

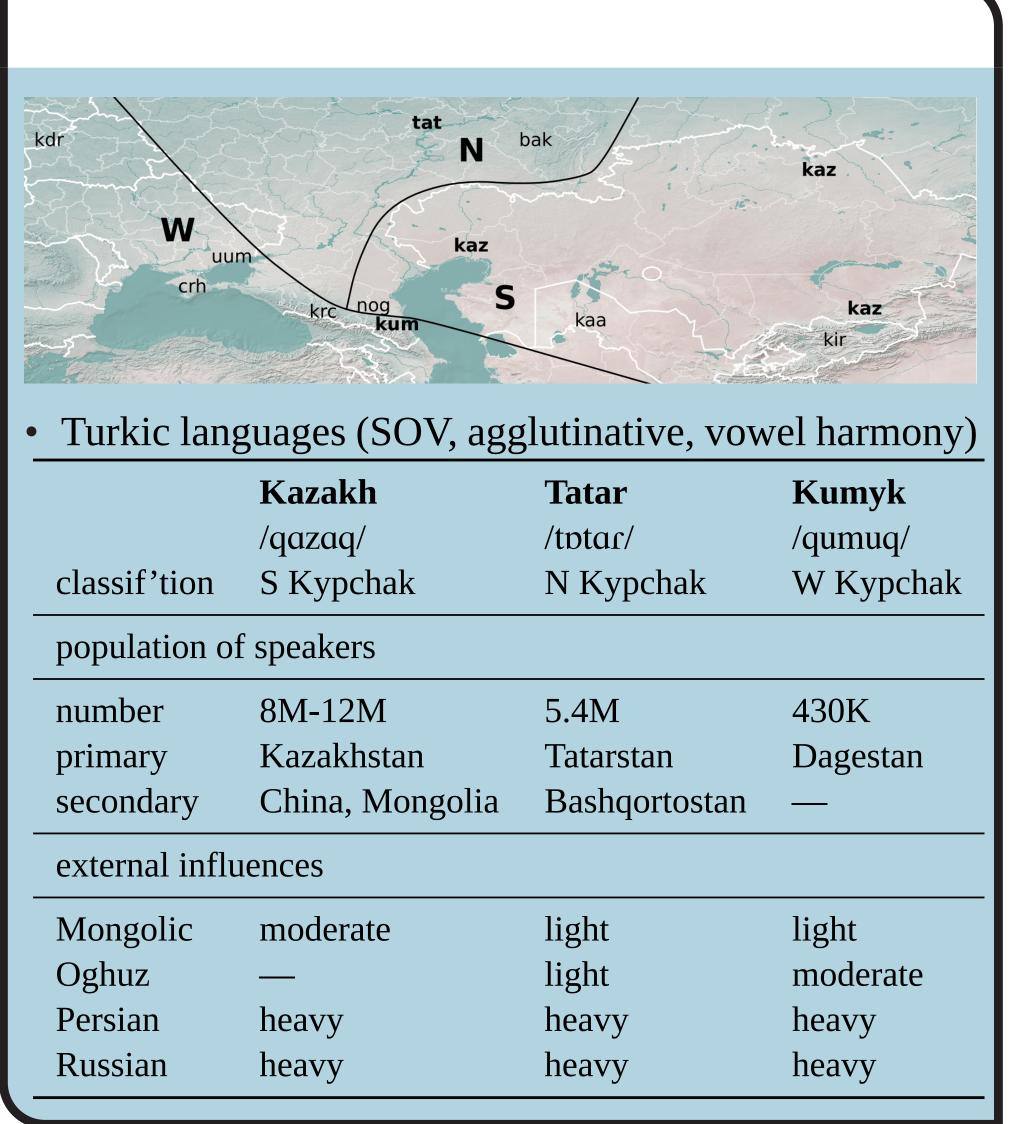
FINITE-STATE MORPHOLOGICAL TRANSDUCERS FOR THREE KYPCHAK LANGUAGES

Jonathan North Washington Ilnar Salimzyanov

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

Francis M. Tyers Special thanks to UiT Norgga Árktalaš Universitehta Áida Sundetova francis.tyers@uit.no A. Sultanmurado A. Sultanmuradov





Morphological transducers

Indiana University

jonwashi@indiana.edu

• • •		Worphological transducers	• • • • • •
• Ta	ake a	surface form, and produce valid lexical	form(s)
• Ta	ake a	lexical form, and produce valid surface	form(s)
'ал,	дым ²	' ↔ ал <v><tv><ifi><pl><sg>, алд<n><px< td=""><td>lsg><nom></nom></td></px<></n></sg></pl></ifi></tv></v>	lsg> <nom></nom>
		Transducers for Turkic languages	
• T	urkis	h (Çöltekin, 2010 & 2014; <mark>Öflazer, 199</mark> 4	4)
• C	rime	an Tatar (Altıntaş, 2001)	
• T	urkm	nen (Tantuğ et al., 2006)	
• K	yrgy	z (Washington et al., 2012)	
• K	azak	h (Бекманова & Махимов, 2013)	
• 01	ır Ka	azakh, Tatar, Kumyk: all GPL (=free an	d open)!
		Framework: HFST	
• R	eimp	lements Xerox FST formalisms (lexc 8	k twol)
• A	lso	provides a wrapper around popular fr	ee/open-
SC	ource	FST toolkits: SFST, OpenFST, and For	ma
		Development effort	
• K	umy	k transducer based on Kazakh, Tatar trai	nsducers
• ~	1 we	ek to reach 80% coverage, +1 week to re	ach 90%
	100		5000
	80		4000
	00		4000
(%)	00		0000
) ef	60		χ Σ
Coverage			Stems
Co	40	_ 	2000
	20		1000
		Coverage (%) ●	
	0	Stems	0

			rs: 0-derivation (zation (e.g., adje	
Type	Gloss	<adj>(<comp>)</comp></adj>	<adj>(<comp>)<subst></subst></comp></adj>	<adj>(<comp>)<advl></advl></comp></adj>
A1	'good'	яхшы (яхшырак)	яхшы (яхшырак)	яхшы (яхшырак)
A2	ʻold'	иске (искерэк)	иске (искерэк)	
A3	'dead'	үле (—)	үле (—)	— (—)
A4	'basic'	төп (—)		

01 02 03 04 05 06 07 08 09 10 11 12 13

Day

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!

• Part of Apertium Turkic project:

Gloss... Құдай Өзінің жаратқандарының бәріне өте жақсы екенін көрді. қарап, Аллаһ Үзе нәрсәләргә аларның бик яхшы икәнен күрде. яраткан карап, Аллагь Оьзю бек яхшы экенин гёрген. къарап, яратгъан затлагъа олар [everything/thing-s]-to looked.at, they/their very good own-his created God being saw. 'God looked at everything he had created and saw that it was very good.' Kazakh (kaz) Kumyk (kum) Tatar (tat) Құдай Өзінің жаратқандарының Аллаh Үзе яраткан нәрсәләргә карап, Аллагь Оьзю яратгъан затлагъа бәріне қарап, өте жақсы екенін көрді. аларның бик яхшы икәнен күрде. къарап, олар бек яхшы экенин гёрген. Құдай<n><nom> Аллаh<n><nom> Аллагь<n><nom> θ3<prn><ref><px3sp><gen> Y3<prn><ref><px3sp><nom> 0ьз<prn><ref><px3sp><nom> жapaт<v><tv><ger past><pl><px3sp><gen> ярат<v><tv><gpr_past> ярат<v><tv><gpr past> 6əpiprn><qnt><px3sp><dat> зат<n><pl><dat> нәрсә<n><pl><dat> қара<v><tv><gna_perf> кара<v><tv><gna perf> къapa<v><tv><gna perf> аларprn><pers><p3><pl><gen> oлapcprn><pers><p3><pl><nom> өте<adv> бик<adv> бек<adv> жақсы<adj> яхшы<adj> яхшы<adj> e<cop><ger past><px3sp><acc> и<cop><ger past><px3sp><acc> ><cop><ger past><px3sp><acc> көр<v><tv><ifi><p3><sq> κγp<v><tv><past><p3><sq> rëp<v><tv><past><p3><sg> <sent> <sent> .<sent> Tagset ... 3rd person poss. Noun Third person <pers> Personal <px3sp> <p3> Verb Plural (Singular/Plural) Comma <gna_perf> Verbal adverb 'Nominative' <sent> Sentence Pronoun <nom> Genitive <past> Past (General) (Perfect) Determiner <gen> <gpr past> Verbal adjective Adjective <ifi> Past <adj> Accusative <acc> Dative (Eyewitness/Recent) (Past) <ger past> Verbal noun (Past)

<adv></adv>	Adverb			Dative	
	Intransit			Quantifier	
<tv></tv>	Transitiv	7 e	<ref></ref>	Reflexive	
					1
			•		
		•		cters	-
• Have	e front- ar	nd back-vow	vel readi	ngs in native words	
	letters	values	exa	mples	
kaz	и, у, ю	/əj, əw, jəv /ej, əw, jev	v/ қию v/ кию	д <mark>а</mark> 'chopping down' д <mark>е</mark> 'getting dressed'	
± a ±		э/С_	дәр	есл ә р 'lessons'	
tat	e	/j/+ы /j/+э		ap 'years' гл <mark>ә</mark> р 'boys'	
		/ø, y/ / C _	ГЮН	лер 'days'	
	••	·	гёзд	i <mark>e</mark> p 'eyes' кл <mark>е</mark> р 'hearts'	
kum	ë, ю	/jø, jy/	ёнк	юл <mark>е</mark> р 'darlings'	
		/jo, ju/	ЮЛД	узл <mark>ар 'stars'</mark>	
			CJIJI	ap 'roads'	
solut	tion: hairy	twol rules	for majo	ority of cases	
unac	counted-f	for words m	arked ha	armony-forcing char	
adjus	st rules fo	r harmony-	forcing o	characters	
		Loan	words		
				s in native words may	V
		vowels in		·	
				n word example	
	елдің 'со			аевтың 'Nazarbayev's'	
tat	галимнат	'scientists'	артист	л <mark>ар 'artists'</mark>	
	сёзлер 'м			тл <mark>ар 'airplanes'</mark>	
• solut	tion: sepa	rate continu	ation lex	kicon (messy rules)	
LEX]	CON N1-	RUS			
:%{;	%} N1 ;				
		nc			
	CON Nou			/ o w + i o + //	
		CT N1-RUS			
I dill	им. галим	N1 ; !	SCIEU	. T 2 C	
	А	cronvms a	nd nu	merals	
				spelled-out words	
		n thirty', бес			
	•	,	•	•	
_				e in numerals (etc.)	
	•	30', 5- те н ';			
• solut	tion: phor	ology-trigg	ering ch	aracters	
1 - 10	(,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,	O.) NIIIM DTA	CTT .	" - 0 0 - "	

kum d	ё, ю /jø, jy/	юреклер 'hearts' ёнкюлер 'darlings' юлдузлар 'stars' ёллар 'roads'
	/jo, ju/	ёллар 'roads'
 solution 	n: hairy twol rules f	for majority of cases
		arked harmony-forcing char
b	rules for harmony-f	
		words
	ent back vowels in F	vowels in native words may
		Russian word example
kaz e tat Γ	алимн <mark>әр 'scientists'</mark>	Назарбаевт <mark>ы</mark> ң ' <i>Nazarbayev's</i> '
• solutio	n: separate continua	ation lexicon (messy rules)
	ON N1-RUS } N1 ;	
артис	ON Nouns т:артист N1-RUS :галим N1 ; ! "	
twol г отыздаno pho 30-дан	ules handle phonolo ан 'from thirty', бес	vailable in numerals (etc.) from 5'
5:5%{	9%}%{c%} NUM-DIG 9%}%{c%} NUM-DIG %0%{a%}%{3%} NUM	
	A resulting m	essy twol rule
"A front %{A%}:e [[%{3%}	[[\[.#. :Vow]] [:RdYotVow й:0 :RdYotV :0 %{γ%}:0] :Cns*]/[[:0 except [:RdYotVow :Cns* %{&% [:Cr [:Vow - :RdYotVow	\${0%}; [:Vow:ь]]:Cns:Cns*]/:0_; :RdVow:ь]:Cns:Cns*]/:0_; :RdYotVow]:Cns:Cns*]/:0_; /ow:Cns:Cns*]/[:0-й:0]_; - AbstractVow:] %-:]*_; s}:0:Cns*]/[:0-%{A%}:0]_; ns:p%{A%}:%>::Cns*]/:0_; v]:RdYotVow:Cns:Cns*]/:0_; ow]/[[:0-й:0] %>:]_;

raitu	of speech		mber of st	
		Kazakh	Tatar	Kumyk
Noun		2640	2795	2568
Verb		1470	1143	386
Adjec	tive	_754	816	219
	r noun	5701	5361	1443
Adver	'b	171	177	63
Nume		63	63	44
	nction	46	45	13
	osition	50	43	12
Prono		32	28	17
Deterr	niner	39	34	<u>C</u>
Total:		11224	10737	4845
		. Test corp	ora	
Wiki	ipedia N	lews	Religion	1
	pedia az	zattyk.org	Quran +	
at Wiki um —	u ll a,	nt.tatar-inform.ru oldash.etnosmi.r		New Testan + New Testa
Naïve co	verage -	percentage c	of surface f	orms in a
	•	≥ 1 analysis		
	n - of a f	orm's analyse alyses provid	es, % corr	
Precision Recall - correct fo	n - of a f % of and or a form	orm's analys alyses provid , by comparii	es, % correled by tran	sducer the
Precision Recall - correct fo	n - of a f % of and or a form	orm's analys alyses provid , by comparid valuation	es, % correled by tran ng against esults	sducer the
Precision Recall - correct fo	n - of a f % of and or a form E Corpus	orm's analys alyses provid , by comparia valuation a Tokens	es, % correled by transt against Coverage	a gold sta
Precision Recall - correct fo	n - of a f % of and or a formE Corpus Wikiped	orm's analysalyses providents, by comparing the second sec	es, % correled by transtogen againstogen Coverage 85.61 ±	a gold start (%) An
Precision Recall - correct fo	n - of a f % of and or a formE Corpus Wikiped News	orm's analys alyses provid , by comparing valuation research Tokens dia 25.6M 3.8M	es, % correled by transfing against Coverage 85.61 ± 92.12 ±	sducer the a gold started (%) And 1.37 2.72 2.72
Precision Recall - correct for	n - of a f % of and or a formE Corpus Wikiped	orm's analys alyses provid , by comparing valuation research Tokens dia 25.6M 3.8M	es, % correled by transfing against Coverage 85.61 ± 92.12 ± 92.49 ±	sducer the a gold star (%) An 1.37 2.72 2.1.66 2.
Precision Recall - correct for	n - of a f % of and or a formE Corpus Wikiped News	orm's analys alyses provided, by comparing the second seco	es, % correled by transfing against Coverage 85.61 ± 92.12 ±	sducer the a gold star (%) An 1.37 2.72 2.1.66 2.
Precision Recall - Correct for	n - of a f % of and or a formE Corpus Wikiped News Religion Average Wikiped	orm's analys alyses provided, by comparing the second seco	es, % correled by transfer against esults Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ±	sducer the a gold started (%) And 1.37 2.72 2.1.66 2.1.91 2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.2.
Precision Recall - correct for Kazakh (r50547)	n - of a f % of and or a formE Corpus Wikiped News Religion Average Wikiped News	orm's analys alyses provided, by comparing the second seco	es, % correled by transfer against esults Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ± 89.75 ±	sducer the a gold star (%) An (2.72 2.72 2.1.66 2.1.91 2.1
Precision Recall - correct for Kazakh (r50547)	n - of a f % of and or a formE Corpus Wikiped News Religion Average Wikiped	orm's analys alyses provided, by comparing the second seco	es, % correled by transfer against esults Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ±	sducer the a gold star (%) An (2.72 2.72 2.1.66 2.1.91 2.2.0.07 2.2.0.07
Precision Recall - correct for Azakh r50547) Tatar	n - of a f % of and or a formE Corpus Wikiped News Religion Average Wikiped News	orm's analys alyses provided, by comparing aluation and aluation aluation and aluation and aluation and aluation and aluation aluation and aluation aluation and aluation aluation and aluation aluatio	es, % correled by transfer against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ±	1.37 2.72 2.1.66 2.17 2.20.07 2.55 2.55
Precision Recall - correct for Azakh r50547) Tatar	n - of a f % of and or a formE Corpus Wikiped News Religion Average Wikiped News Religion	orm's analys alyses provided, by comparing aluation and aluation aluation and aluation and aluation and aluation and aluation aluation and aluation aluation and aluation aluation and aluation aluatio	es, % correled by transit against Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ± 89.75 ± 91.25 ±	sducer the a gold star (%) An (1.37 2.72 2.72 1.66 2.72 2.25 2.25 2.55 2.3 1.60 2.55
Precision Recall - correct for Azakh r50547) Tatar r50260)	n - of a few of and or a form of a form of a form of a form of a few serage. Wikiped News Religion Average. Wikiped News Religion Average. Average.	alyses provided by comparing the state of th	es, % correled by transit against Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ± 89.75 ± 91.25 ± 89.12 ±	sducer the a gold star (%) And (1.37) 2.72 2.1.66 2.1.91 2
Precision Recall - Correct for Cazakh Cr50547) Catar Cr50260) Cumyk	n - of a few and the second of a form	alyses provided, by comparing the second sec	es, % correled by transit against Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ± 89.75 ± 91.25 ± 91.25 ± 91.10 ±	sducer the a gold start and start an
Precision Recall - Correct for Cazakh (r50547) Tatar (r50260) Kumyk (r50300)	n - of a few of and or a form of a f	alyses provided, by comparing the transfer of	es, % correled by transing against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ± 91.25 ± 91.25 ± 91.78 ± 91.78 ±	sducer the a gold star (%) And (1.37) 2.72 2.72 1.66 2.72 2.55 2.55 2.55 2.55 2.55 2.55 2.55
Recall - correct for Kazakh r50547) Catar r50260) Kumyk r50300) Selected &	n - of a few of and or a form of a few of a	alyses provided, by comparing the Tokens Tokens dia 25.6M 3.8M 851K dia 159K 5.2M 382K 286K 227K dique random su	es, % correled by training against Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ± 91.25 ± 91.10 ± 92.47 ± 91.78 ± arface forms form	sducer the a gold star (%) And (1.37) 2.72 2.72 1.66 2.72 2.55 2.75 2.75 2.75 2.75 2.75 2.75
Recall - correct for Kazakh r50547) Catar r50260) Kumyk r50300) Selected &	n - of a few of and or a form of a few of a	alyses provided, by comparing the Tokens Tokens dia 25.6M 3.8M 851K dia 159K 5.2M 382K 286K 227K dique random su	es, % correled by transing against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ± 91.25 ± 91.25 ± 91.78 ± 91.78 ±	sducer the a gold start and start an
Precision Recall - Correct for Correct for Cazakh r50547) Catar r50260) Kumyk r50300) Selected & Languag Kazakh	n - of a few of and or a form of a form of a form of a form of a few of a f	orm's analys alyses provided, by comparing a luation of the second state of the second	es, % correled by transing against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ± 91.25 ± 91.25 ± 91.78 ± 91.78 ± erface forms for (%) 98.61	sducer the a gold star (%) An (1.37 2.72 2.72 1.66 2.72 2.55 2.55 2.66 1.03 1.03 1.03 1.03 1.03 1.03 1.03 1.03
Precision Recall - Correct for Correct for Cazakh r50547) Catar r50260) Cumyk r50300) Selected & Languag Kazakh	n - of a few of and or a form of a form of a form of a form of a few of a f	lorm's analysed lyses provided, by comparing the second se	es, % correled by transing against cesults Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 91.25 ± 91.25 ± 91.25 ± 91.78 ± 91.78 ± erface forms for	sducer the a gold star (%) An (1.37 2.72 2.72 1.66 2.72 2.55 2.55 2.66 1.03 1.03 1.03 1.03 1.03 1.03 1.03 1.03
Precision Recall - Correct for Cazakh r50547) Catar r50260) Cumyk r50300) Selected & Languag Kazakh Tatar	n - of a few of and or a form of a form of a form of a form of a few serage of the ser	orm's analys alyses provided, by comparing a luation of the second state of the second	es, % correled by transing against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ± 91.25 ± 91.25 ± 91.78 ± 91.78 ± erface forms for (%) 98.61	1.37 2 2 2 1.66 2 1.91 2 2 2.55 2 1.60 2 1 1.03 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Precision Recall - Correct for Cazakh (r50547) Fatar (r50260) Kumyk (r50300)	n - of a few of and or a form of a form of a form of a form of a few serage of the ser	dia 25.6M 3.8M 851K 159K 5.2M 382K 1 286K 227K 1 2000 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	es, % correled by training against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ± 91.25 ± 91.10 ± 92.47 ± 91.78 ± arface forms fo	1.37 2.72 2.72 1.66 2.17 2.55 2.160 2.55 2.160 2.17 0.94 1.16 1.03 1.16 1.16 1.03 1.16 1.16 1.16 1.16 1.16 1.16 1.16 1.1
Precision Recall - Correct for Kazakh (r50547) Fatar (r50260) Kumyk (r50300) Selected & Languag Kazakh Tatar	n - of a few of and or a form of a form of a form of a form of a few serage of the ser	dia 25.6M 3.8M 851K 159K 5.2M 382K 1 286K 227K 1 2000 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	es, % correled by training against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ± 89.75 ± 91.25 ± 91.10 ± 92.47 ± 91.78 ± arface forms fo	sducer the a gold star (%) And (1.37) 2.72 2.72 1.66 2.72 2.55 2.75 2.75 2.75 2.75 2.75 2.75

................Number of stems......

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