

FINITE-STATE MORPHOLOGICAL TRANSDUCERS FOR THREE KYPCHAK LANGUAGES

Jonathan North Washington Ilnar Salimzyanov

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

Francis M. Tyers

Special thanks to UiT Norgga Árktalaš Universitehta Aida Sundetova francis.tyers@uit.no A. Sultanmuradov francis tyers@uit.no



Kypchak languages



Indiana University

jonwashi@indiana.edu

Turkic languages (SOV, agglutinative, vowel harmony)

classif'tion	Kazakh /qazaq/ S Kypchak	Tatar /tɒtɑɾ/ N Kypchak	Kumyk /qumuq/ W Kypchak
population of	f speakers		
number primary	8M-12M Kazakhstan	5.4M Tatarstan	430K Dagestan
secondary	China, Mongolia	Bashqortostan	—
external influ	iences		
Mongolic	moderate	light	light
Oghuz		light	moderate
Persian	heavy	heavy	heavy
		5	

Morphological transducers

Morphological transducers

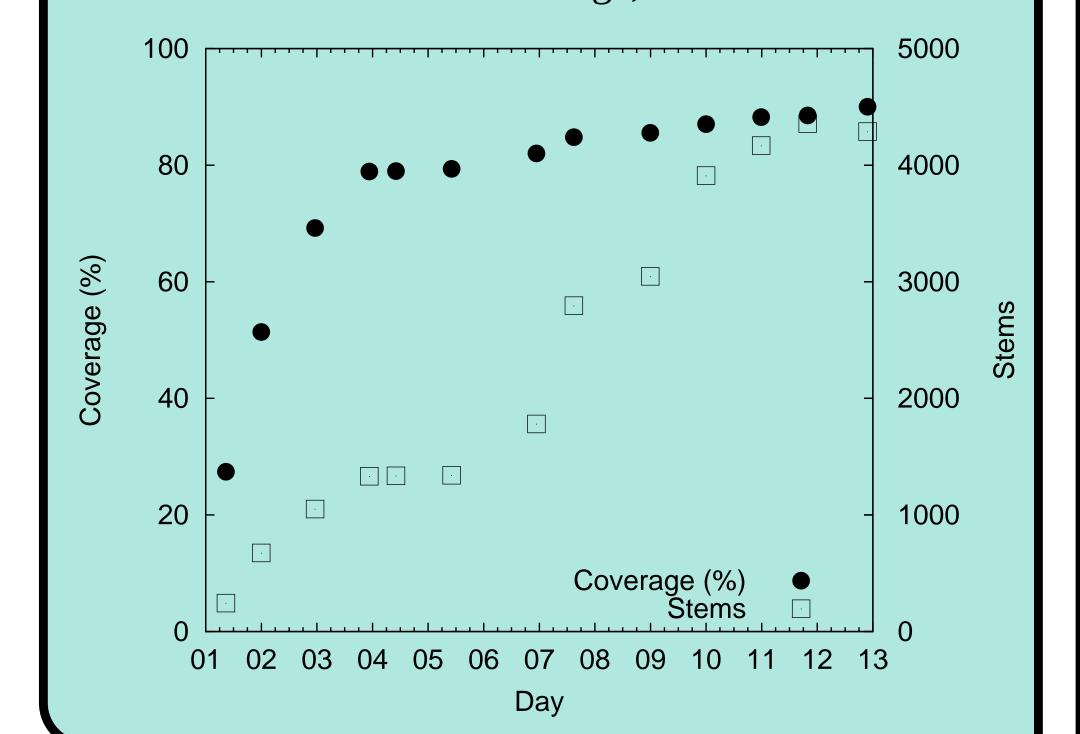
- Take a surface form, and produce valid lexical form(s)
- Take a lexical form, and produce valid surface form(s) 'алдым' ↔ ал<v><tv><ifi><p1><sg>, алд<n><px1sg><nom> Transducers for Turkic languages.....
- Turkish (Çöltekin, 2010 & 2014; Öflazer, 1994)
- Crimean Tatar (Altıntaş, 2001)
- Turkmen (Tantuğ 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/opensource FST toolkits: SFST, OpenFST, and Foma Development effort......

Kumyk transducer based on Kazakh, Tatar transducers

• \sim 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)

	uppre	dem eategon	zation (e.g., aaje	ett ves, below,
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 (искерәк)	иске (искерэк)	 ()
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. Құдай Өзінің жаратқандарының бәріне өте жақсы екенін көрді. қарап, Аллаh Yзе бик яхшы икәнен күрде. яраткан аларның нәрсәләргә карап, Аллагь Оьзю яратгъан бек яхшы экенин гёрген. къарап, олар затлагъа own-his created [everything/thing-s]-to looked.at, they/their very good God being

'God looked at everything he had created and saw that it was very good.'

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.	Аллагь Оьзю яратгъан затлагъа къарап, олар бек яхшы экенин гёрген		
<pre>Kyдaй<n><nom> 03<prn><ref><px3sp><gen> жарат<v><tv><ger_past><pl><ppx3sp><gen> бәрі<prn><qnt><px3sp><dat> кара<v><tv><gna_perf> ,<cm> — 0те<adv> жақсы<adj> e<cop><ger_past><px3sp><acc> көр<v><tv><ifi><p3><sent> .<sent></sent></sent></p3></ifi></tv></v></acc></px3sp></ger_past></cop></adj></adv></cm></gna_perf></tv></v></dat></px3sp></qnt></prn></gen></ppx3sp></pl></ger_past></tv></v></gen></px3sp></ref></prn></nom></n></pre>	Аллаh <n><nom> Y3<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><sent> .<sent></sent></sent></p3></past></tv></v></acc></px3sp></ger_past></cop></adj></adv></gen></pl></p3></pers></prn></cm></gna_perf></tv></v></dat></pl></n></gpr_past></tv></v></nom></px3sp></ref></prn></nom></n>	Аллагь <n><nom> Oьз<prn><ref><px3sp><nom> ярат<v><tv><gpr_past> зат<n><pl><pre> зат<n><pl><pre> хьара<v><tv><gna_perf> ,<cm> олар<pre> олар<pre> prn><pers><p3><pl><nom> бек<adv> яхшы<adj> э<cop><ger_past><px3sp><acc> гёр<v><tv><past><p3><sent> </sent></p3></past></tv></v></acc></px3sp></ger_past></cop></adj></adv></nom></pl></p3></pers></pre></pre></cm></gna_perf></tv></v></pre></pl></n></pre></pl></n></gpr_past></tv></v></nom></px3sp></ref></prn></nom></n>		
	Tagget			

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

Orthography-phonology mapping issues

.. Ambiguous characters.....

Have front- and back-vowel readings in native words

	letters	values	examples
kaz	и, у, ю	/wej, we, jew/ /wej, we, jew/	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'
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 for majority of cases
- 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 tat kum	елдің 'country's' галимнәр 'scientists' сёзлер 'words'	Назарбаевтың 'Nazarbayev's' артистлар 'artists' самолётлар 'airplanes'		
solution: separate continuation lexicon (messy rules				

LEXICON N1-RUS :%{\angle \cdots\} N1; LEXICON Nouns артист:apтист N1-RUS ; ! "artist" галим:галим N1 ; ! "scientist"Acronyms and numerals.....

- twol rules handle phonology for spelled-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%{3%}%{c%} NUM-DIGIT; ! "τθρτ" 5:5%{3%}%{c%} NUM-DIGIT; ! "6ec" 3%0:3%0%{a%}%{3%} NUM-DIGIT ; ! "отыз"

..... A resulting messy twol rule......

RdYotVow = ë ю Ë Ю ; AbstractVow = %{a%} %{э%} %{γ%} %{o%} ;
"A front unrounded harmony" %{A%}:e <=> [[:FrontVow [:Vow :ь]] :Cns :Cns*]/:0 _ ;
except [:RdYotVow :Cns* %{¾%}:0 :Cns*]/[:0 - %{¾%}:0] _ ; [:Cns :p %{¾%}: %>: :Cns*]/:0 _ ; [[:Vow - :RdYotVow] :RdYotVow :Cns :Cns*]/:0 _ ; [:Vow]/[[:0 - й:0] %>:] _ ;

Evaluation

Number of stems.

Part of speech	Nun	nber of st	of stems	
r art or specen	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

kaz Wikipedia azattyk.org Quran + Bible		Wikipedia	News	Religion
	t	at Wikipedia	tat.tatar-inform.ru	Quran + Bible Quran + New Testament 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 News Religion	25.6M 3.8M 851K	85.61 ± 1.37 92.12 ± 2.72 92.49 ± 1.66	2.43 2.88 2.63
(r50547)	Average		90.07 ± 1.91	2.64
Tatar	Wikipedia News Religion	159K 5.2M 382K	$86.35 \pm 2.17 89.75 \pm 0.07 91.25 \pm 2.55$	2.24 2.30 2.24
(r50260)	Average		89.12 ± 1.60	2.26
Kumyk	News Religion	286K 227K	91.10 ± 0.86 92.47 ± 1.03	1.53 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

Future Work

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