

## FINITE-STATE MORPHOLOGICAL TRANSDUCERS FOR THREE KYPCHAK LANGUAGES

Jonathan North Washington Ilnar Salimzyanov

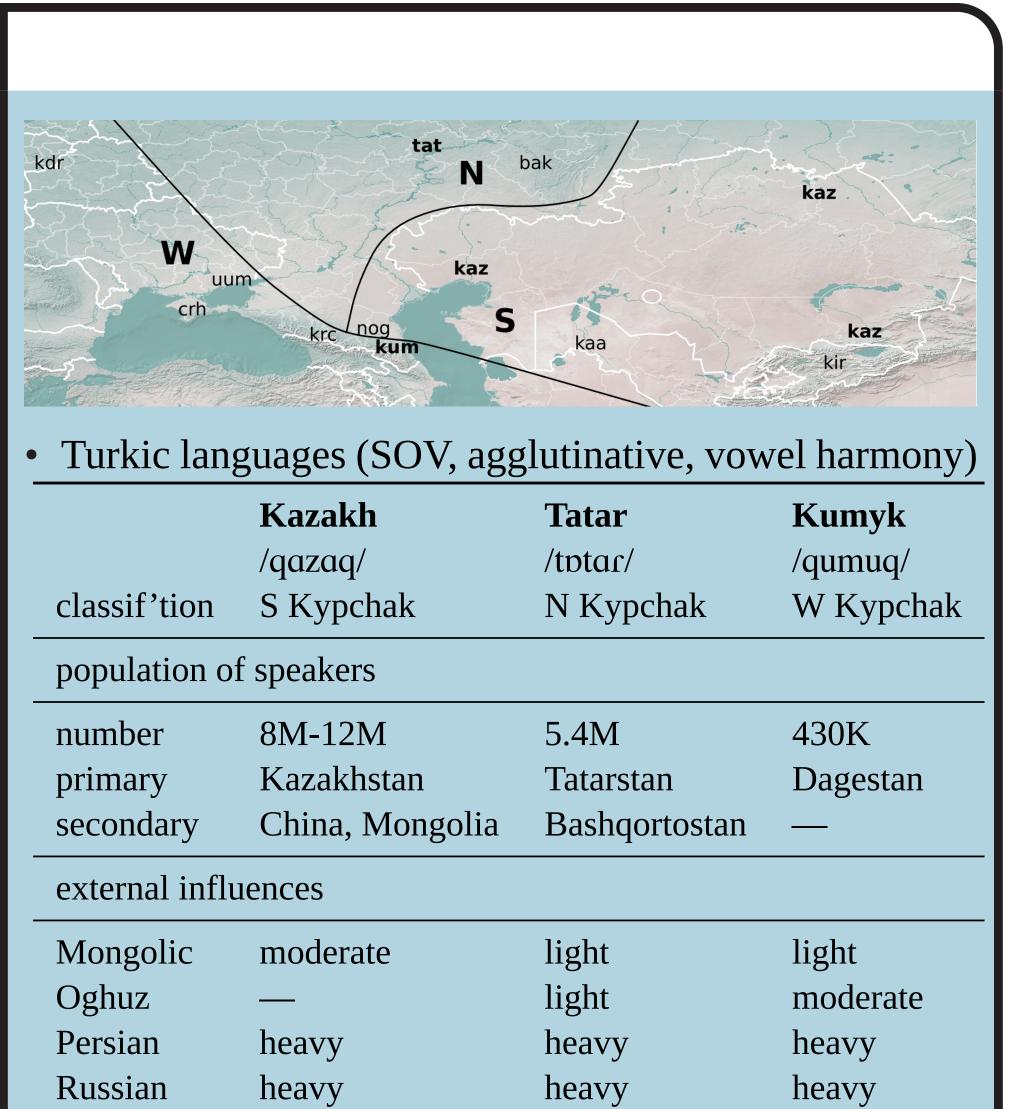
Казан (Идел буе) федераль университеты ilnar.salimzyan@gmail.com

Francis M. Tyers

UiT Norgga Árktalaš Universitehta francis.tyers@uit.no Aida Sundetova sun27aida@gmail.com

Special thanks to





Indiana University

jonwashi@indiana.edu

		Morphological transducers						
• T	ake a	surface form, and produce valid lexica	l form(s)					
• T	• Take a lexical form, and produce valid surface form(s)							
'ал	дым	' ↔ ал <v><tv><ifi><p1><sg>, алд<n><px< td=""><td>1sg&gt;<nom></nom></td></px<></n></sg></p1></ifi></tv></v>	1sg> <nom></nom>					
		Transducers for Turkic languages	5					
• T	urkis	sh (Çöltekin, 2010 & 2014; <mark>Öflazer, 19</mark> 9	4)					
• C	rime	an Tatar (Altıntaş, 2001)						
• T	urkn	nen (Tantuğ et al., 2006)						
• K	yrgy	z (Washington et al., 2012)						
• K	azak	th (Бекманова & Махимов, 2013)						
• 01	ur Ka	azakh, Tatar, Kumyk: all GPL (=free ar	ıd open)!					
		Framework: HFST						
• R	eimp	olements Xerox FST formalisms (lexc	& twol)					
• A	lso	provides a wrapper around popular f	ree/open-					
SO	ource	FST toolkits: SFST, OpenFST, and Fo	ma					
		Development effort						
• K	umy	k transducer based on Kazakh, Tatar tra	nsducers					
• ~	1 we	ek to reach 80% coverage, +1 week to re	each 90%					
	100		<sub>7</sub> 5000					
	80		4000					
(%	60		2000					
ge ( <sup>9</sup>	60		- 3000 SE					
Soverage (%)			Stems					
S	40	<b>-</b> ⊡	2000					
	20		1000					
	20	Coverage (%) ● Stems	- 1000					

• Oth	er Tur	kic transduce	rs: 0-derivation (	(overgenerates)
• Oui	r appro	ach: categoriz	zation (e.g., adje	ctives, below)
Type	Gloss	<adj>(<comp>)</comp></adj>	<adj>(<comp>)<subst></subst></comp></adj>	<adj>(<comp>)<advl></advl></comp></adj>
<b>A1</b>	'good'	яхшы (яхшырак)	яхшы (яхшырак)	яхшы (яхшырак)
<b>A2</b>	ʻold'	иске (искерэк)	иске (искерэк)	<del></del>
<b>A</b> 3	'dead'	үле (—)	үле (—)	<del></del>
A4	'basic'	төп (—)	<del></del>	<del></del>

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

Day

- 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 PRMT mailing list (>25 subscribors);
- Turkic RBMT mailing list (>25 subscribers): apertium-turkic@lists.sourceforge.net Feel free to post in any language!
- Feel free to post in any language!
   See our paper in the LREC 2014 proceedings
- See our paper in the LREC 2014 proceedings
  And feel free to contact the authors any time!

Gloss... Құдай Өзінің жаратқандарының бәріне өте жақсы екенін көрді. қарап, Аллаһ Үзе аларның бик яхшы икәнен күрде. яраткан нәрсәләргә карап, бек яхшы экенин гёрген. Аллагь Оьзю яратгъан затлагъа къарап, 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>  $\theta$ 3prn><ref><px3sp><gen> Y3<prn><ref><px3sp><nom> 0ьз<prn><ref><px3sp><nom> mapar<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> олаponapon бек<adv> бик<adv> өте<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><sg> rëp<v><tv><past><p3><sq> κγp<v><tv><past><p3><sq> <sent> <sent> <sent> Tagset ... Noun Third person <pers> Personal <px3sp> 3rd person poss. <p3> Verb Plural (Singular/Plural) Comma <pl><pl> <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> (Eyewitness/Recent) Adverb Dative <dat> (Past) <adv> <qnt> Quantifier <ger past> Verbal noun (Past) Intransitive Transitive Reflexive <ref> <tv>

	/	Ambiguou	s characters	
Have	e front- aı	nd back-vov	wel readings in	native words
	letters	values	examples	
kaz	и, у, ю	/əj, əw, jəv vej, we, jev	w/ қиюд <mark>а</mark> 'ch w/ киюд <mark>е</mark> 'ge	opping down' tting dressed'
tat	e	э / С _ /j/+ы /j/+э	дәресләр ' еллар 'уес егетләр 'b	irs'
		/ø, y/ / C _	гюнлер ' <i>d</i> - гёзлер ' <i>ey</i>	ays'
kum	ё, ю	/jø, jy/	юреклер ' ёнкюлер '	hearts' darlings'
		/jo, ju/	юлдузл <mark>а</mark> р ёлл <mark>а</mark> р ' <i>rod</i>	
"A fro				
%{A%}:	[ [ [ \ [ :RdYo %}:0 %{γ%}: except [ :RdY	[ :FrontVow   [ .#.   :Vow ] tVow й:0 :RdYo 0] :Cns* ]/[[ : otVow :Cns* %{, [ :0	[ :Vow :ь ] ] :Cns [ :RdVow :ь ] :Cns ] :RdYotVow ] :Cns tVow :Cns :Cns* ]/ :0 - AbstractVow: &%}:0 :Cns* ]/[ :0 Cns :p %{&%}: %>: ow ] :RdYotVow :Cns Vow ]/[ [ :0 - й:0	S:Cns*]/:0_; S:Cns*]/:0_; [:0-й:0]_; ]
%{A%}:	e <=> [ [ [ [ \ [ :RdYo %}:0 %{γ%}: except [ :RdY	[ :FrontVow   [ .#.   :Vow ] tVow й:0 :RdYor 0] :Cns* ]/[[ : otVow :Cns* %{/ [ :0 :Vow - :RdYotVo	[ :RdVow :ь ] :Cns ] :RdYotVow ] :Cns tVow :Cns :Cns* ]/:0 - AbstractVow:	S:Cns*]/:0_; S:Cns*]/:0_; [:0-й:0]_; ]
%{A%}:	e <=> [	[ :FrontVow	[ :RdVow :ь ] :Cns ] :RdYotVow ] :Cns tVow :Cns :Cns* ]/ :0 - AbstractVow: ^%}:0 :Cns* ]/[ :0 Cns :p %{^%}: %>: ow ] :RdYotVow :Cns	S:Cns*]/:0_; S:Cns*]/:0_; [:0-й:0]_; ]  %-:]*_; -%{Д%}:0]_; :Cns*]/:0_; S:Cns*]/:0_; ]  %>:]_;
%{A%}: [[%{3	e <=> [	[:FrontVow   [.#.  :Vow] tVow й:0:RdYo 0]:Cns*]/[[: otVow:Cns*%{/ [::Vow -:RdYotVo [::Vow]	[ :RdVow :ь ] :Cns ] :RdYotVow ] :Cns tVow :Cns :Cns* ]/:0 - AbstractVow:    % } : 0 :Cns* ]/[ :0   Cns :p % { % } : % > : ow ] :RdYotVow :Cns Vow ]/[ [ :0 - й:0	s :Cns* ]/:0 _ ; s :Cns* ]/:0 _ ; [ :0 - й:0 ] _ ; ]
%{A%}:  [[%{3}]  Russ	e <=> [         [ [ ] \         [ :RdYo         [ :RdYo         [ :RdY         [ :RdY         [ :RdY         [ ian 'fron         native wo         елдің 'со	[:FrontVow   [.#.   :Vow ] tVow й:0 :RdYo 0] :Cns* ]/[[: otVow :Cns* %{/ [:: :Vow - :RdYotVo [:: ot' vowels have ord ountry's' o 'scientists'	[:RdVow:ь]:Cns]:RdYotVow]:Cns tVow:Cns:Cns*]/ :0 - AbstractVow:  A%}:0:Cns*]/[:0 Cns:p%{A%}: %>: ow]:RdYotVow:Cns Vow]/[[:0 - й:0   ****  *****  *****  *****  *****  ****	s :Cns* ]/:0 _ ; s :Cns* ]/:0 _ ; [ :0 - й:0 ] _ ; ]
*{A*}:  [[%{3}  Russ  kaz tat kum  solut	e <=> [     [ [ ] \     [ :RdYo     ] : 0   % { y % } :     except     [ :RdY     [ [     ian 'fron     native wo     eлдің 'co     галимнәр     сёзлер 'w tion: sepa	[:FrontVow   [.#.  :Vow ] tVow ŭ:0 :RdYo 0]:Cns* ]/[[: otVow :Cns* %{ [:: otVow -:RdYotVo [:: ord ord ountry's' o'scientists' vords' rate continu cronyms a	[:RdVow:b]:Cns ]:RdYotVow]:Cns tVow:Cns:Cns*]/ :0 - AbstractVow:  A*}:0:Cns*]/[:0 Cns:p%{A*}: %>: ow]:RdYotVow:Cns Vow]/[[:0 - й:0   ***Twords  ***Thermonise as base borrowed word  ***Haзapбaeвтын артистлар 'art самолётлар 'art camonise as base  ation lexicon (  ***Thermonise and numeral	S: Cns* ]/:0 _ ; S: Cns* ]/:0 _ ; S: Cns* ]/:0 _ ; I
%{A%}:  [[%{3}  Russ  kaz tat kum solut  тызда	e <=> [     [ [ ] \     [ :RdYo     ] : 0  % { y % } :     except     [ :RdY     [ :RdY     ]     ian 'fron     native wo     eлдің 'co     галимнәр     cёзлер 'w tion: sepa	[:FrontVow   [.#.   :Vow ] tVow й:0 :RdYo 0] :Cns* ]/[[: otVow :Cns* %{ :Vow - :RdYotVo [::Vow - :RdYotVo cronyms in the continuence cont	[:RdVow:b]:Cns ]:RdYotVow]:Cns tVow:Cns:Cns*]/ :0 - AbstractVow:  %}:0 :Cns*]/[:0 Cns:p%{%}: %>: ow]:RdYotVow:Cns Vow]/[[:0 - й:0	s: Cns* ]/:0 _ ; s: Cns* ]/:0 _ ; [:0 - й:0] _ ; ]   %-: ]* _ ; - %{,2%}:0] _ ; :Cns* ]/:0 _ ; s: Cns* ]/:0 _ ; ]   %>: ] _ ;  ck vowels d ('Nazarbayev's' rists' irplanes' (messy rules)  S
%{A%}:  [[%{s}]  Russ  kaz tat kum  solut  тызда  0-дан	e <=> [	[:FrontVow] [.#.   :Vow] tVow ŭ:0 :RdYor 0] :Cns* ]/[[: otVow :Cns* %{/ :Vow - :RdYotVo [::Vow] t' vowels have ord ountry's' o 'scientists' vords' rate continue cronyms a  becten ten not	[:RdVow:b]:Cns ]:RdYotVow]:Cns tVow:Cns:Cns*]/ :0 - AbstractVow:  As :0:Cns*]/[:0 Cns:p%{As}: %>: ow]:RdYotVow:Cns Vow]/[[:0 - й:0	S: Cns* ]/:0 _ ; S: Cns* ]/:0 _ ; S: Cns* ]/:0 _ ; I
%{A%}:  [[%{s}]  Russ  kaz tat kum  solut  тызда  0-дан	e <=> [     [ [ ] \     [ :RdYo     ] : 0  % { y % } :     except     [ :RdY     [ :RdY     ]     ian 'fron     native wo     eлдің 'co     галимнәр     cёзлер 'w tion: sepa	[:FrontVow] [.#.   :Vow] tVow ŭ:0 :RdYor 0] :Cns* ]/[[: otVow :Cns* %{/ :Vow - :RdYotVo [::Vow] t' vowels have ord ountry's' o 'scientists' vords' rate continue cronyms a  becten ten not	[:RdVow:b]:Cns ]:RdYotVow]:Cns tVow:Cns:Cns*]/ :0 - AbstractVow:  %}:0 :Cns*]/[:0 Cns:p%{%}: %>: ow]:RdYotVow:Cns Vow]/[[:0 - й:0	s: Cns* ]/:0 _ ; s: Cns* ]/:0 _ ; [:0 - й:0] _ ; ]   %-: ]* _ ; - %{,2%}:0] _ ; :Cns* ]/:0 _ ; s: Cns* ]/:0 _ ; ]   %>: ] _ ;  ck vowels d ('Nazarbayev's' rists' irplanes' (messy rules)  S

		M	nber of s	tomo	
Part	of speech	Kazakh	Tatar	Kumy	
NI					
Nour Verb	l	2640 1470	2795 1143	256 38	58 36
Adje	ctive	754	816		19
	er noun	5701	5361	$1\overline{4}$	
Adve		171	177		53
Num		63	63		44
	unction osition	46 50	45 43	_	13 12
Pron		32	28	_	17
	rminer	39	34	-	9
Total	•	11224	10737	484	45
		Test corp	ora		
Wil	kipedia New	/S	Religion	n	
	kipedia tat.ta	tyk.org atar-inform.ru lash.etnosmi.rı	_	Bible New Testa + New Tes	
	Eval	uation me	easures		
<b>Mean a</b> surface	receiving > mbiguity - form found n - of a for	average nui in analysed	d corpus		or ea
Mean a surface Precision Recall - correct for	mbiguity - form found on - of a for solution of a form, b	average numerin analysed manalysed manalysed manalysed with the second s	d corpuses, % corred by tran	rect nsducer t a gold st	hat a
Mean a surface Precision Recall -	mbiguity - form found on - of a form form form form form, but the contract of	in analysed must an analysed m's analysed yes provided by comparing aluation results.	d corpuses, % corred by transfagainstes	rect isducer t a gold st	hat a
Mean a surface Precision Recall - correct for	mbiguity - form found on - of a form form form form, but the corpus	in analysed mains analysed mains analysed yes provided by comparing aluation representation repr	d corpus es, % corr ed by tran ng against esults Covera	rect nsducer t a gold st ge (%)	hat a anda
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a form form form form form, but the contract of	in analysed mains analysed mains analysed yes provided by comparing aluation representation repr	d corpus es, % corr ed by tran ng against esults Covera	rect isducer t a gold st	hat a and a and a <b>Aml</b>
Mean a surface Precision Recall -	mbiguity - form found on - of a for a form, because Wikiped	in analysed myses analysed analysed yes provided by comparing the Tokens are always analysed at 15.6 M analysed analysed at 15.6 M analysed at 15.	d corpus es, % corr ed by tran ng against esults Covera 85.61 92.12	ect sducer t a gold st ge (%) ± 1.37	hat a and a and a <b>Aml</b>
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a for a form, because Wikiped News	in analysed myses analysed analysed yes provided by comparing the Tokens are always analysed at 15.6 M analysed analysed at 15.6 M analysed at 15.	d corpus es, % corr ed by tran ng against esults Covera  85.61 92.12 92.49	rect nsducer t a gold st <b>ge</b> (%) ± 1.37 ± 2.72	hat a and a
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a for of analy for a form, b  Corpus Wikiped News Religion	in analysed myself analysed myself analysed yes provided by comparing aluation relation relation relation analysed analysed by comparing aluation relation relation analysed analysed by comparing aluation relation relation relation relation analysed by comparing aluation relation relation relation analysed by comparing aluation relation relation analysed by the relation relation relation analysed by the relation relation analysed by the relation relation relation and relation analysed by the relation relation relation and relation analysed by the relation relation and relation analysed by the relation relation relation relation and relation analysed by the relation relation relation and relation relation and relation	d corpus es, % corr ed by tran g against esults Covera  85.61 92.12 92.49 90.07	ect nsducer t a gold st <b>ge</b> (%) ± 1.37 ± 2.72 ± 1.66	hat a and a
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a for of analy for a form, b L. L. Eva e Corpus Wikiped News Religion Average Wikiped News	in analysed m's analysed yes provided yes provided a luation result of the second states of t	d corpus es, % corr ed by tran g against esults. Covera  85.61 92.12 92.49 90.07 86.35 89.75	rect isducer to a gold st  ge (%)  ± 1.37  ± 2.72  ± 1.66  ± 1.91  ± 2.17  ± 0.07	hat a and a
Mean a surface Precision Recall - correct for Language Kazakh (r50547)  Tatar	form found on - of a for of analy for a form, b Corpus Wikiped News Religion Average Wikiped News Religion Eva	in analysed m's analysed yes provided yes provided a luation result of the second states of t	d corpus es, % corr ed by tran g against esults Covera  85.61 92.12 92.49 90.07  86.35 89.75 91.25	ect sducer t a gold st  ge (%)  ± 1.37 ± 2.72 ± 1.66  ± 1.91 ± 2.17 ± 0.07 ± 2.55	hat a and a
Mean a surface Precision Recall - correct for Language Kazakh (r50547)	mbiguity - form found on - of a for of analy for a form, b L. L. Eva e Corpus Wikiped News Religion Average Wikiped News	in analysed m's analysed yes provided by comparing a luation range of the second secon	d corpus es, % corr ed by tran g against  Covera  85.61 92.12 92.49  90.07  86.35 89.75 91.25  89.12	ect sducer t a gold st  ge (%)  ± 1.37 ± 2.72 ± 1.66  ± 1.91 ± 2.17 ± 0.07 ± 2.55 ± 1.60	hat a and a
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a for - % of analy for a form, b - Leva  Eva  Corpus  Wikiped News Religion Average  Wikiped News Religion Average  News Religion Average  News	in analysed m's analysed yes provided by comparing aluation results and 159K 5.2M 382K 286K	d corpus es, % corr ed by tran against esults. Covera  85.61 92.12 92.49  90.07  86.35 89.75 91.25  89.12  91.10	ect sducer t a gold st  ge (%)  ± 1.37 ± 2.72 ± 1.66  ± 1.91 ± 2.17 ± 0.07 ± 2.55 ± 1.60 ± 1.60	hat a and a
Mean a surface Precision Recall - correct for the correct for	form found on - of a for of analy or a form, b corpus Wikiped News Religion Average Wikiped News Religion Average News Religion Average News Religion	in analysed m's analysed yes provided by comparing aluation results and 159K 5.2M 382K 286K	d corpus es, % corred by transe against esults Covera  85.61 92.12 92.49  90.07  86.35 89.75 91.25  89.12  91.10 92.47	ect sducer t a gold st  ge (%)  ± 1.37 ± 2.72 ± 1.66  ± 1.91 ± 2.17 ± 0.07 ± 2.55 ± 1.60 ± 1.03	hat a and a
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a for - % of analy for a form, b - Leva  Eva  Corpus  Wikiped News Religion Average  Wikiped News Religion Average  News Religion Average  News	in analysed m's analysed yes provided by comparing aluation results and 159K 5.2M 382K 286K	d corpus es, % corr ed by tran g against  Covera  85.61 92.12 92.49  90.07  86.35 89.75 91.25  89.12  91.10 92.47	ect sducer t a gold st  ge (%)  ± 1.37 ± 2.72 ± 1.66  ± 1.91 ± 2.17 ± 0.07 ± 2.55 ± 1.60 ± 1.60	hat a and a
Mean a surface Precision Recall - correct for the correct for	form found on - of a for of analy or a form, be corpus  Wikiped News Religion Average Wikiped News Religion Average  News Religion Average	in analysed m's analysed yes provided by comparing a luation relation relation relation relation substitution	d corpus es, % corred by trans against esults Covera  85.61 92.12 92.49 90.07  86.35 89.75 91.25  89.12  91.10 92.47  91.78  rface forms	rect a gold st a gold st ge (%) ± 1.37 ± 2.72 ± 1.66 ± 1.91 ± 2.17 ± 0.07 ± 2.55 ± 1.60 ± 1.03 ± 0.94 from news	hat a and a
Mean a surface Precision Recall - correct for the correct for	form found on - of a for of analy or a form, be corpus  Wikiped News Religion Average Wikiped News Religion Average  News Religion Average	in analysed m's analysed yes provided by comparing a luation relation relation relation relation substitution	d corpus es, % corred by transparents ed by transparents esults Covera  85.61 92.12 92.49 90.07 86.35 89.75 91.25 89.12 91.10 92.47 91.78	ect sducer t a gold st  ge (%)  ± 1.37 ± 2.72 ± 1.66  ± 1.91 ± 2.17 ± 0.07 ± 2.55 ± 1.60 ± 1.60 ± 1.03 ± 0.94	hat a and a
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a for % of analy for a form, b	in analysed m's analysed yes provided by comparing a luation random sure a luation random sure as Precise 10	d corpus es, % corred by trans against esults.  Covera  85.61 92.12 92.49  90.07  86.35 89.75 91.25  89.12  91.10 92.47  91.78  rface forms for (%)  98.61	rect a gold st a gold st ge (%)  ± 1.37 ± 2.72 ± 1.66 ± 1.91 ± 2.17 ± 0.07 ± 2.55 ± 1.60 ± 1.03 ± 0.94  from news Recall 5	hat a and a and a and a and a and a and a a and a a and a a a a
Mean a surface Precision Recall - correct for the correct for	mbiguity - form found on - of a for of analy of analy for a form, b  L	in analysed m's analysed yes provided by comparing aluation random summer aluation random summer aluation	d corpus es, % corred by trans against esults. Covera  85.61 92.12 92.49  90.07  86.35 89.75 91.25  89.12  91.10 92.47  91.78  rface forms sion (%)	rect a gold st  ge (%)  ± 1.37  ± 2.72  ± 1.66  ± 1.91  ± 2.17  ± 0.07  ± 2.55  ± 1.60  ± 1.03  ± 0.94  from news  Recall  58	hat a and a

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