Finite-state morphological transducers for three Kypchak languages

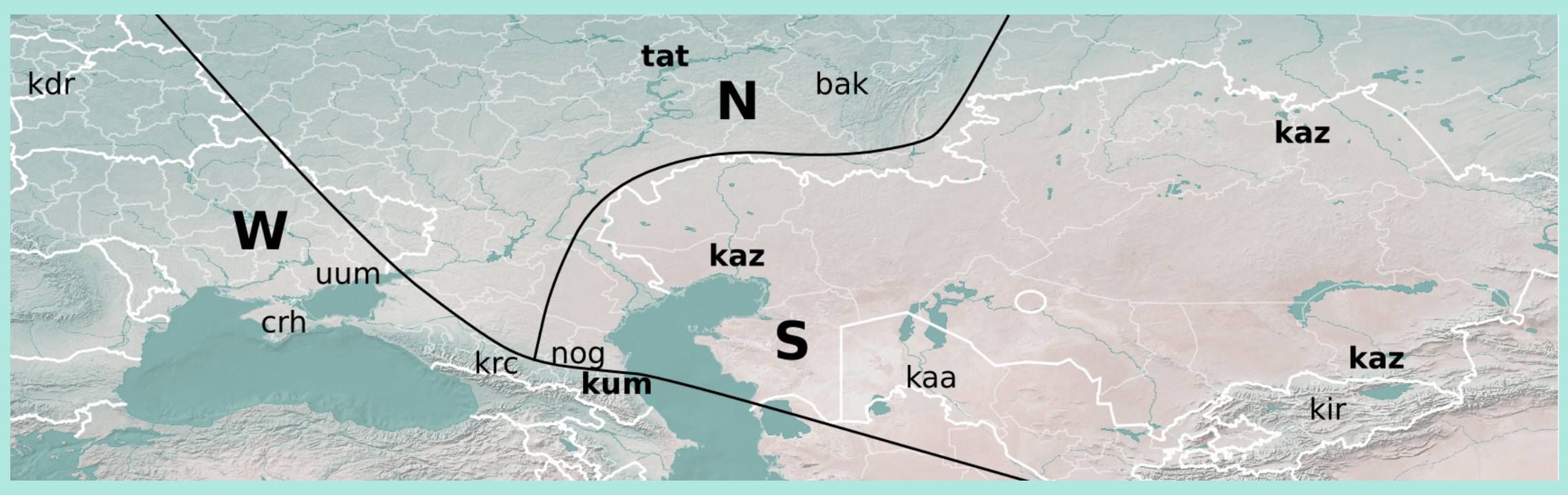
Jonathan North Washington
Indiana University
jonwashi@indiana.edu

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
A. Sultanmuradov

Kypchak languages



- Turkic languages (SOV, agglutinative, vowel harmony)

	Kazakh	Tatar	Kumyk
	/qɑzɑq/	/tɒtɑr/	/qumuq/
classification	Southern	Northern	Western
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

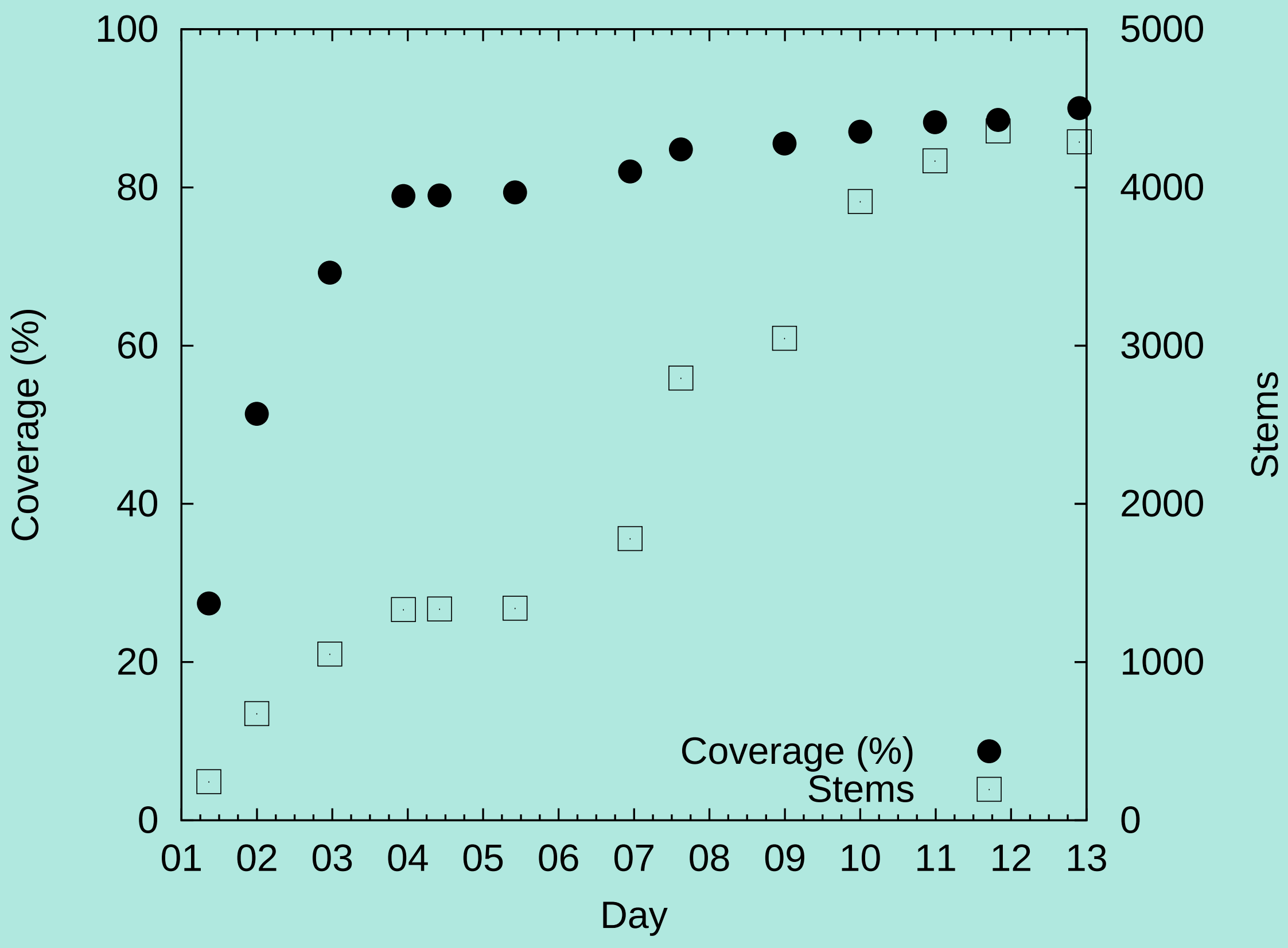
- Turkish (Çöltekin, 2010 & 2014; Öflazer, 1994)
- Crimean Tatar (Altıntaş, 2001)
- Turkmen (Tantuğ et al., 2006)
- Kyrgyz (Washington et al., 2012)
- Kazakh (Бекманова & Махимов, 2013)
- our Kazakh, Tatar, Kumyk: all GPL (=free and open)!

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 is 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.
	Аллаh	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>
,<cm>
—
өте<adv>
жақсы<adj>
е<cop><ger_past><px3sp><acc>
көр<v><tv><ifi><p3><sg>
.<sent>
```

Tatar (tat)

```
Аллаh Үзе яраткан нәрсәләргә карап,
аларның бик яхшы икәнен күрде.
```

```
Аллаh<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	<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/ /ɘj, ɘw, jɘw/	қиюда *'chopping down'* киюде *'getting dressed'*
tat	е	э / C _ /j/+ы /j/+э	дәресләр *'lessons'* еллар *'years'* егетләр *'boys'*
kum	ё, ю	/ø, y/ / C _ /jø, jy/ /jo, ju/	гюнлер *'days'* гёзлер *'eyes'* юреклер *'hearts'* ёнкюлер *'darlings'* юлдузлар *'stars'* ёллар *'roads'*

- solution: hairy twol rules cover majority of examples
- unaccounted-for words marked harmony-forcing char
- 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
:%{☭%} 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 (etc.) 30-дан *'from 30'*, 5-тен *'from 5'*
- solution: phonology-triggering characters

```
4:4%{э%}%{с%} NUM-DIGIT ; !  ”төрт”
5:5%{э%}%{с%} NUM-DIGIT ; !  ”бес”
3%0:3%0%{а%}%{з%} NUM-DIGIT ; !  ”отыз”
```

A resulting messy twol rule

```
RdYotVow = ё ю Ё Ю ;
AbstractVow = %{а%} %{э%} %{у%} %{о%} ;

”A front unrounded harmony”
%{A%}:е <=>     [ [ :FrontVow | [ :Vow :ь ] ] :Cns :Cns* ]/:0 _ ;
                                  [ [ :RdVow :ь ] :Cns :Cns* ]/:0 _ ;
           [ [ [ \[ .#. | :Vow ] ] :RdYotVow ] :Cns :Cns* ]/:0 _ ;
           [ :RdYotVow й:0 :RdYotVow :Cns :Cns* ]/[ :0 - й:0 ] _ ;
   [[%{э%}:0|%{у%}:0] :Cns* ]/[[ :0 - AbstractVow:  ]  | %-:  ]* _ ;
        except
            [ :RdYotVow :Cns* %{☭%}:0 :Cns* ]/[ :0 - %{☭%}:0 ] _ ;
                               [ :Cns :р %{☭%}:  %>:  :Cns* ]/:0 _ ;
              [ [ :Vow - :RdYotVow ] :RdYotVow :Cns :Cns* ]/:0 _ ;
                               [ :Vow ]/[ [ :0 - й: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

	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

	Corpus	Tokens	Coverage (%)	Amb.
Kazakh	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
(r50547)	Average		90.07 ± 1.91	2.64
Tatar	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
(r50260)	Average		89.12 ± 1.60	2.26
Kumyk	News	286K	91.10 ± 0.86	1.53
	Religion	227K	92.47 ± 1.03	1.53
(r50300)	Average		91.78 ± 0.94	1.53

- 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

Ongoing and future work

- Disambiguation, more stems
- Machine translation between these languages
- Other Turkic lgs.: Nogay, Bashqort, Uzbek, Chuvash