







Jonathan North Washington Ilnar Salimzyanov Indiana University jonwashi@indiana.edu

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

Francis M. Tyers

DESIGNING FINITE-STATE MORPHOLOGICAL TRANSDUCERS

FOR KYPCHAK LANGUAGES

UiT Norgga Árktalaš Universitehta francis.tyers@uit.no

Part of speech

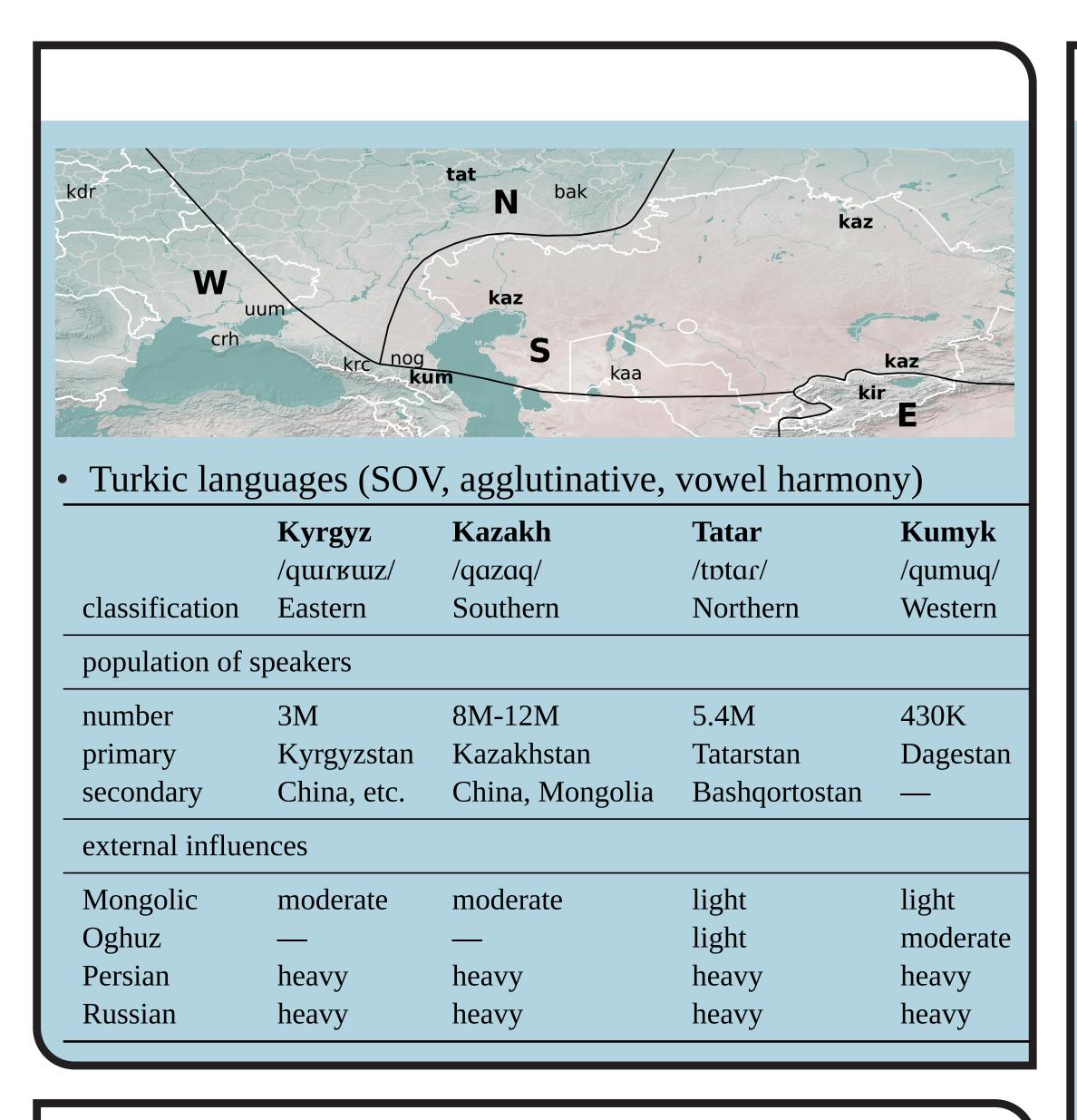
Special thanks to: Tolgonay Kubatova Aida Sundetova Ağarahim Sultanmuradov

. Number of stems

Number of stems

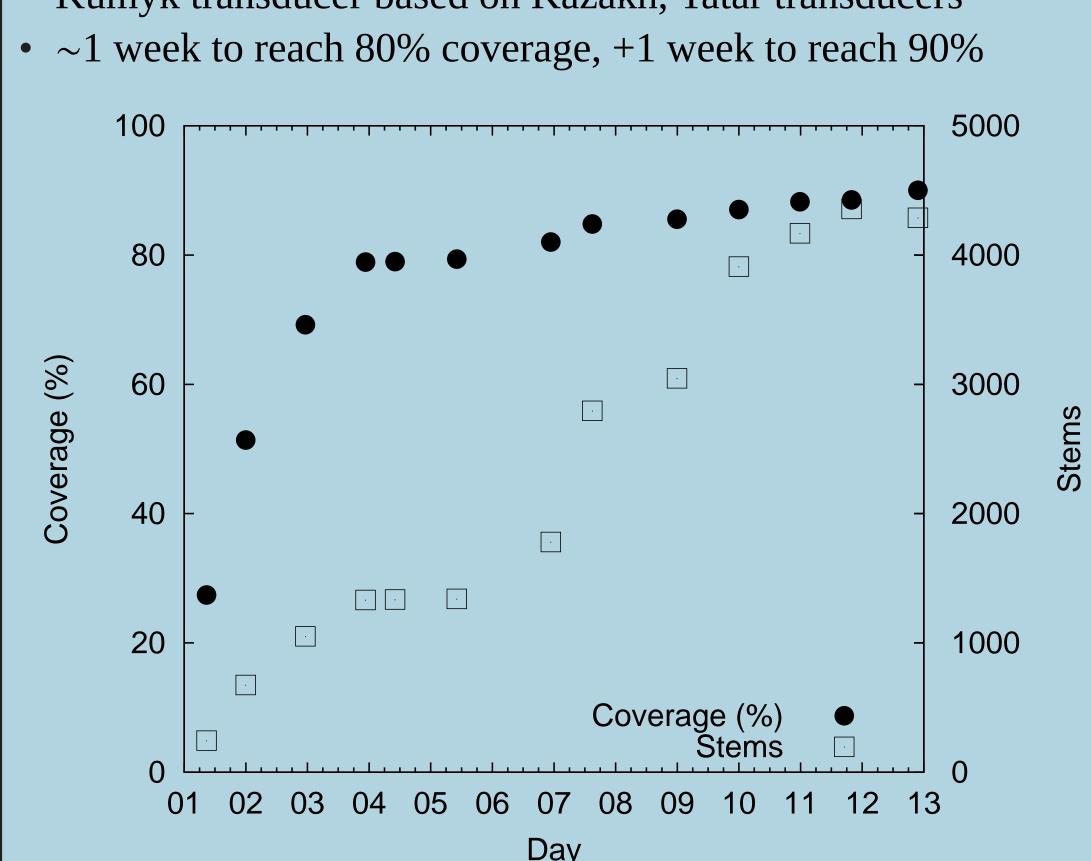
.

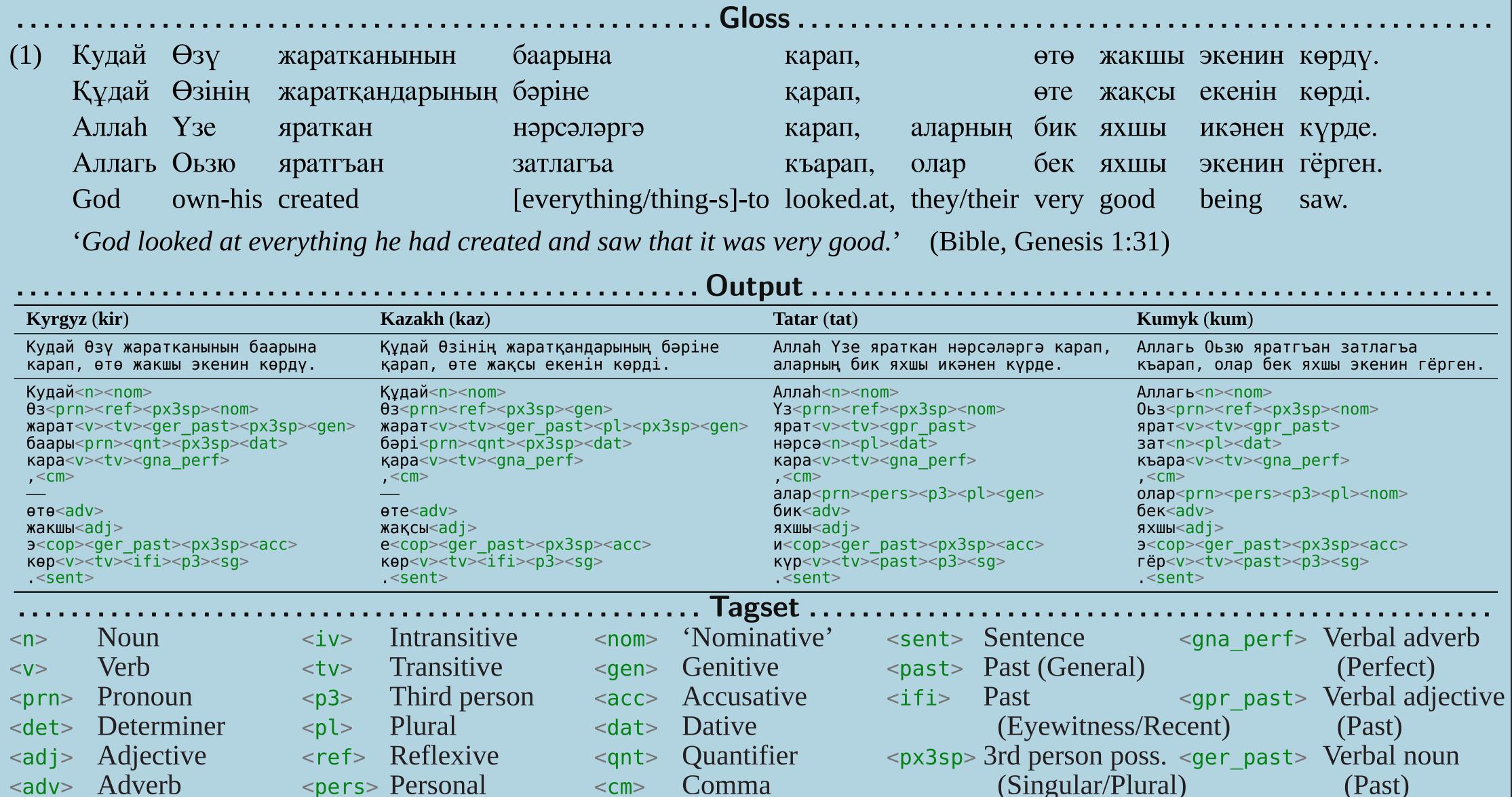




..... Morphological transducers

- 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; Oflazer, 1994)
- Crimean Tatar (Altıntaş, 2001)
- Turkmen (Tantuğ et al., 2006)
- Kazakh (Бекманова & Махимов, 2013)
- our Kyrgyz, 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
- Kumyk transducer based on Kazakh, Tatar transducers





- Example: morphologically distinct adjective classes (not documented elsewhere)
- Other Turkic transducers: 0-derivation (overgenerates)
- Our approach: categorisation (generates/analyses only correct

A1 'good' яхшы (яхшырак) яхшы (яхшырак) яхшы (яхшы (яхшырак) A2 'old' иске (искерэк) — (—) A3 'dead' үле (—) — (—) A4 'basic' төп (—) — (—)	ыр ак)

Have front- and back-vowel readings in native words				
	letters	values	examples	
kaz	и, у, ю	/wej, we, jew/ /wej, we, jev/	қиюд <mark>а</mark> 'chopping down' киюд <mark>е</mark> 'getting dressed'	
tat	e	э / С _ /j/+ы /j/+э	дәресләр 'lessons' еллар 'years' егетләр 'boys'	
		14 x1 1 C	гюнлер 'days'	

..... Ambiguous characters

	ë, ю	/ø, y/ / C _	гюнл <mark>е</mark> р 'days' гёзл <mark>е</mark> р 'eyes'
kum		/jø, jy/	юреклер 'hearts' ёнкюлер 'darlings'
		/jo, ju/	ёнкюлер 'darlings' юлдузлар 'stars' ёллар 'roads'
• colut	ion: hai	ry tuol rules e	over majority of evar

- solution: hairy twol rules cover majority of examples
- unaccounted-for words get a harmony-forcing character
- 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
- Назарбаевтың 'Nazarbayev's' kaz елдің 'country's' tat галимнәр 'scientists' артистлар 'artists' kum сёзлер 'words' самолётлар 'airplanes'
- solution: separate continuation lexicon (messy rules) LEXICON N1-RUS
- LEXICON Nouns артист:артист N1-RUS ; ! "artist"

:%{»%} N1 ;

- галим:галим N1 ; ! "scientist"
- Acronyms and numerals twol rules handle phonology for spelt-out words

		t of sp							
	1 41	t or sp		Kazal	kh T	atar	Kur	nyk	
	Nou	ın		264	40	2795	2	2568	
	Ver			147	70	1143		386	
		ective			54	816		219	
	Proper noun			570		5361	1	.443	
	Adverb				71	177		63	
	Numeral Conjunction				53 46	63 45		44 13	
	Postposition				50	43		12	
	Pronoun				32	28		17	
	Determiner				39	34		9	
	Tota	al:		1122	24 1	0737	4	845	
			т	est co	rpora				
	Wikip	edia N	News		Relig	ion			
kaz	Wikip		zattyk.or			ı + Bib			
tat kum	Wikip —		at.tatar-in oldash.et	itorm.ru :nosmi.ru	•		v Testam ew Testa		
	Evaluation measures								
Naï	ve cov	verage	- perce	entage o	of surfa	ce fo	rms in	a giv	
		/ing ≥ 1	_					O	
Mea	an am	biguit	y - ave	rage nu	ımber o	of ana	lyses f	or ea	
face	e form	found	in anal	ysed co	orpus				
Pre	cision	- of a	form's	analyse	es, % c	orrect			
Rec	all - 9	% of an	alyses p	provide	d by tr	ansdu	cer tha	it are	
for a	a form	ı, by co	mparir	ng again	ıst a go	old sta	ındard		
Evaluation results									
		Corpu	s T	okens	Cove	rage (%) A	mb.	
		Wikipe	edia :	25.6M		$1\pm1.$		2.43	
Kaza	akh	News		3.8M		2 ± 2 .		2.88	
		Religio		851K		$9 \pm 1.$		2.63	
(r505	547)	Averag	<u></u>			7 ± 1.9		2.64	
		Wikipe News	edia	159K 5.2M		$5\pm 2.5\pm 0.5$		2.24	
Tata	r	Religio	n	5.2111	09./3	$\supset \perp \cup .$		γ	
			/11	382K	91 2	5 + 2	55	2.30 2.24	
	- 260)		e	382K		$\frac{5 \pm 2}{2 \pm 1}$.		2.30 2.24 2.26	
(r502	260)	Averag	je		89.1	$2\pm1.$	60	2.24 2.26	
				382K 286K 227K	89.12 91.10		60 86	2.24	
(r502	nyk _	Averag News	n	286K	89.12 91.10 92.4	$2 \pm 1.$ $0 \pm 0.$	60 86 03	2.242.261.53	
(r502 Kun (r503	nyk - 300)	Averag News Religio	e je	286K 227K	89.12 91.10 92.4 91.78	$2 \pm 1.$ $0 \pm 0.$ $7 \pm 1.$ $8 \pm 0.$	60 86 03 94	2.24 2.26 1.53 1.53	
(r502 Kun (r503	nyk - 300)	Average News Religion Average	e je	286K 227K ndom sur	89.12 91.10 92.4 91.78	$2\pm1.$ $0\pm0.$ $7\pm1.$ $8\pm0.$	60 86 03 94 n news o	2.24 2.26 1.53 1.53	
(r502 Kun (r503 select	nyk 300) cted & p	Average News Religion Average roofed under the second seco	n e nique rai	286K 227K ndom sur	89.12 91.10 92.4 91.78 face form	$2\pm1.$ $0\pm0.$ $7\pm1.$ $8\pm0.$	60 86 03 94 n news o	2.24 2.26 1.53 1.53 2.000000000000000000000000000000000000	
(r502 Kun (r503 select	nyk 300) eted & p	Average News Religion Average roofed under the second seco	n e nique rai	286K 227K ndom sur	89.12 91.10 92.4 91.78 face form	2 ± 1 . 0 ± 0 . 7 ± 1 . 0 ± 0 . ms from 0 ± 0 .	60 86 03 94 n news o	2.24 2.26 1.53 1.53 2.20 2.20 2.20 2.20 2.20 2.20 2.20 2.2	

- Disambiguation, more stems
- Machine translation between these languages
- Bring our other Kypchak transducers to comparable perfor-
- Qaraqalpaq, Bashqort, Nogay, Crimean Tatar
- Other Turkic lgs.: Uzbek, Uyghur, Chuvash, Yakut, Tuvan,

- 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 papers in LREC proceedings (2012: Kyrgyz, 2014: Kazakh, Tatar, Kumyk) And feel free to contact the authors any time!