









#### Jonathan North Washington Indiana University jonwashi@indiana.edu

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

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

Obsrn><ref><px3sp><nom>
spat<v><tv><gpr\_past>
sat<n><pl><dat>
къара<v><tv><gna\_perf>

олаpconapconapconapconapconapconap

э<cop><ger\_past><px3sp><acc> гёр<v><tv><past><p3><sg>

<gna perf> Verbal adverb

<gpr past> Verbal adjective

өтө жакшы экенин көрдү.

өте жақсы екенін көрді.

бек яхшы экенин гёрген.

Kumyk (kum)

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





# Kvpchak languages

Kypchak languages	Exa	ample	output	
• Turkic languages (SOV, agglutinative, vowel harmony)	(1)		Өзінің Үзе	9

	<u> </u>	, 00		
	Kyrgyz	Kazakh	Tatar	Kumyk
	\dmurmz\	/qazaq/	/totar/	/qumuq/
classification	Eastern	Southern	Northern	Western
population of s	peakers			
number	3M	8M-12M	5.4M	430K
primary	Kyrgyzstan	Kazakhstan	Tatarstan	Dagestan
secondary	China, etc.	China, Mongolia	Bashqortostan	_

external influ	iences			
Mongolic	moderate	moderate	light	light
Oghuz	<u>—</u>	<del></del>	light	moder
Persian	heavy	heavy	heavy	heavy

## 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  $\leftrightarrow$  aл>{D}{I}>м  $\leftrightarrow$  aл<v><tv><ifi><p1><sg> алдым  $\leftrightarrow$  алд $>{I}м <math>\leftrightarrow$  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!

### Құдай Өзінің жаратқандарының бәріне Аллаһ Үзе

<adv> Adverb

Morphophonology

"L Desonorisation"

"N Desonorisation"

• {N} desonorises to д after a consonant

алма- $\{N\}\{I\}$   $\rightarrow$  алманы 'apple-ACC'

 $\{L\}$  desonorises to д after cons. of sonority  $\leq l$ 

%{L%}:д <=> :VoicedLowSonCns %>: \_\_ ;

сыр- $\{N\}\{I\}$  → сырды 'secret—ACC'

сыр- $\{L\}\{A\}$ р → сырлар 'secret-PL'

%{N%}:д <=> :VoicedCns %>: \_\_ ;

 $myp{y}H+{I}M \rightarrow mypдym 'my nose'$ 

<del>- matched ; -</del>

follow the following consonant:

 $мур{y}н \rightarrow мурун 'nose'$ 

Other uses

кыз- $\{L\}\{A\}p \rightarrow$  кыздар 'girl–PL'

Аллагь Оьзю яратгъан God own-his created

нәрсәләргә затлагъа

аларның бик яхшы икәнен күрде. [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) 

Kazakh (kaz) Kyrgyz (kir) Құдай Өзінің жаратқандарының бәріне қарап, өте жақсы екенін көрді. карап, өтө жакшы экенин көрдү θ3<prn><ref><px3sp><nom> θ3<prn><ref><px3sp><gen> қара<v><tv><gna\_perf> κapa<v><tv><gna\_perf>

жаратканынын

e<cop><ger\_past><px3sp><acc> көр<v><tv><ifi><p3><sg>

..... Desonorisation (kaz & kir)......

..... Epenthesis .....

Turn {y} into a harmonised high vowel when a vowel doesn't

[ :Cns [ .#. | :Cns ] ]/[ :0 | %>:] ;

Noun <gen> Genitive Third person Determiner <dat> Dative <ref> Reflexive <adj> Adjective

acc> Accusative <qnt> Quantifier

<past> Past (General) (Eyewitness/Recent) <px3sp> 3rd person poss. <ger past> Verbal noun (Singular/Plural) <pers> Personal <cm> Comma

## Morphotactics

Аллаһ Үзе яраткан нәрсәләргә ка аларның бик яхшы икәнен күрде.

аларprn><pers><p3><pl><gen>

Y3<prn><ref><px3sp><nom

кара<v><tv><gna\_perf>

- ...... Morphological & orthographical words......
- өнүк<v><tv><caus><aor><pl><pl>+бы<qst>
- келатсаң 'if you come'
- ... 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 accusative genitive locative ablative	— -NI -NIн -DA -DAн	-(I)м -(I)мдI -(I)мдIн -(I)мдА -(I)мдАн, -(I)мАн	-(I)ң -(I)ңдI -(I)ңдIн -(I)ңдА -(I)ндАн,	-(c)I -(c)Iн -(c)ІнІн -(c)ІндА -(c)ІнАн
dative	-GA	-(I)mA	-(I)ңАн -(I)ңА	-(с)ІнА

- underlying <px3sp> form used: {s}{I}{n}
- morphophonology more complicated, morphotactics simpler
- ..... Noun-noun compounds ......

• Segmenter, e.g. көргөзгөндөрдөнсүңбү:

%{y%}:Vy <=> [ :LastVowel :Cns\* :Cns ]/[:0] \_\_

where Vy in (иүииүыыууыуу)

LastVowel in (иүеэөяаёоыюу)

 $\kappa \in P^{G}_{A}3>\{G}_{A}H>\{L}_{A}p>\{D}_{A}H>c\{I}_{H}>\{B}_{I}$ 

HFST transducers are trivially converted to spell checkers

- өнүктүрөбүзбү? 'will we develop [it]?'
- кел<v><iv><prc impf>+жат<vaux><gna cnd><p2><sg>

case	form	18G	2 <b>S</b> G	3SP
nominative	<u>—</u>	-(I)M	-(І)ң	-(c)I
accusative	-NI	-(І)мдІ	-(Ï)ңдI	-(c)IH
genitive	-NIH	-(І)мдІн	-(I)ндIн	-(с)ІнІн
locative	-DA	-(І)мдА	-(I)ңдA	-(с)ІндА
ablative	-DAн	-(І)мдАн,	-(I)ндAн,	-(с)ІнАн
		-(І)мАн	-(І)ңАн	
dative	-GA	-( <b>I</b> ) <b>MA</b>	-(І)ңА	-(c)IHA
INDG have va	arious allor	hones: (I) null	after vowels: (d	r) null after cor

- {s} and {n} default to c and н; rules map to null by context
- a N-N compund type: N2 has <px3> and related morphology e.g., аба ырайы<n><loc>: аба ырайында, \*аба ырайыда

чакыруу% кагазы:чакыруу% кагаз N-INFL-3PX-COMPOUND

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

; ! "invitation"

## Orthography-phonology mapping issues

..... Ambiguous characters ......... Have front- and back-vowel readings in native words

	letters	values	examples
kaz	и, у, ю	/wej, we, jev/ /wej, we, je/	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	ë, 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

CIIC	buch vowels in russi	an words
	native word example	Russian word example
az at um	елдің 'country's' галимнәр 'scientists' сёзлер 'words'	Назарбаевтың 'Nazarbayev's 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

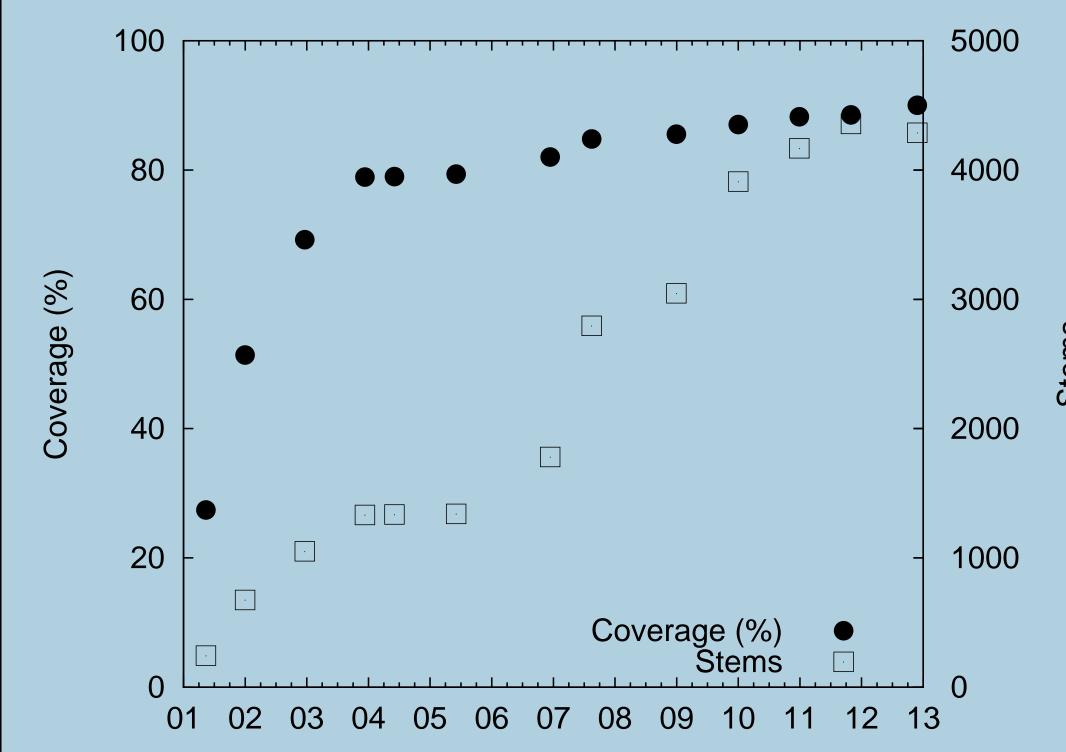
"Deletion of й before yoticised vowels" й:0 <=> \_\_ [ :YotVow ]/[ :0 | %>: ] ; ..... A resulting messy twol rule....

RdYotVow = ë ω Ë Ю ; AbstractVow = %{a%} %{э%} %{γ%} %{o%} ; [ [:FrontVow | [:Vow:ь]]:Cns:Cns\*]/:0 \_ [ [:RdVow:ь]:Cns:Cns\*]/:0 \_ [ [[[\[.#.|:Vow]]:RdYotVow]:Cns:Cns\*]/:0 \_ [[%{3%}:0|%{γ%}:0] :Cns\*]/[[ :0 - AbstractVow: ] | %-: ]\* \_ ; except [ :RdYotVow :Cns\* %{¾%}:0 :Cns\* ]/[ :0 - %{¾%}:0 ] [ :Cns :p %{2%}: %>: :Cns\* ]/:0 [ :Vow - :RdYotVow ] :RdYotVow :Cns :Cns\* ]/:0 [ :Vow ]/[ [ :0 - й:0 ] | %>: ]

## Development effort

Kyrgyz	Kazakh	Tatar	Kumyk
Apr. 2011 Aug.? 2011			Oct. 2013 Oct. 2013
4 months	19 months	7 months	1 week
	Apr. 2011 Aug.? 2011	Apr. 2011 Dec. 2010 Aug.? 2011 Aug. 2012	Apr. 2011 Dec. 2010 Dec. 2011 Aug.? 2011 Aug. 2012 Aug. 2012

- (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
- 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

	<b>G10</b> 55	taaj (teemp )	taajs ( teomps ) toabbes	day ( deomps ) day es		
A1 A2	ʻgood' ʻold'	яхшы (яхшырак) иске (искерәк)	яхшы (яхшырак) иске (искерәк)	яхшы (яхшырак)		
<b>A3</b>	'dead' 'basic'	уле (—)	үле (—)			
A4	'basic'	төп (—)	<u> </u>	<u> </u>		

- Certain adverbs have special attributive and ablative forms
- Mostly time adverbs

ADV-ABL;

 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;

## Evaluation

Number of stems					
Part of speech	Number of stems				
	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	

Test corpora			
	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  $\geq 1$  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.
Kyrgyz	Wikipedia News Religion	5.3M 4.1M 215K	$84.51 \pm 2.27$ $91.43 \pm 0.51$ $91.66 \pm 1.81$	3.56 4.19 3.99
(r54474)	Average		$89.20 \pm 3.48$	3.91
Kazakh	Wikipedia News Religion	25.6M 3.8M 851K	$85.61 \pm 1.37$ $92.12 \pm 2.72$ $92.49 \pm 1.66$	2.43 2.88 2.63
(r50547)	Average		$90.07 \pm 1.91$	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		$89.12 \pm 1.60$	2.26
Kumyk	News Religion	286K 227K	$91.10 \pm 0.86$ $92.47 \pm 1.03$	1.53 1.53
(r50300)	Average		$91.78 \pm 0.94$	1.53

### selected & proofed unique random surface forms from news corpora Language Forms Precision (%) Recall (%)

		( ,	(
Kyrgyz	200	90.77	69.
Kazakh	1000	98.61	57.
Tatar	1000	95.03	85.
Kumyk	500	96.57	69.

## 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.