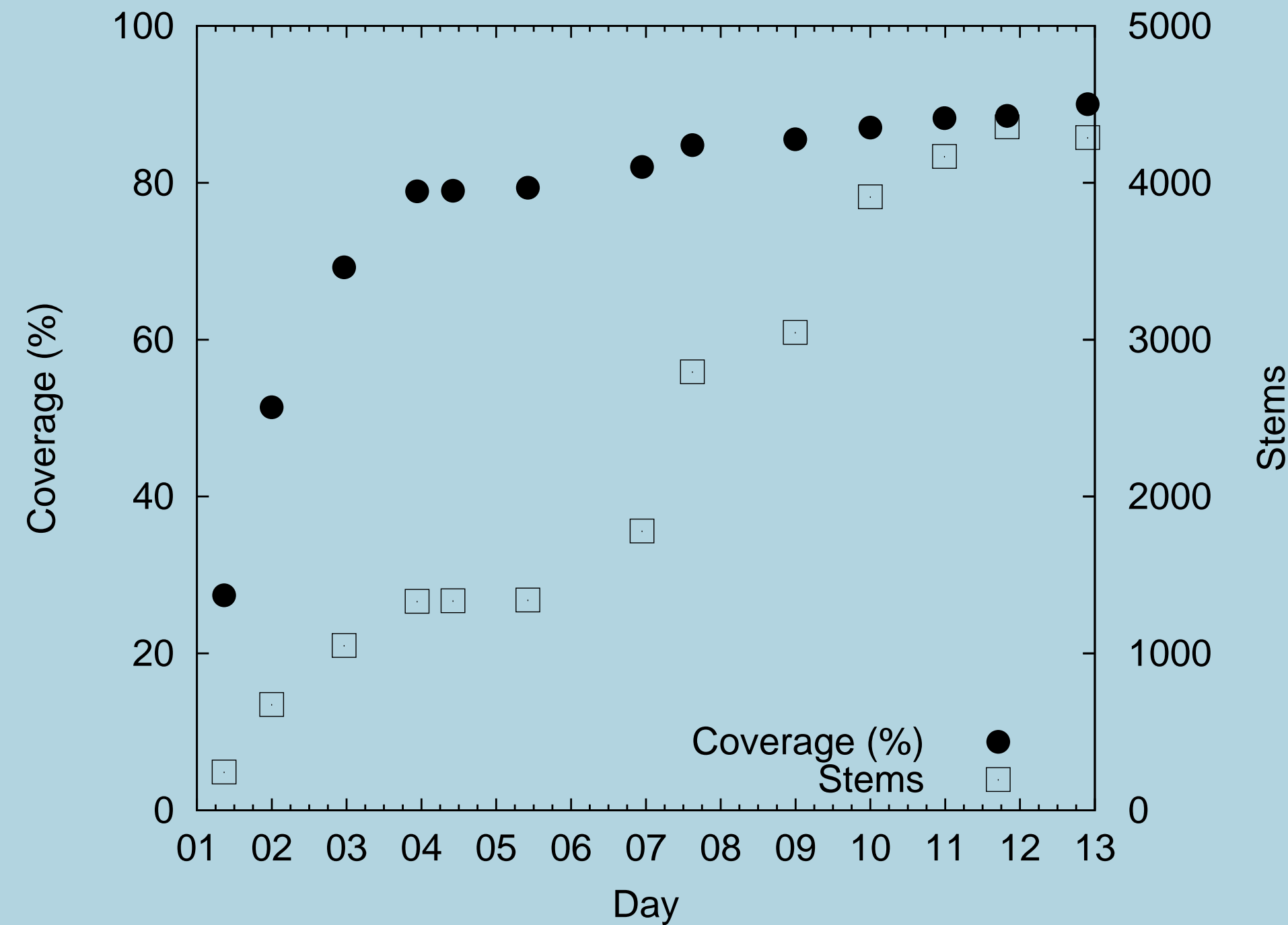
- Other Turkic transducers: 0-derivation (overgenerates)
- Our approach: categorization (e.g., adjectives, below)

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

- 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!

.... Gloss ....

(1)
Құдай	Өзінің	жаратқандарының	бәріне		қарап,		өте	жақсы	екенін	көрді.
Аллаһ	Үзе	яраткан	нәрсәләргә		карап,	аларның	бик	яхшы	икәнен	күрде.
Аллагь	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)

Құдай өзінің жаратқандарының бәріне қарап, өте жақсы екенін көрді.

```
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>
—
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 икәнен күрде.

```
Аллаһ<n><nom>
Үз<prn><ref><px3sp><nom>
ярат<v><tv><gpr_past>
нәрсә<n><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>
```

Kumyk (kum)

Аллагь Оьзю яратгъан затлагъа къарап, олар бек яхшы экенин гёрген.

```
Аллагь<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>
```

.... Tagset ....

<n>	Noun	<p3>	Third person	<pers>	Personal
<v>	Verb	<pl>	Plural	<cm>	Comma
<prn>	Pronoun	<nom>	'Nominative'	<sent>	Sentence
<det>	Determiner	<gen>	Genitive	<past>	Past (General)
<adj>	Adjective	<acc>	Accusative	<ifi>	Past (Eyewitness/Recent)
<adv>	Adverb	<dat>	Dative		
<iv>	Intransitive	<qnt>	Quantifier		
<tv>	Transitive	<ref>	Reflexive		

<px3sp>	3rd person poss. (Singular/Plural)
<gna_perf>	Verbal adverb (Perfect)
<gpr_past>	Verbal adjective (Past)
<ger_past>	Verbal noun (Past)

.... Ambiguous characters ....

-

Tatar - e (non-initial orthographic variant of ә, e.g. дәреслǝр) - initial position /j/+ә (егетлǝр) and /j/+ы (еллар) Kumyk - ё ю (front round vowels between consonants гёзлер, гюнлер) - word-initially /j/ followed by front (юреклер, ёнкюлер) or back vowel (юлдузлар, ёллар) Kazakh - и у ю (front/back vowels, along with glides) - ǝиюда/киюде **.. Loanwords ..** Kazakh - елдіǝ / Назарбаевтыǝ Tatar - галимнǝр, 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}{c}>-{D}{A}н
```

.... 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
Kazakh	Wikipedia	azattyk.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** - 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

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