

## FINITE-STATE MORPHOLOGICAL TRANSDUCERS FOR THREE KYPCHAK LANGUAGES

Jonathan North Washington Ilnar Salimzyanov

Kазан (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. Sultanmurado A. Sultanmuradov



<ger\_past> Verbal noun (Past)

**Number of stems** 

**Tatar** 

2795

1143

816

Kumyk

2568

386

219

................Number of stems......

Kazakh

2640

1470

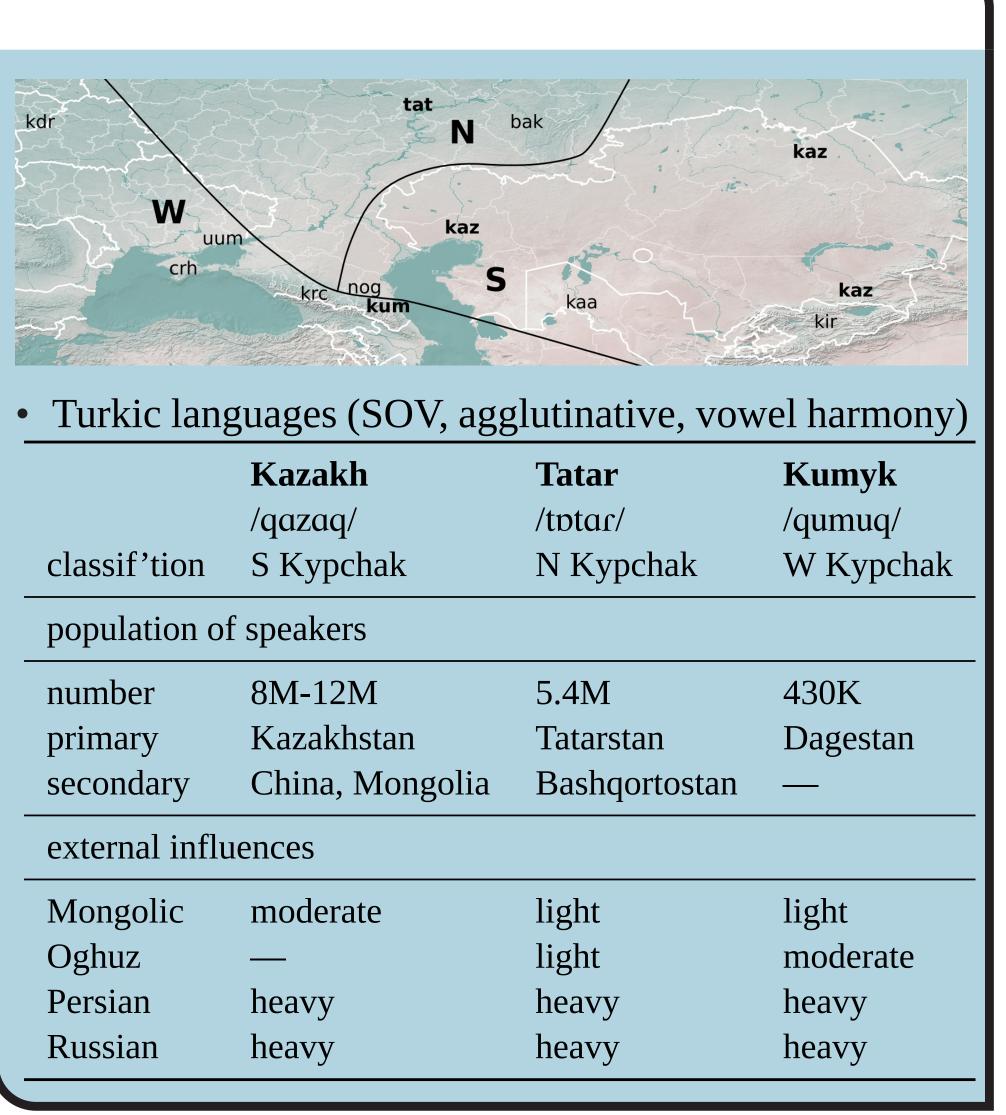
754

Part of speech

Noun

Verb

Adjective



Morphological transducers .....

Indiana University

jonwashi@indiana.edu

• Ta	ake a	surface form, and produce valid lexical	form(	s)
• Ta	ake a	lexical form, and produce valid surface	form(	s)
'алд	дым	' ↔ ал <v><tv><ifi><p1><sg>, алд<n><px1< td=""><td>Lsg&gt;<no< td=""><td>&gt;m&gt;</td></no<></td></px1<></n></sg></p1></ifi></tv></v>	Lsg> <no< td=""><td>&gt;m&gt;</td></no<>	>m>
		Transducers for Turkic languages		
• Tı	urkis	h (Çöltekin, 2010 & 2014; <mark>Öflazer, 199</mark> 4	<b>4</b> )	
• C	rime	an Tatar (Altıntaş, 2001)		
• Tı	urkn	nen (Tantuğ et al., 2006)		
• K	yrgy	z (Washington et al., 2012)		
• K	azak	h (Бекманова & Махимов, 2013)		
• oı	ır Ka	azakh, Tatar, Kumyk: all GPL (=free and	d open	)!
		Framework: HFST		
• R	eimp	olements Xerox FST formalisms (lexc &	z twol	)
• A	lso	provides a wrapper around popular fr	ee/ope	n-
SC	urce	FST toolkits: SFST, OpenFST, and For	na	
		Development effort		
• K	umy	k transducer based on Kazakh, Tatar trar	ısduce	rs
• ~	1 we	ek to reach 80% coverage, +1 week to re	ach 90	%
	100		5000	
	80		4000	
(%) eb	60		3000	
	00		3000	ms
Coverage				Stems
S	40	- _	2000	
	20		1000	
		Coverage (%) ●		
	0	Stems	0	

• Other Turkic transducers: 0-derivation (overgenerates) • Our approach: categorization (e.g., adjectives, below) <adj>(<comp>)<subst> <adj>(<comp>)<advl> Type Gloss <adj>(<comp>) 'good яхшы (яхшырак) яхшы (яхшырак) яхшы (яхшырак) 'old' иске (искерәк) иске (искерэк) 'dead' үле (**—**) уле (—) 'basic' төп (—) — (—)

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

Day

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!

Part of Apertium Turkic project:

Gloss... Құдай Өзінің жаратқандарының бәріне өте жақсы екенін көрді. қарап, Аллаһ Үзе аларның бик яхшы икәнен күрде. яраткан нәрсәләргә карап, Аллагь Оьзю бек яхшы экенин гёрген. яратгъан къарап, олар затлагъа own-his created [everything/thing-s]-to looked.at, they/their very good 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. Аллаh<n><nom> Құдай<n><nom> Аллагь<n><nom> θ3<prn><ref><px3sp><gen> Y3<prn><ref><px3sp><nom> Oьз<prn><ref><px3sp><nom> жapaт<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> Hapca<n><pl><dat>қара<v><tv><gna perf> kapa<v><tv><gna perf> къapa<v><tv><gna perf> аларprn><pers><p3><pl><gen> олаponapon бик<adv> бек<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><sg> κγp<v><tv><past><p3><sg> <sent> .<sent> .<sent> Tagset ..... Noun Third person <pers> Personal 3rd person poss. <px3sp> <p3> (Singular/Plural) Verb Plural Comma <gna\_perf> Verbal adverb 'Nominative' <sent> Sentence Pronoun <nom> Determiner Genitive <past> Past (General) (Perfect) <gen> Past <gpr past> Verbal adjective ifi> (Eyewitness/Recent) (Past)

adj>	A diactiv	70	40.00	Genitive	
241/2	Adjectiv Adverb		<acc> <dat></dat></acc>	Accusative Dative	
	Intransit			Quantifier	
	Transitiv		•	Reflexive	
		Ambiguous (	chara	cters	
				ngs in native wor	
Tiuv	letters	values		mples	<u>—</u>
	letters				?
kaz	и, у, ю	/wej, aw, jaw/ /wej, we, je/	кию кию	д <mark>а 'chopping dowl</mark> д <mark>e 'getting dresse</mark> d	n d'
		э/С_		есл <mark>ә</mark> р 'lessons'	
tat	e	/j/+ы	елла	ap 'years'	
		/j/+э	егет	л <mark>әр 'boys'</mark>	
		/ø, y/ / C _	гюн	лер 'days'	
1			гезл юре	<mark>e</mark> p 'eyes' кл <mark>е</mark> р 'hearts'	
kum	ë, ю	/jø, jy/	ёнк	кл <b>е</b> р 'darlings' узл <mark>а</mark> р 'stars'	
		/jo, ju/	ЮЛД	узлар 'stars'	
		J , J	елла	ap 'roads'	
solut	ion: hairy	y twol rules fo	r maic	rity of cases	
	· ·			rmony-forcing c	har
		or harmony-for			iiui
aujus			icing c	liaracters	
			orda		
T 44		Loanw			
	ers that re	present front v	vowels	in native words	 may
	ers that re esent back	present front v vowels in Ru	vowels ussian	words	may
	ers that re esent back	present front v vowels in Ru	vowels ussian		may
repre kaz	ers that reesent back native wo	present front versions of the version of the versio	vowels issian <b>Russia</b> Hasapõ	words n word example баевтын 'Nazarbaye	
repre kaz tat	ers that reesent back native wo елдің 'со галимнәр	present front versions of the version of the versio	vowels issian <b>Russia</b> Hasapõ	words n word example баевтын 'Nazarbaye	
repre kaz tat kum	ers that reesent back native wo елдің 'со галимнар сёзлер 'м	present front very very solution of the content of	vowels issian Russia Назарб артист самолё	words n word example aeвтың 'Nazarbaye пар 'artists' тлар 'airplanes'	2v's'
repre	ers that reesent back native wo елдің 'со галимнар сёзлер 'wation: sepa	present front vowels in Runtry's' o'scientists' vords'	vowels issian Russia Назарб артист самолё	words n word example баевтын 'Nazarbaye	2v's'
kaz tat kum solut	ers that reesent backesent backesent backesent backesent were ending to certain and the certain and the certain separation: separation: separation with the certain separation and the certain and the certai	present front vowels in Rus  ord example  untry's' o'scientists' yords'  rate continuation	vowels issian Russia Назарб артист самолё	words n word example aeвтың 'Nazarbaye пар 'artists' тлар 'airplanes'	2v's'
kaz tat kum solut	ers that reesent back native wo елдің 'со галимнар сёзлер 'wation: sepa	present front vowels in Rus  ord example  untry's' o'scientists' yords'  rate continuation	vowels issian Russia Назарб артист самолё	words n word example aeвтың 'Nazarbaye пар 'artists' тлар 'airplanes'	2v's'
kaz tat kum solut LEXI	ers that reesent backesent backesent backes we ending to canumner cesses ion: sepandon: sepando	present front vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuation  RUS	vowels issian Russia Назарб артист самолё	words n word example aeвтың 'Nazarbaye пар 'artists' тлар 'airplanes'	2v's'
kaz tat kum solut LEXI	ers that reesent backesent backesent backes native we елдің 'со галимнәр сёзлер 'м сіоп: sepa (CON N1-18%) N1;	present front vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuation  RUS	vowels ussian Russia Назарбартист. самолё	words n word example раевтың 'Nazarbaye пар 'artists' тлар 'airplanes' xicon (messy rule	2v's'
kaz tat kum solut LEXI :%{;	ers that reesent backesent backesent backes at the continum of the certain separation: separation: separation: separation of the certain separation separation of the certain separation separat	present front vowels in Rusord example untry's' o'scientists' vords' rate continuations RUS  Ins ICT N1-RUS;	vowels ussian Russia Hазарбартист. самолё ion lex	words n word example aeвтың 'Nazarbaye nap 'artists' тлар 'airplanes' xicon (messy rule	2v's'
kaz tat kum solut LEXI :%{;	ers that reesent backesent backesent backes at the continum of the certain separation: separation: separation: separation of the certain separation separation of the certain separation separat	present front vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuation  RUS	vowels ussian Russia Hазарбартист. самолё ion lex	words n word example aeвтың 'Nazarbaye nap 'artists' тлар 'airplanes' xicon (messy rule	2v's'
repretatat kum solut LEXI :%{A	ers that reesent back native we елдің 'со галимнәр сёзлер 'м сіоп: sepa (CON N1-12%) N1; CON Nou ет: артим: галимиталим	present front vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuate  RUS  INS ICT N1-RUS; IN1; ! "s	vowels ussian Russia Назарбартист самолё ion lex	words n word example aeвтың 'Nazarbaye nap 'artists' тлар 'airplanes' xicon (messy rule	2v's'
repretatat kum solut LEXI :%{A	ers that reesent back native we елдің 'со галимнәр сёзлер 'м сіоп: sepa (CON N1-12%) N1; CON Nou от:артим:галим	present front volume vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuati  RUS  ICT N1-RUS; IN1; ! "s  cronyms an	vowels ussian Russia Назарбартист самолё ion lex cient d nur	words n word example раевтын 'Nazarbaye nap 'artists' тлар 'airplanes' xicon (messy rule rist" rist" nerals	2v's'
kaz tat kum solut LEXI :%{;	ers that reesent back native we елдің 'со галимнәр сёзлер 'м сіоп: sepa (CON N1-128) N1; СОN Nou от:арти ом:галим галим	present front very vowels in Rus  ord example  untry's' o'scientists' vords' rate continuation  RUS  INS INT N1-RUS; IN1; ! "s  cronyms and	vowels ussian Russia Hазарбартист самолё ion lex scient	words n word example баевтын 'Nazarbaye пар 'artists' тлар 'airplanes' cicon (messy rule sist" nerals spelled-out words	2v's'
kaz tat kum solut LEXI :%{A LEXI арти гали	ers that resent back native we елдің 'со галимнәр сёзлер 'м сіоп: sepa (CON N1- 12%) N1; СОN Nоим галим м: гал	present front very vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuate  RUS  INS ICT N1-RUS; IN1; ! "s  cronyms an  ndle phonolog on thirty', бесте	vowels ussian Russia Hазарбартист самолё ion lex scient d nur y for se	words n word example баевтың 'Nazarbaye пар 'artists' тлар 'airplanes' xicon (messy rule sist" nerals spelled-out words om five'	es)
kaz tat kum solut LEXI apти гали twol	ers that resent back native we елдің 'со галимнәр сёзлер 'м сіоп: sepa CON N1- %} N1; CON Nou ст:артим: галим лим галим	present front very vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuate  RUS  INS INT   "S  cronyms an  ndle phonology of thirty', becreated triggers available.	Nowels  Issian  Russia  Назарбартист  самолё  ion lex  scient  d nur  y for se  ailable	words n word example баевтын 'Nazarbaye пар 'artists' тлар 'airplanes' cicon (messy rule sist" nerals spelled-out words	es)
kaz tat kum solut LEXI :%{A LEXI арти гали тали twol отыз no pl	ers that resent back native we елдің 'со галимнәр сёзлер 'м тіоп: sepa cesne native we result ion: sepa cesne native we result in the result is a result in the result is a result in the result is a result in the result in the result is a result in the result in the result in the result is a result in the re	present front very vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuation  RUS  IN1 ; ! "s  cronyms and and and a phonology of thirty', becreated triggers available of the continuation	vowels ussian Russia Hазарбартист самолё ion lex d nur gy for s en 'fro ailable om 5'	words n word example Taeвтың 'Nazarbaye Tap 'artists' Tлар 'airplanes' Tist" Therals Spelled-out words om five' e in numerals (etc.)	es)
kaz tat kum solut LEXI :%{A LEXI арти гали тали тали	ers that resent back native we елдің 'со галимнәр сёзлер 'м тіоп: sepa cesne native we result ion: sepa cesne native we result in the result is a result in the result is a result in the result is a result in the result in the result is a result in the result in the result in the result is a result in the re	present front very vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuate  RUS  INS INT   "S  cronyms an  ndle phonology of thirty', becreated triggers available.	vowels ussian Russia Hазарбартист самолё ion lex d nur gy for s en 'fro ailable om 5'	words n word example Taeвтың 'Nazarbaye Tap 'artists' Tлар 'airplanes' Tist" Therals Spelled-out words om five' e in numerals (etc.)	es)
kaz tat kum solut LEXI apти гали twol oтыз no pl 30-д solut	ers that resent back native we елдің 'со галимнәр сёзлер 'м сіоп: sepa (CON NOU CT: артим: галим м: г	present front very vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuate  RUS  IN1 ; ! "s  cronyms an  ndle phonology n thirty', бесте cal triggers ave 30', 5-тен 'fronology-triggers' nology-triggers	Nowels  Issian  Russia  Назарбартист  самолё  ion lex  scient  d nur  gy for se  en 'fro  ailable  om 5'  ing ch	words n word example баевтың 'Nazarbaye пар 'artists' тлар 'airplanes' cicon (messy rule spelled-out words om five' e in numerals (etc.	es)
kaz tat kum solut LEXI :%{ LEXI apти гали twol oтыз no pl 30-д solut 4:4%	ers that resent back native we елдің 'со галимнар сёзлер 'м сіоп: sepa cöлер 'м сіоп: sepa con Nou ст: артим: галим м:	present front very vowels in Rus  ord example  untry's' o'scientists' vords'  rate continuate  RUS  cronyms an  ndle phonology thirty', becreated triggers available a	Nowels  Issian  Russia  Назарбартист  самолё  ion lex  scient  d nur  y for se  en 'fro  ailable  m 5'  ing ch  T ; !	words n word example Gaeвтың 'Nazarbaye Tap 'artists' Tлар 'airplanes' dicon (messy rule spelled-out words om five' in numerals (etc. aracters "төрт"	es)
kaz tat kum solut LEXI :%{A  LEXI apти гали тали twol oтыз no pl 30-д solut 4:4% 5:5%	ers that resent back native we елдің 'со галимнәр сёзлер 'м сіоп: sepa CON N1- (СОN NOU СТ: артим: галим на галим на галим на галим ноповоді ан 'from tion: phore (3%)% { Сом	present front very vowels in Rus  ord example  untry's' o'scientists' ords'  rate continuation  RUS  IN1 ; ! "s  cronyms an  ndle phonolog of thirty', becreated triggers available and	Nowels  Issian  Russia  Назарбартист  самолё  ion lex  scient  d nur  gy for s  en 'fro  ailable  om 5'  ing ch  T; !  T; !	words n word example Gaeвтың 'Nazarbaye Tap 'artists' Tлар 'airplanes' dicon (messy rule spelled-out words om five' in numerals (etc. aracters "төрт"	es)

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

 $[[%{3\%}:0]%{\gamma\%}:0] : Cns*]/[[:0 - AbstractVow:] | %-:$ 

RdYotVow =  $\ddot{e}$   $\ddot{b}$   $\ddot{b}$  ; AbstractVow =  $%\{a\%\}$   $%\{y\%\}$   $%\{o\%\}$  ;

"A front unrounded harmony"

%{A%}:e <=>

	/j/+ы /j/+э	еллар 'years' егетләр 'boys'	I	Proper nou Adverb	ın	5701 171	5361 177	1443 63
		гюнл <mark>е</mark> р 'days'		Numeral		63	63	44
	/ø, y/ / C _	гёзл <mark>е</mark> р 'eyes'		Conjunctio		46	45	13
<b>;</b> , ю	/jø, jy/	юрекл <mark>е</mark> р 'hearts' ёнкюл <mark>е</mark> р 'darlings'		Postpositio	n	50 32	43 28	12 17
		юлдузл <mark>ар 'stars'</mark>		Pronoun Determine	r	39	26 34	9
	/jo, ju/	ёллар 'roads'	_	Fotal:		11224	10737	4845
•		or majority of cases			Т	est corp	ora	
		rked harmony-forcing char		Wikipedia	News		Religion	
	_	orcing characters	kaz	Wikipedia			Quran + Bib	
		vords	tat kum	Wikipedia —		ar-inform.ru sh.etnosmi.r		v Testament ew Testament
that re	epresent front	vowels in native words may	_					
nt bac	k vowels in R	ussian words					easures	
ative w	ord example	Russian word example	• Nai	ive covera	<b>ge -</b> per	centage o	f surface form	ns in a given
лдің ' <i>сс</i> алимнә	ountry's' p 'scientists'	Назарбаевтың 'Nazarbayev's' артистлар 'artists'		pus receiv		<b>J</b>	mboy of analy	sos for onch
езлер 'и	words'	самолётлар 'airplanes'		•		•	mber of analy	ses for each
n: sepa	rate continua	tion lexicon (messy rules)		face form		_	es, % correct	
ON N1-	-RUS					J	ed by transdu	icor that are
N1 ;						_	ng against a go	
ON Nou	ıns					-	esults	
г:арти	ист N1-RUS	; ! "artist"			pus	Tokens	Coverage (%	
•	4 N1 ; ! "s			Wik	ipedia	25.6M	$85.61 \pm 1.37$	7 2.43
Δ	cronyms ar	nd numerals	Kaz		vs igion	3.8M 851K	$92.12 \pm 2.72$ $92.49 \pm 1.66$	
	•	gy for spelled-out words	(r50	547) Ave	rage		$90.07 \pm 1.92$	L 2.64
	•	сен 'from five'		Wik	cipedia	159K	$86.35 \pm 2.17$	
•		vailable in numerals (etc.)	Tata	ar Nev		5.2M	$89.75 \pm 0.07$ $91.25 \pm 2.55$	
	саг итддегэ ач 30', 5-тен 'fr		(r50		igion rage	382K	$89.12 \pm 1.60$	
•		ring characters		Nev		286K	$91.10 \pm 0.86$	
-	<u> </u>		Kur	<b>nyk</b> Reli	igion	227K	$92.47 \pm 1.03$	
	C%} NUM-DIG:	IT ; ! "төрт" IT ; ! "бес"	(r50	300) Ave	rage		$91.78 \pm 0.94$	1.53
		-DIGIT ; ! "отыз"	• selec	cted & proof	ed unique	e random su	rface forms from	news corpora
Δ,	esulting me	essy twol rule	La	nguage	Forms	Precis	sion (%) R	ecall (%)
			Kaz	zakh	1000		98.61	57.98
=ёюЁ /ow =%{a	Ю ; a%} %{э%} %{γ%} %{	[0%} ;	Tat		1000		95.03	85.65
unrounde	ed harmony"		Ku	myk ———	500		96.57	69.11
	] ]	:Vow :ь ] ] :Cns :Cns* ]/:0 _ ; :RdVow :ь ] :Cns :Cns* ]/:0 _ ;						
[:RdYc	otVow й:0 :RdYotVo	:RdYotVow ] :Cns :Cns* ]/:0 _ ; ow :Cns :Cns* ]/[ :0 - й:0 ] _ ; - AbstractVow: ]  %-: ]* ;						
xcept		}:0 :Cns* ]/[ :0 - %{,2%}:0 ] ;	D.			40 ct -		
	[:Cns	s :p %{,2%}: %>: :Cns* ]/:0 _ ; ] :RdYotVow :Cns :Cns* ]/:0 _ ;		ambiguati	•		,	
ι ι		v ]/[ [:0 - й:0 ]   %>: ] _ ;	• Ma			between t	hese language	es

- Machine translation between these languages
- Other Turkic lgs.: Nogay, Bashqort, Uzbek, Chuvash