







Example output



Jonathan North Washington Indiana University onwashi@indiana.edu

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

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

өтө жакшы экенин көрдү.

DESIGNING FINITE-STATE MORPHOLOGICAL TRANSDUCERS

FOR KYPCHAK LANGUAGES

Francis M. Tyers UiT Norgga Árktalaš Universitehta francis.tyers@uit.no

Special thanks to: Tolgonay Kubatova Aida Sundetova Ağarahim Sultanmuradov





Kypchak languages • Turkic languages (SOV, agglutinative, vowel harmony) Kumyk classification Eastern population of speakers China, etc. China, Mongolia Bashqortostan external influences

Morphological transducers

- Morphological transducers Efficient (in speed & size) models of a language's morphology
- Take a surface form, and produce valid lexical form(s)
- Take a lexical form, and produce valid surface form(s) алдым \leftrightarrow an<v><tv><ifi><p1><sg>, anд<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
- morphotactics implemented in lexc
- morphophonology implemented in twol
- compiled separately; compose-intersected to single transducer алдым \leftrightarrow an>{D}{I}>м \leftrightarrow an<v><tv><ifi><p1><sg алдым ↔ алд>{I}м ↔ алд<n><px1sg><nom>

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

Kyrgyz (kir)

Аллаһ Үзе яраткан нәрсәләргә Аллагь Оьзю яратгъан затлагъа God own-his created

Kazakh (kaz)

Құдай Өзіні

Құдай Өзінің жаратқандарының бәріне

- өте жақсы екенін көрді. аларның бик яхшы икәнен күрде. бек яхшы экенин гёрген. [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.'* (Bible, Genesis 1:31)

	Tatar (tat)
қ жаратқандарының бәріне кақсы екенін көрді.	Аллаһ Үзе яратка аларның бик яхшы
n> <px3sp><gen> <ger_past><pl><px3sp><gen> nt><px3sp><dat> <gna_perf></gna_perf></dat></px3sp></gen></px3sp></pl></ger_past></gen></px3sp>	Аллаh <n><nom> Y3<prn><ref><px3 ярат<v=""><tv><gpr_ нәрсә<n=""><pl><dat кара<v=""><tv><gna_ ,<cm=""> алар<prn><pers></pers></prn></gna_></tv></dat></pl></gpr_></tv></px3></ref></prn></nom></n>

къapa<v><tv><gna_perf> олаpconapconapconapconapconapconap э<cop><ger_past><px3sp><acc> гёр<v><tv><past><p3><sg> e<cop><ger_past><px3sp><acc>

Verbal adverb Noun <gen> Genitive <past> Past (General) <gpr past> Verbal adjective Third person acc> Accusative <ifi> Past (Eyewitness/Recent) Determiner <dat> Dative <px3sp> 3rd person poss. <ger past> Verbal noun <ref> Reflexive <qnt> Quantifier <adj> Adjective (Singular/Plural) <pers> Personal <cm> Comma <adv> Adverb

Morphophonology

көр<v><tv><ifi><p3><sg>

- Desonorisation (kaz & kir)......
- {N} desonorises to д after a consonant алма- $\{N\}\{I\}$ \rightarrow алманы 'apple-ACC' сыр- $\{N\}\{I\}$ → сырды 'secret—ACC'
- $\{L\}$ desonorises to д after cons. of sonority $\leq l$
- сыр- $\{L\}\{A\}$ р → сырлар 'secret-PL' кыз- $\{L\}\{A\}p \rightarrow$ кыздар 'girl–PL'
- "L Desonorisation" %{L%}:д <=> :VoicedLowSonCns %>: __ ;
- "N Desonorisation"
- %{N%}:д <=> :VoicedCns %>: __ ; Epenthesis
- Turn {y} into a harmonised high vowel when a vowel doesn't follow the following consonant: $myp{y}H \rightarrow mypyh 'nose'$ $мур{y}H+{I}M \rightarrow мурдум 'my nose'$
- %{y%}:Vy <=> [:LastVowel :Cns* :Cns]/[:0] __ [:Cns [.#. | :Cns]]/[:0 | %>:] ; where Vy in (иүииүыыууыуу) LastVowel in (иүеэөяаёоыюу) matched;

Other uses

- HFST transducers are trivially converted to **spell checkers**
- Segmenter, e.g. көргөзгөндөрдөнсүңбү: $\kappa \in P^{G}_{A}3>\{G}_{A}H>\{L}_{A}p>\{D}_{A}H>c\{I}_{H}>\{B}_{I}$

Morphotactics

- Morphological & orthographical words......
- өнүктүрөбүзбү? 'will we develop [it]?' өнүк<v><tv><caus><aor><pl><pl>+бы<qst>
- келатсаң 'if you come'
- кел<v><iv><prc impf>+жат<vaux><gna cnd><p2><sg> ... Irregular [noun + possessive + case] forms
- Some combinations of possessive + case morphemes are unpredicted (i.e., not formed simply by concatenation and application of phonology):

case	form	1SG	2SG	3SP
nominative	_	-(I)M	-(I)ң	-(c)I
accusative	-NI	-(І)мдІ	-(I)ңдI	-(с)Ін
genitive	-NIH	-(І)мдІн	-(І)ңдІн	-(с)ІнІн
locative	-DA	-(Ï)мдA	-(Ï)ндA	-(с)ІндА
ablative	-DAн	-(I)мдAн,	-(I)ндAн,	-(c)InAn
		-(I) MA H	-(I) ңА н	` '
dative	-GA	-(I)MA	-(I)́ңА	-(с)ІнА
TNID C 1	• 11	1 /T\ 11	C. 1 /	\ 11 C.

- A,I,N,D,G have various allophones; (I) null after vowels; (c) null after cons.
- underlying <px3sp> form used: {s}{I}{n} • {s} and {n} default to c and н; rules map to null by context
- morphophonology more complicated, morphotactics simplerNoun-noun compounds......
- a N-N compund type: N2 has <px3> and related morphology e.g., аба ырайы<n><loc>: аба ырайында, *аба ырайыда

LEXICON N-INFL-3PX-COMPOUND %<n%>:%>%{S%}%{I%}%{n%} GEN-POS ; LEXICON Nouns аба% ырайы:аба% ырай N-INFL-3PX-COMPOUND

"weather" чакыруу% кагазы:чакыруу% кагаз N-INFL-3PX-COMPOUND ; ! "invitation"

Ambiguous characters

		. Ambiguous	Characters
• Have	e front - an	d back-vowel re	adings 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'
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 get a harmony-forcing character
- adjust rules for harmony-forcing characters
- Letters that represent front vowels in native words may represent "back" vowels in Russian words

JCIIC	buch vovicio ili russi	an words
	native word example	Russian word example
tat	елдің 'country's' галимнәр 'scientists' сёзлер 'words'	Назарбаевтың 'Nazarbayev'з артистлар 'artists' самолётлар 'airplanes'

- solution: separate continuation lexicon (messy rules) LEXICON N1-RUS
- :%{*>*%} N1 ; LEXICON Nouns артист:apтист 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 (incorrect phonological triggers in acronyms)
- 30-дан 'from 30', 5-тен 'from 5'
- solution: phonology-triggering characters • simplified: e.g., {c} for all voiceless ostruents
- 4:4%{9%}%{c%} NUM-DIGIT ; ! "τθρτ" 5:5%{9%}%{c%} NUM-DIGIT ; ! "бес" 3%0:3%0%{a%}%{3%} NUM-DIGIT ; ! "отыз"
- + vowel letters..... • [a o y] become [я ё ю] after й and й deletes
- й incorporated into the context of many rules • additional rules to change the characters and delete original i
- "Deletion of й before yoticised vowels" й:0 <=> __ [:YotVow]/[:0 | %>:] ;

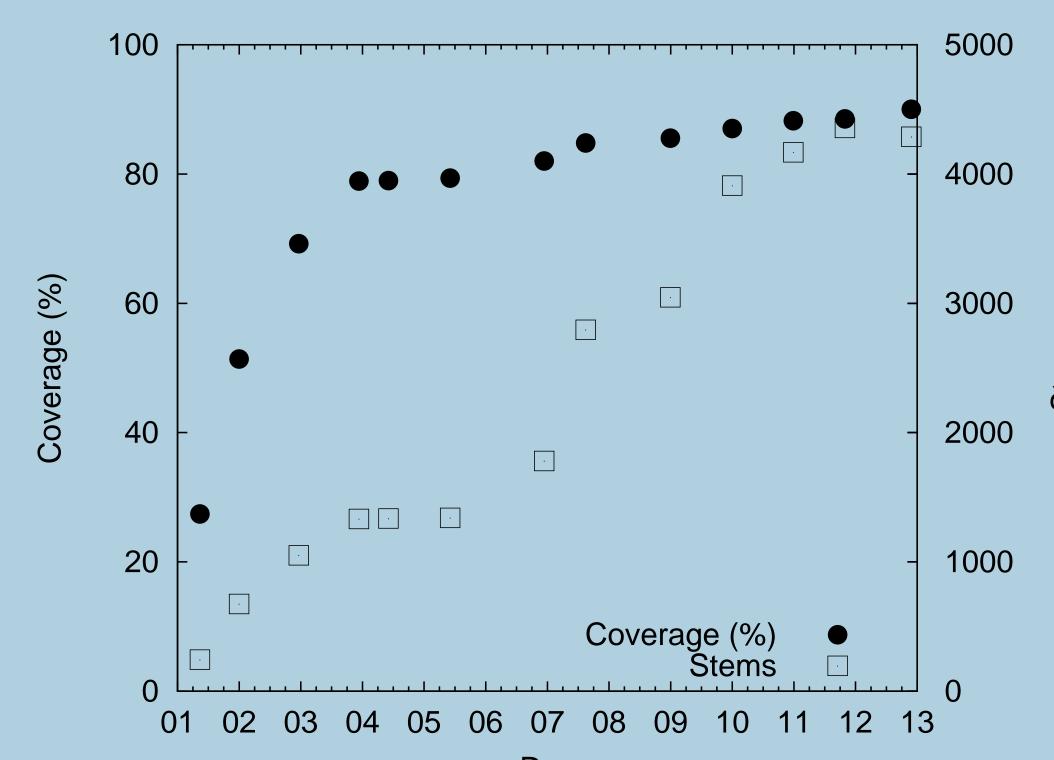
..... A resulting messy twol rule......

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

Development effort

	Kyrgyz	Kazakh	Tatar	Kumyk	
begun	Apr. 2011	Dec. 2010	Dec. 2011	Oct. 2013	
80% cov.	Aug.? 2011	Aug. 2012	Aug. 2012	Oct. 2013	
time	4 months	19 months	7 months	1 week	
• (various periods of intermission, various rewrites)					
Kazakh transducer based on Kyrgyz transducer					
Kyrgyz transducer currently being rewritten based on insights					
gained while writing other Turkic transducers					

- gamed while writing other Turkic transducers • Kumyk transducer based on Kazakh, Tatar transducers: ~1 week to reach 80% coverage, +1 week to reach 90%



Categorisation

- morphologically distinct adjective classes
- most sources claim: adjectives can be used substantively and adverbially
- Other Turkic transducers: 0-derivation (overgenerates) but not all adjectives have all of the following:
- comparative forms, substantive readings, adverbial readings
- Our approach: categorisation if properly categorised, only correct forms are analysed and generated

-	G10 55	(Comp)	-adj/(-comp/)-sabst/	-adj (-comp / -adv c	
A1	ʻgood'	яхшы (яхшырак)	яхшы (яхшырак)	яхшы (яхшырак)	
A2	ʻold'	иске (искерэк)	иске (искерэк)	— (—)	
A 3	'dead'	үле (—)	үле (—)		
A4	'basic'	төп (—)	 ()		
Adverbs					

Certain adverbs have special attributive and ablative forms

- Mostly time adverbs
- Some also have noun readings: regular ablative, other cases:
- быйыл кечээ жана 'this year' 'yesterday' 'just now' <attr> form бүгүнкү быйылкы кечээги жанагы<adv><abl> form бүгүнтөн быйылтан кечээтен жанатан

<n><abl> form бүгүндөн быйылдан — LEXICON ADV-WITH-KI-ABL ADV-KI; ADV-ABL;

Evaluation

Number of stems				
Part of speech	Number of stems			
r ar c or speccir	Kyrgyz	Kazakh	Tatar	Kumyk
Noun	4582	2640	2795	2568
Verb	1193	1470	1143	386
Adjective	1211	754	816	219
Proper noun	5887	5701	5361	1443
Adverb	312	171	177	63
Numeral	66	63	63	44
Conjunction	77	46	45	13
Postposition	50	50	43	12
Pronoun	51	32	28	17
Determiner	64	39	34	9
Total:	13749	11224	10737	4845

	Wikipedia	News	Religion
Kyrgyz Kazakh Tatar Kumyk	Wikipedia Wikipedia Wikipedia —	azattyk.org azattyq.org tat.tatar-inform.ru yoldash.etnosmi.ru	Bible 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 sur-
- face form found in analysed corpus • **Precision** - probability that a provided analysis is valid
- **Recall** probability that a certain valid analysis is among those provided by the transducer

..... Evaluation results

	Corpus	Tokens	Coverage (%)	Amb.
T 7	Wikipedia News	5.3M 4.1M	84.51 ± 2.27 91.43 ± 0.51	3.56 4.19
Kyrgyz	Religion	215K	91.66 ± 1.81	3.99
(r54474)	Average		89.20 ± 3.48	3.91
	Wikipedia	25.6M	85.61 ± 1.37	2.43
Kazakh	News	3.8M	92.12 ± 2.72	2.88
1 tuzum	Religion	851K	92.49 ± 1.66	2.63
(r50547)	Average		90.07 ± 1.91	2.64
	Wikipedia	159K	86.35 ± 2.17	2.24
Tatar	News	5.2M	89.75 ± 0.07	2.30
Tatai	Religion	382K	91.25 ± 2.55	2.24
(r50260)	Average		89.12 ± 1.60	2.26
	News	286K	91.10 ± 0.86	1.53
Kumyk	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 (%)

-38		(/)	(,,
Kyrgyz	200	90.77	69.15
Kazakh	1000	98.61	57.98
Tatar	1000	95.03	85.65
Kumyk	500	96.57	69.1

Ongoing and future work

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