







DESIGNING FINITE-STATE MORPHOLOGICAL TRANSDUCERS

Jonathan North Washington Indiana University jonwashi@indiana.edu

Ilnar Salimzyanov Казан (Идел буе) федераль университеть ilnar.salimzyan@gmail.com

FOR KYPCHAK LANGUAGES

francis.tyers@uit.no

Francis M. Tyers UiT Norgga Árktalaš Universitehta

kum сёзлер 'words'

LEXICON N1-RUS

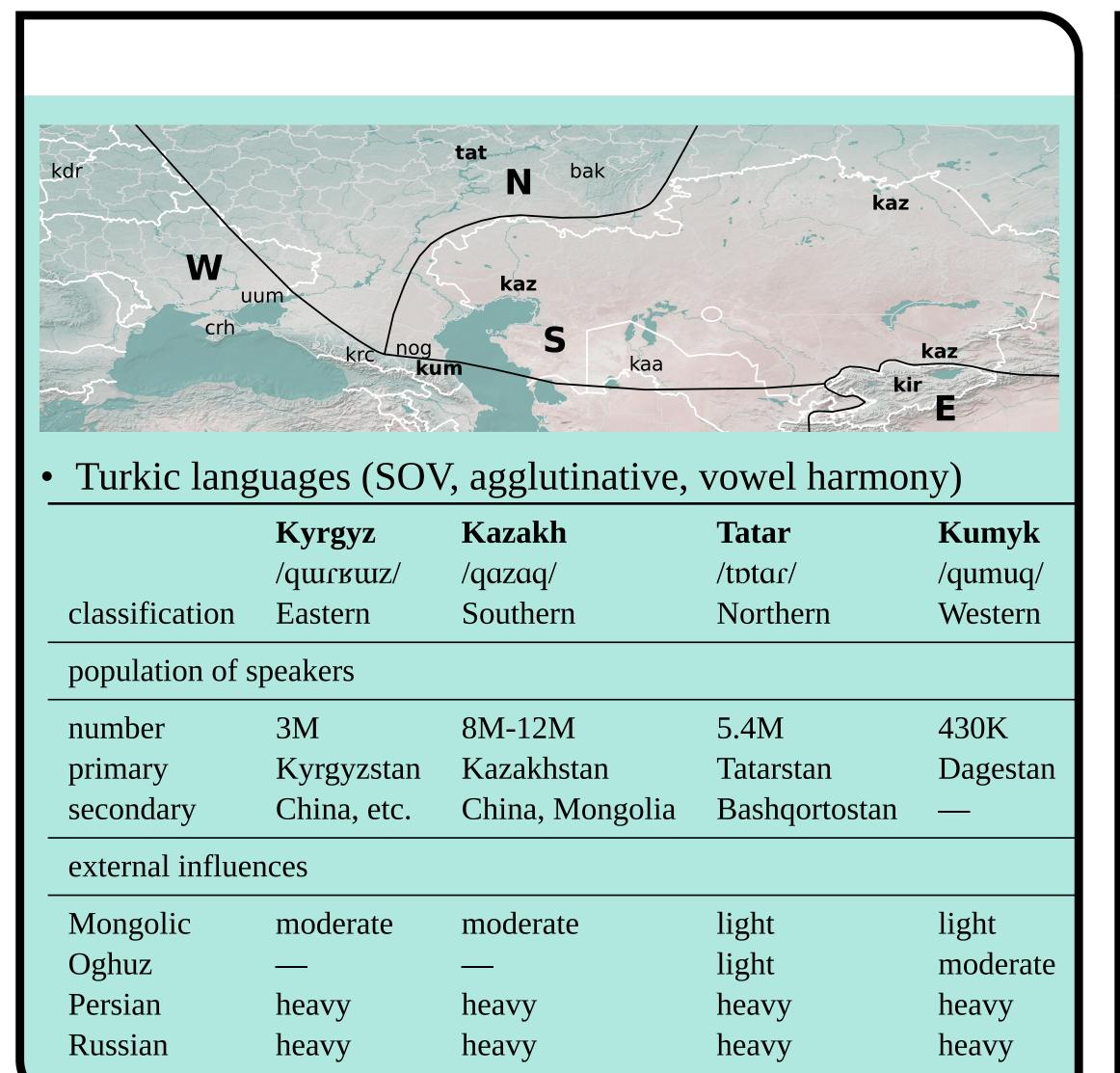
:%{\angle \chi_8\} N1 ;

LEXICON Nouns

Special thanks to: Tolgonay Kubatova Aida Sundetova Ağarahim Sultanmuradov







..... 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' ↔ 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
- 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!

						Glo	SS						
(1)	Кудай	Өзү	жаратка	НЫНЫН	баарына		карап,		өтө	жакшы	экенин	көрдү.	
	Құдай	Өзінің	жаратқа	ндарының	бәріне		қарап,		өте	жақсы	екенін	көрді.	
	Аллаһ	Yзе	яраткан	-	нәрсәләргә		карап,	аларны	ң бик	яхшы	икәнен	күрде.	
	Аллагь	Овзю	яратгъан	ł	затлагъа		къарап,	олар	бек	яхшы	экенин	гёрген.	
	God	own-his	created		[everything/t	hing-s]-to	looked.at,	they/the	eir very	good	being	saw.	
	'God loo	oked at ev	verything i	he had crea	ited and saw t				J		S		
								`		ŕ			
Kyrgy	yz (kir)			Kazakh (kaz)			Tatar (tat)				k (kum)		
_	-	тканынын ба шы экенин к		the contract of the contract o	жаратқандарының ақсы екенін көрді		Аллаһ Үзе яра аларның бик я					гъан затлагъа яхшы экенин гё	рген.
Өз<рг жарат баары кара< , <cm> — өтө<а</cm>	<pre><v><tv><ge i<pre="">c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<pre>c<</pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></ge></tv></v></pre>	x3sp> <nom> er_past><px ><px3sp><da a_perf></da </px3sp></px </nom>	3sp> <gen> t></gen>		<px3sp><gen> <ger_past><pl> <><px3sp><dat></dat></px3sp></pl></ger_past></gen></px3sp>	(3sp> <gen></gen>	Aллah <n><nom> Yз<prn><ref>< ярат<v><tv><g нәрсә<n=""><pl>< кара<v><tv><g ,<cm=""> алар<prn><per бик<adv=""> яхшы<adj></adj></per></prn></g></tv></v></pl></g></tv></v></ref></prn></nom></n>	<pre>px3sp><nom pr_past=""> dat> na_perf></nom></pre>		0ьз<ри ярат< зат <n> къара< ,<cm></cm></n>	dv>	past>	
э<сор	> <ger_pas v><tv><ifix< th=""><th>t><px3sp><a ><p3><sg></sg></p3></a </px3sp></th><th>CC></th><th></th><th>ast><px3sp><acc> i><p3><sg></sg></p3></acc></px3sp></th><th></th><th>и<cop><ger_pa күр<v><tv><pa .<sent></sent></pa </tv></v></ger_pa </cop></th><th></th><th></th><th>э<сор></th><th>><ger_past> ><tv><past></past></tv></ger_past></th><th><px3sp><acc> <p3><sg></sg></p3></acc></px3sp></th><th></th></ifix<></tv></ger_pas 	t> <px3sp><a ><p3><sg></sg></p3></a </px3sp>	CC>		ast> <px3sp><acc> i><p3><sg></sg></p3></acc></px3sp>		и <cop><ger_pa күр<v><tv><pa .<sent></sent></pa </tv></v></ger_pa </cop>			э<сор>	> <ger_past> ><tv><past></past></tv></ger_past>	<px3sp><acc> <p3><sg></sg></p3></acc></px3sp>	
						Tag	set						
<n></n>	Noun Verb	um.	<iv><</iv>	Intransitive Transitive	<gen></gen>	Genitiv	7e <	past> P	entence Past (Ger	neral)	gna_perf	(Perfect)	
<pre><pre><det></det></pre></pre>	_		<p3></p3>	Third pers Plural	son <acc> <dat></dat></acc>	_	ilive <		ast (Evewit	ness/Rec	gpr_past ent)	Verbal adj (Past)	ecuve
<adj></adj>	A 1.		<ref></ref>	Reflexive			fier <		\ <u>\</u>	n poss. <			un
<adv></adv>	A 1	_		Personal	<cm></cm>	Comma				ar/Plural)		(Past)	

..... 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 μ after cons. of sonority $\leq \ln m$ 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 %>: __ ;
- Turn {y} into a harmonised high vowel when a vowel doesn't follow the following consonant:
- $мур{y}н \rightarrow мурун 'nose'$
- $myp{y}H+{I}M \rightarrow mypдym 'my nose'$

```
%{y%}:Vy <=> [ :LastVowel :Cns* :Cns ]/[:0] __
            [ :Cns [ .#. | :Cns ] ]/[ :0 | %>:] ;
where Vy in (иүииүыыууыуу)
LastVowel in (иүеэөяаёоыюу)
         matched ;
```

- ӨНҮК<v><tv><caus><aor><pl><pl><pl>+бы<qst>
- келатсаң 'if you come'
- Some combinations of possessive and case morphemes are

case	form	1SG	2SG	3SP
nom	_	-(I)M	-(І)ң	-(S)I
acc	-NI	-(І)мдІ	-(I)ңдI	-(S)Iн
gen	-NIH	-(І)мдІн	-(І)ңдІн	-(S)IнIн
loc	-DA	-(I)мдA	-(I)ндA	-(S) І нд А
abl	-DAн	-(I)мдAн,	-(I)ндАн,	-(S)IHAH
		-(І)мАн	-(І)ңАн	
dat	-GA	-(I)MA	-(I)ңA	-(S) І н А

- phonological rules delete {n}, {S} by context
- one type of N-N compunds: N2 has <px3> and related morphology

LEXICON N-INFL-3PX-COMPOUND

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

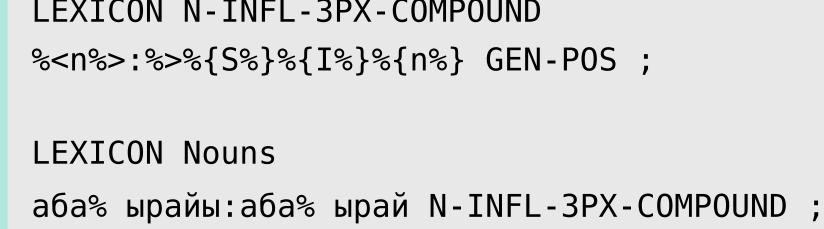
......Morphological & orthographical words......

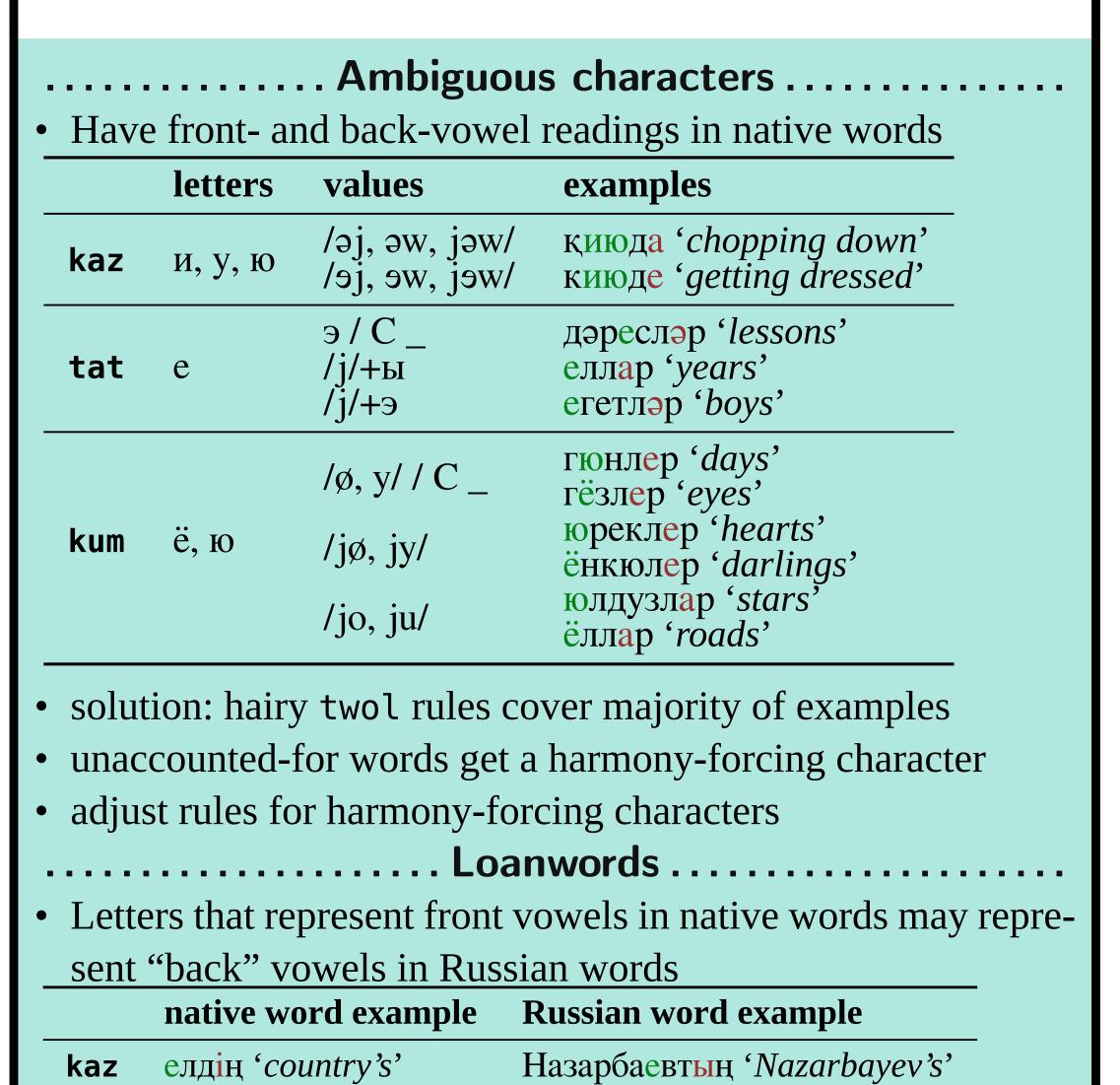
- өнүктүрөбүзбү ? 'will we develop [it]?'
- кел<v><iv><prc impf>+жат<vaux><gna_cnd><p2><sg>
- Irregular [noun + possessive + case] forms
- non-canonical (i.e., not formed simply by concatenation):

case	form	1SG	2SG	3SP
nom	— -NI	-(I)M	-(I)H	-(S)I
acc gen loc	-NIH	-(I)мдI -(I)мдIн	-(I)ңдI -(I)ңдIн	-(S)Ін -(S)ІнІн
loc abl	-DA -DAн	-(I)мдА -(I)мдАн,	-(I)ңдА -(I)ндАн,	-(S)ІндА -(S)ІнАн
		-(I)MAH	-(І)ңАн	
dat	-GA	-(I)MA	-(I)ңA	-(S)IHA

morphophon. complicateder, morphotactics simpler underlying form used: {S}{I}{n}

-Noun-noun compounds...........





самолётлар 'airplanes'

solution: separate continuation lexicon (messy rules)

..... Acronyms and numerals

артист:артист N1-RUS ; ! "artist"

twol rules handle phonology for spelt-out words

no phonological triggers available in numerals (etc.)

галим:галим N1 ; ! "scientist"

отыздан 'from thirty', бестен 'from five'

solution: phonology-triggering characters

4:4%{9%}%{c%} NUM-DIGIT ; ! "τθρτ"

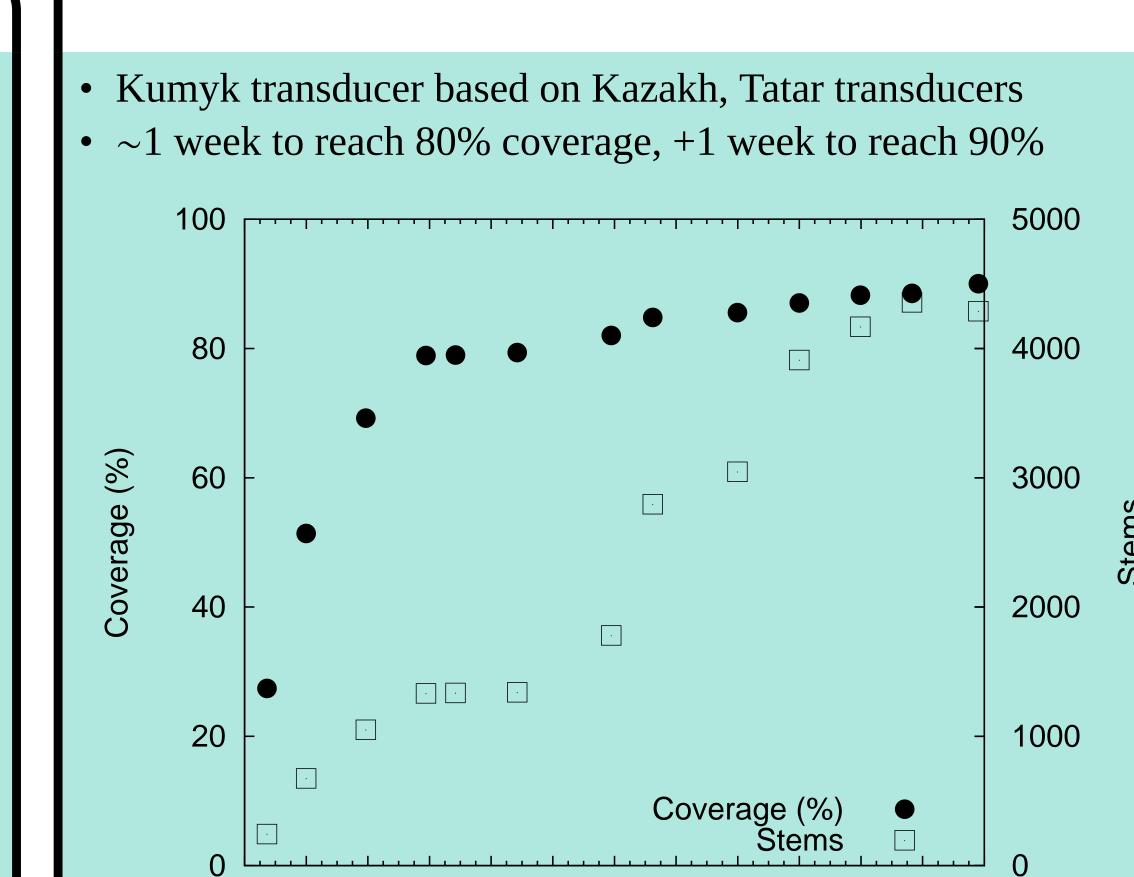
3%0:3%0%{a%}%{3%} NUM-DIGIT ; ! "отыз"

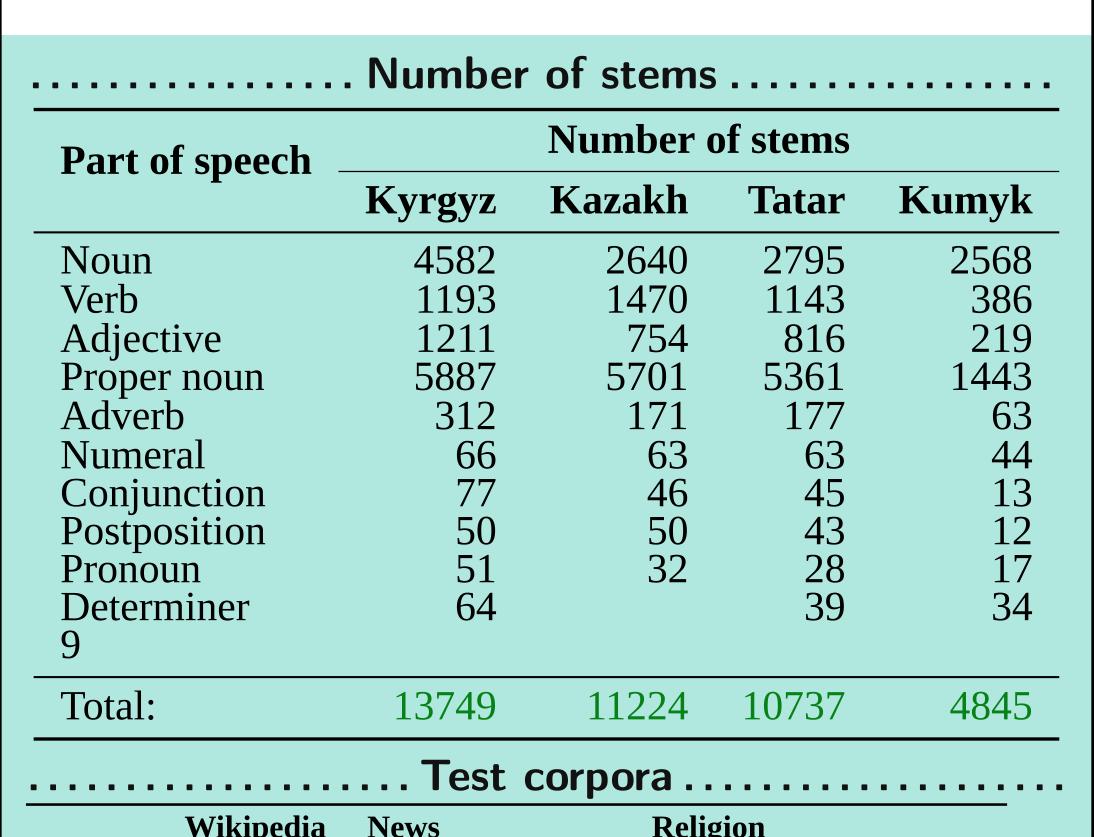
..... A resulting messy twol rule......

5:5%{3%}%{c%} NUM-DIGIT ; ! "бес"

30-дан 'from 30', 5-тен 'from 5'

RdYotVow = ë ю Ë Ю ; AbstractVow = %{a%} %{3%} %{γ%} %{o%} ;





	Wikipedia	TICWS	Kengion
Kyrgyz	Wikipedia	azattyk.org	Quran + Bible
	Wikipedia	azattyq.org	Quran + Bible
Tatar	Wikipedia	tat.tatar-inform.ru	Quran + New Testament
Kumyk	*	yoldash.etnosmi.ru	Genesis + New Testament

- 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

	Corpus	Tokens	Coverage (%)	Amb.
I /	Wikipedia News	M	$84.51 \pm 2.27 \\ 91.43 \pm 0.51$	
Kyrgyz	Religion		土	
(r54474)	Average		土	
	Wikipedia	25.6M	85.61 ± 1.37	2.43
Kazakh	News Religion	3.8M 851K	92.12 ± 2.72 92.49 ± 1.66	2.88 2.63
(r50547)	Average		90.07 ± 1.91	2.64
	Wikipedia	159K	86.35 ± 2.17	2.24
Tatar	News Religion	5.2M 382K	89.75 ± 0.07 91.25 ± 2.55	2.30 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 Forms Precision (%) Recall (%) Kyrgyz Kazakh 57.98 85.65 98.61 95.03 Kumyk
- Disambiguation, more stems
- Clean up Kyrgyz transducer
- 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,