

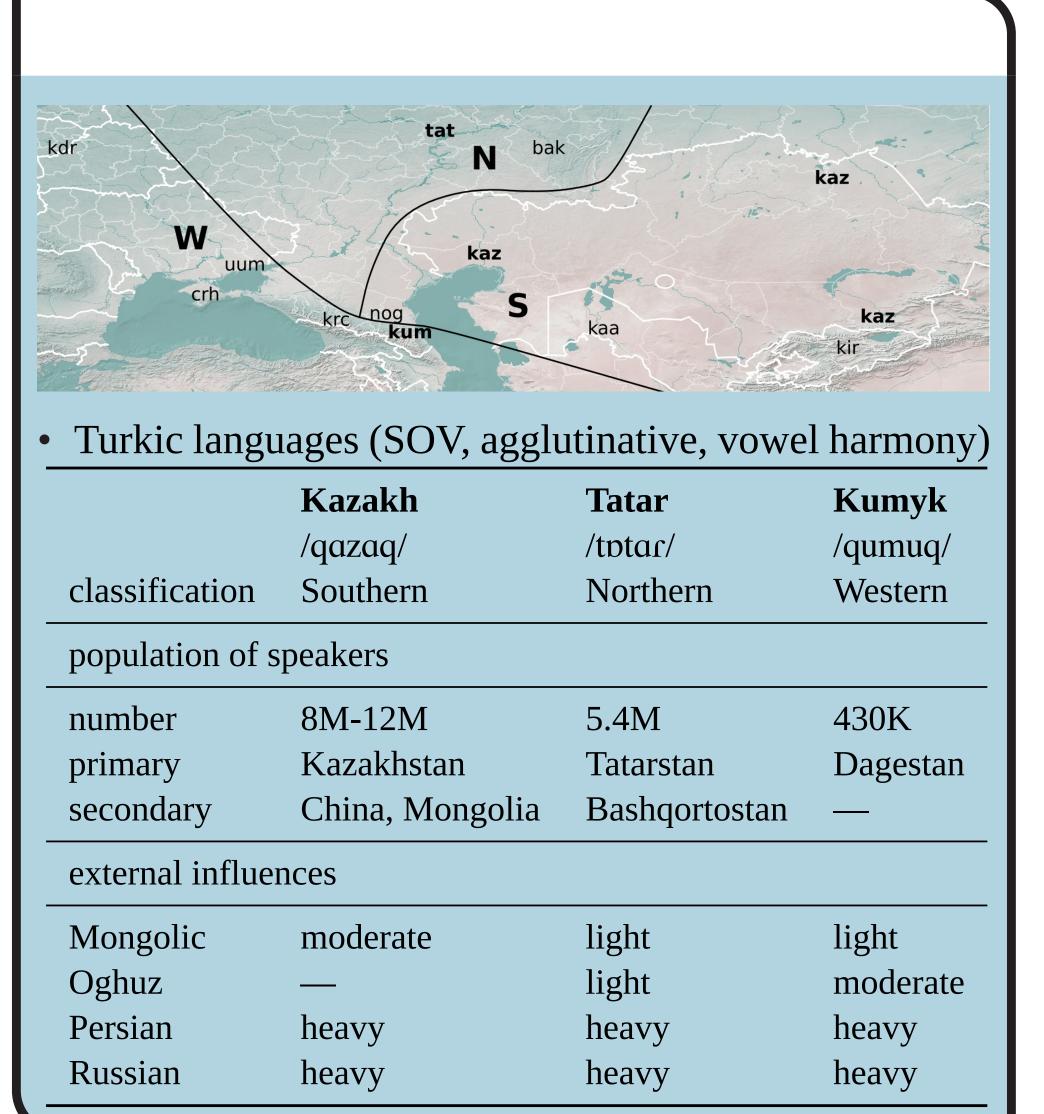
FINITE-STATE MORPHOLOGICAL TRANSDUCERS FOR THREE KYPCHAK LANGUAGES

Jonathan North Washington Ilnar Salimzyanov Indiana University jonwashi@indiana.edu

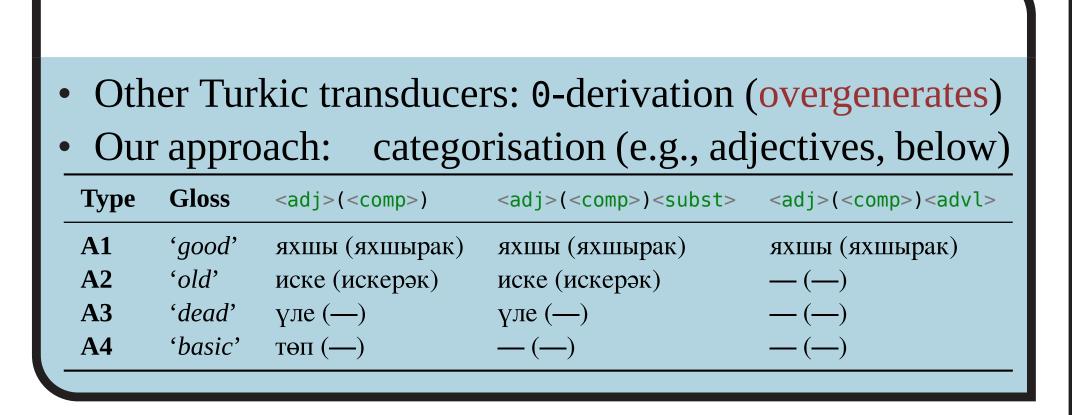
Kaзaн (Идел буе) федераль университеты ilnar.salimzyan@gmail.com

Francis M. Tyers Special thanks to UiT Norgga Árktalaš Universitehta Aida Sundetova francis.tyers@uit.no A. Sultanmuradov





N /
Morphological transducers
 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></nom></px1sg></n></sg></p1></ifi></tv></v>
Transducers for Turkic languages
• Turkish (Çöltekin, 2010 & 2014; Öflazer, 1994)
Crimean Tatar (Altıntaş, 2001)
• Turkmen (Tantuğ et al., 2006)
• Kyrgyz (Washington et al., 2012)
 Kazakh (Бекманова & Махимов, 2013)
• our 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
Development effort
 Kumyk transducer based on Kazakh, Tatar transducers
• ~1 week to reach 80% coverage, +1 week to reach 90%
100 [5000
80 - 4000
€ 60 - 3000
/erage
S 40 - 2000



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

Day

Coverage (%) Stems

1000

• 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!

http://wiki.apertium.org/wiki/Apertium_Turkic

• Part of Apertium Turkic project:

						Gloss						
(1)	Құдай	Өзінің	жаратқанда	рының	бәріне		қарап,		өте	жақсы	екенін	көрді.
	Аллаh	Yзе	яраткан		нәрсәлә	ргә	карап,	аларның	бик	яхшы	икәнен	күрде.
	Аллагь	Овзю	яратгъан		затлагъа	a	къарап,	олар	бек	яхшы	экенин	гёрген.
	God	own-his	created		[everyth	ing/thing-s]-t	o looked.at,	they/their	very	good	being	saw.
	'God loc	oked at ev	verything he h	nad crea	ited and s	saw that it wo	ıs verv aood.	,	-			
Kon		• • • • • • • • • • • • • • • • • • • •				•						• • • • • • • • • • • • • • • • • • • •
	akh (kaz)				Tatar (tat)			Kumyk				
Құд бәр	ай Өзінің іне қарап	, жаратқан , өте жан	ндарының қсы екенін көр			яраткан нәрс ик яхшы икәне	and the control of th				затлагъа ы экенин	
жар бәр қар ,<с Өте жақ е<с көр	aт <v><tv> i<prn><qn a<v><tv>< n> <adv> сы<adj> op><ger_p< th=""><th><pre><px3sp><g <ger_past="" t=""><px3sp> gna_perf> ast><px3se fi=""><p3><g< pre=""></g<></p3></px3se></px3sp></g></px3sp></pre></th><th>t><pl></pl></th><th>sgen> s</th><th>apaт<v><tapar<v><tapar<v><tapar<v><tapa<tapa<tapa<tapa<tapa<tapa<tapa<tap< th=""><th>v><gna_perf> <pers><p3><pl r_past=""><px3sp><past><p3><s< th=""><th>><gen> ><acc> g></acc></gen></th><th>ярат<v> зат<n>< къара<v ,<cm=""> олар<рг бек<adv э<cop="" яхшы<ad="">< гёр<v>< -<sent></sent></v></adv></v></n></v></th><th><tv><q pl><da ><tv>< n><per j> ger_pa tv><pa< th=""><th><gna_perrs><p3><</p3></gna_perrs></th><th><pre>> f> pl><nom> sp><acc></acc></nom></pre></th><th></th></pa<></per </tv></da </q </tv></th></s<></p3></past></px3sp></pl></p3></pers></gna_perf></th></tapa<tapa<tapa<tapa<tapa<tapa<tapa<tap<></tapar<v></tapar<v></tapar<v></v></th></ger_p<></adj></adv></tv></v></qn </prn></tv></v>	<pre><px3sp><g <ger_past="" t=""><px3sp> gna_perf> ast><px3se fi=""><p3><g< pre=""></g<></p3></px3se></px3sp></g></px3sp></pre>	t> <pl></pl>	sgen> s	apaт <v><tapar<v><tapar<v><tapar<v><tapa<tapa<tapa<tapa<tapa<tapa<tapa<tap< th=""><th>v><gna_perf> <pers><p3><pl r_past=""><px3sp><past><p3><s< th=""><th>><gen> ><acc> g></acc></gen></th><th>ярат<v> зат<n>< къара<v ,<cm=""> олар<рг бек<adv э<cop="" яхшы<ad="">< гёр<v>< -<sent></sent></v></adv></v></n></v></th><th><tv><q pl><da ><tv>< n><per j> ger_pa tv><pa< th=""><th><gna_perrs><p3><</p3></gna_perrs></th><th><pre>> f> pl><nom> sp><acc></acc></nom></pre></th><th></th></pa<></per </tv></da </q </tv></th></s<></p3></past></px3sp></pl></p3></pers></gna_perf></th></tapa<tapa<tapa<tapa<tapa<tapa<tapa<tap<></tapar<v></tapar<v></tapar<v></v>	v> <gna_perf> <pers><p3><pl r_past=""><px3sp><past><p3><s< th=""><th>><gen> ><acc> g></acc></gen></th><th>ярат<v> зат<n>< къара<v ,<cm=""> олар<рг бек<adv э<cop="" яхшы<ad="">< гёр<v>< -<sent></sent></v></adv></v></n></v></th><th><tv><q pl><da ><tv>< n><per j> ger_pa tv><pa< th=""><th><gna_perrs><p3><</p3></gna_perrs></th><th><pre>> f> pl><nom> sp><acc></acc></nom></pre></th><th></th></pa<></per </tv></da </q </tv></th></s<></p3></past></px3sp></pl></p3></pers></gna_perf>	> <gen> ><acc> g></acc></gen>	ярат <v> зат<n>< къара<v ,<cm=""> олар<рг бек<adv э<cop="" яхшы<ad="">< гёр<v>< -<sent></sent></v></adv></v></n></v>	<tv><q pl><da ><tv>< n><per j> ger_pa tv><pa< th=""><th><gna_perrs><p3><</p3></gna_perrs></th><th><pre>> f> pl><nom> sp><acc></acc></nom></pre></th><th></th></pa<></per </tv></da </q </tv>	<gna_perrs><p3><</p3></gna_perrs>	<pre>> f> pl><nom> sp><acc></acc></nom></pre>	
	NT.	• • • • • • •		ml	• • • • • • •		D		••••	2	1	
<n></n>	Noun Verb		<p3></p3>	Third p Plural	person	<pre><pers> <cm></cm></pers></pre>	Personal Comma	<	px3sp		d person Singular/	
<pre><pre><pre><pre></pre></pre></pre></pre>	_	ıın	<nom></nom>	'Nomi	native'		Sentence	<	ana n	`	erbal adve	
<det></det>	_		<gen></gen>	Genitiv			Past (Gener		9114_P		Perfect)	
<adj></adj>	Adject	tive	<acc></acc>	Accusa	ative	<ifi></ifi>			gpr_p	`	erbal adje	ctive
<adv></adv>	Adver	b	<dat></dat>	Dative			(Eyewitnes	ss/Recent)		(]	Past)	
<iv></iv>	Intrans		•	Quanti				<	ger_p	ast> Ve	erbal noui	n (Past)
<tv></tv>	Transi	tive	<ref></ref>	Reflex	ive							

<adj></adj>	Adjectiv	re	<acc></acc>	Accusative	<
	Adverb			Dative	
	Intransit		•	Quantifier	
<tv></tv>	Transitiv	/e	<ret></ret>	Reflexive	
					1
		Amhiguaus	chara	cters	
				ngs in native wo	
	letters	values		mples	<u> </u>
	Tetter 5			_	un'
kaz	и, у, ю	/əj, əw, jəw /ej, əw, jəw	// қию // кию	д <mark>a</mark> 'chopping dov д <mark>e</mark> 'getting dresse	ed'
		э/C		есл <mark>ә</mark> р 'lessons'	_
tat	e	/j/+ы	елла	ap 'years'	
		/j/+э		ләр 'boys'	
		/ø, y/ / C _	ГЮН	л <mark>е</mark> р 'days' l <mark>e</mark> p 'eyes'	
kum	ё, ю	lid in	юре	кл <mark>е</mark> р 'hearts'	
TCGIII	C, 10	/jø, jy/	ëhki	юл <mark>е</mark> р 'darlings' узл <mark>а</mark> р 'stars'	
		/jo, ju/	юлд ёлла	ap 'roads'	
	ers that re	Loan present front vowels in I	t vowels	s in native words words	s may
- repre				n word example	
tat	елдің 'со	untry's' o 'scientists'	Назарбартист.	раевтың 'Nazarbay лар 'artists' Этлар 'airplanes'	vev's'
• solu	tion: sepa	rate continua	ation lex	kicon (messy rul	es)
IFX	ICON N1-	RUS			
	%} N1 ;				
LEX	ICON Nou	ns			
-	-	CT N1-RUS			
гали	им:галим	N1 ; ! '	'scient	ist"	
	Δ	cronyms a	nd nu	nerals	
				spelt-out words	
		n thirty', бес		•	
	•		•	in numerals (et	c)
_	•	лаг итддегэ а 30', 5- <mark>те</mark> н ' <i>f</i>		in numeruis (et	·)
	•	nology-trigge		aracters	
		%} NUM-DIO		•	
5:5°	%{	%} NUM-DIO)	oec"	

solution: hairy twol rules cover majority of examples	
unaccounted-for words marked harmony-forcing char	
adjust rules for harmony-forcing characters	Wikip
Loanwords	kaz Wikipe tat Wikipe
Letters that represent front vowels in native words may	kum —
represent back vowels in Russian words	
native word example Russian word example	 Naïve cov
kaz елдің 'country's' Назарбаевтың 'Nazarbayev's'	corpus rec
tat галимнәр 'scientists' артистлар 'artists' kum сёзлер 'words' самолётлар 'airplanes'	• Mean am
solution: separate continuation lexicon (messy rules)	surface for
	 Precision
LEXICON N1-RUS	• Recall - %
:%{,2%} N1 ;	correct for
LEXICON Nouns	
артист:артист N1-RUS ; ! "artist"	
галим:галим N1 ; ! "scientist"	
Acronyms and numerals	Kazakh
twol rules handle phonology for spelt-out words	(r50547)
отыздан 'from thirty', бестен 'from five'	
no phonological triggers available in numerals (etc.)	Tatar
30-дан 'from 30', 5-тен 'from 5'	(F02C0)
solution: phonology-triggering characters	(r50260)
	Kumyk
4:4%{9%}%{c%} NUM-DIGIT ; ! "τθρτ"	(r50300)
5:5%{9%}%{c%} NUM-DIGIT ; ! "δec"	
3%0:3%0%{a%}%{3%} NUM-DIGIT ; ! "отыз"	• selected & p
A resulting messy twol rule	Language
RdYotVow = ёюЁЮ;	Kazakh Tatar
AbstractVow = $%{a%} %{y%} %{0%} ;$	Kumyk
"A front unrounded harmony" %{A%}:e <=> [[:FrontVow [:Vow:ь]]:Cns:Cns*]/:0 _;	
[[:RdVow :ь] :Cns :Cns*]/:0 _ ; [[[\[.#. :Vow]] :RdYotVow] :Cns :Cns*]/:0 _ ;	
[:RdYotVow й:0 :RdYotVow :Cns :Cns*]/[:0 - й:0] _ ; [[%{э%}:0 %{ү%}:0] :Cns*]/[[:0 - AbstractVow:] %-:]* _ ;	
except [:RdYotVow :Cns* %{¾%}:0 :Cns*]/[:0 - %{¾%}:0] _ ;	Disambigu
[:Cns :p %{¾%}: %>: :Cns*]/:0 _ ; [[:Vow - :RdYotVow] :RdYotVow :Cns :Cns*]/:0 _ ;	Machine t
[:Vow]/[[:0 - й:0] %>:] _ ;	Other Turl

Adject		754	816		219
Proper		5701	5361		443
Adver		171	177		63
Nume		63	63		44
Conju		46 50	45 43		13 12
Postpo Pronoi		32	28		17
Detern		39	34		9
Total:		11224	10737	4	845
	T e	est corp	ora		
Wiki	pedia News		Religion	1	
	pedia azattyk	corg r-inform.ru	Quran +		ctomon
tat Wikij		h.etnosmi.ru	Quran + Genesis		
	Evalua verage - pero				
	ceiving ≥ 1	J	Surface i	011115 1	n a grv
•	O		abou of ou]	. .
	nbiguity - av	<u> </u>		iaiyses	s ior ea
	orm found in	analweed	corniis		
surface fo	Jilli Ioulia III	anarysec	Corpus		
	1 - of a form	D	-	ect	
Precision Recall -	1 - of a form? % of analyse	's analyse es provide	es, % corr	ısduce	
Precision Recall - Correct for	n - of a form? % of analyse or a form, byEvalu	's analyse es provide comparin	es, % corred by tranged against esults	sduce a gold	stand
Precision Recall - Correct for	n - of a form? % of analyse or a form, by Evaluation	's analyse es provide comparin uation re Tokens	es, % corred by trans g against esults Coverag	a gold e (%)	stand Amt
Precision Recall - Correct for	n - of a form? % of analyse or a form, by Corpus Wikipedia	's analyse es provide comparin Lation re Tokens 25.6M	es, % corred by trangagainst esults Coverage 85.61 ±	sduce a gold • (%) • 1.37	stand Amt 2.4
Precision Recall - Correct for	a - of a form? % of analyse or a form, by Corpus Wikipedia News	's analyse es provide comparin Lation re Tokens 25.6M 3.8M	es, % corred by transfagagainst esults Coverages 85.61 ± 92.12 ±	sduce a gold • • • • • (%) • 1.37 • 2.72	stand Amt 2.4 2.8
Precision Recall - Correct for	n - of a form? % of analyse or a form, by Corpus Wikipedia	's analyse es provide comparin Lation re Tokens 25.6M	es, % corred by trangagainst esults Coverage 85.61 ±	a gold e (%) 1.37 2.72 1.66	2.4 2.8 2.6
Precision Recall - Correct for Kazakh	a - of a form? % of analyse or a form, by LEvalue Corpus Wikipedia News Religion Average	's analyse es provide comparin Lation re Tokens 25.6M 3.8M 851K	es, % corred by transparent states against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ±	sduce a gold • • • • • (%) • 1.37 • 2.72 • 1.66 • 1.91	2.4 2.8 2.6 2.6
Precision Recall - Correct for Kazakh (r50547)	of analyse of analyse of analyse or a form, by one of the corpus Corpus Wikipedia News Religion Average Wikipedia News	's analyse es provide comparin Lation re Tokens 25.6M 3.8M	es, % corred by trans against esults Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ± 89.75 ±	e (%) 1.37 2.72 1.66 1.91 2.17 0.07	2.4 2.8 2.6 2.6 2.2
Precision Recall - Correct for Kazakh	of a form? of analyse or a form, by or a form, by or a form, by or a forpus Corpus Wikipedia News Religion Average Wikipedia	's analyse es provide comparin Jation re Tokens 25.6M 3.8M 851K	es, % corred by transparent against esults Coverage 85.61 ± 92.12 ± 92.49 ± 90.07 ± 86.35 ±	e (%) 1.37 2.72 1.66 1.91 2.17 0.07	2.4 2.8 2.6 2.6 2.2 2.3
Precision Recall - Correct for Kazakh (r50547)	of analyse of analyse of analyse or a form, by one of the corpus Corpus Wikipedia News Religion Average Wikipedia News	's analyse es provide comparing to the second secon	es, % corred by trans against esults Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 86.35 ± 89.75 ±	a gold a gold (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55	2.4 2.8 2.6 2.6 2.2 2.3 2.2
Precision Recall - 9 correct for Kazakh (r50547) Tatar (r50260)	of analyse or a form, by or a form, by or a form, by or a form, by or a form. Corpus Wikipedia News Religion Average Wikipedia News Religion Average News News	's analyse es provide comparing to the c	85.61 ± 92.12 ± 92.49 ± 86.35 ± 89.75 ± 91.25 ± 91.10 ±	e (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55 1.60 0.86	2.4 2.8 2.6 2.6 2.2 2.3 2.2 2.2
Precision Recall - Correct for Kazakh (r50547) Tatar (r50260) Kumyk	of analyse or a form, by or a form, by or a form, by or a form, by or a form. Corpus Wikipedia News Religion Average Wikipedia News Religion Average	's analyse es provide comparing ation recommendation recommendatio	85.61 ± 92.12 ± 92.49 ± 89.75 ± 91.25 ± 92.47 ± 92.47 ±	e (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55 1.60 0.86 1.03	2.4 2.8 2.6 2.6 2.2 2.3 2.2 2.2 1.5 1.5
Precision Recall - 9 correct for Kazakh (r50547) Tatar (r50260)	of analyse or a form, by or a form, by or a form, by or a form, by or a form. Corpus Wikipedia News Religion Average Wikipedia News Religion Average News News	's analyse es provide comparing to the c	85.61 ± 92.12 ± 92.49 ± 86.35 ± 89.75 ± 91.25 ± 91.10 ±	e (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55 1.60 0.86 1.03	stand Amb
Precision Recall - Correct for	n - of a form % of analyse or a form, by L. L. Evaluation Corpus Wikipedia News Religion Average Wikipedia News Religion Average News Religion	's analyse es provide comparin recomparin Tokens 25.6M 3.8M 851K 159K 5.2M 382K 286K 227K	85.61 ± 92.12 ± 92.49 ± 91.25 ± 91.10 ± 92.47 ± 91.78 ±	e (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55 1.60 1.60 0.86 1.03 0.94	2.4 2.8 2.6 2.6 2.2 2.3 2.2 2.3 2.2 1.5 1.5
Precision Recall - Correct for	of analyse or a form, by or a	's analyse es provide comparin recomparin recomparing ation recomparing 25.6M, 3.8M, 851K. 159K, 5.2M, 382K. 286K, 227K. random sur	85.61 ± 92.12 ± 92.49 ± 91.25 ± 91.10 ± 92.47 ± 91.78 ±	sducer a gold (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55 1.60 1.60 0.86 1.03 0.94 from nev	2.4 2.8 2.6 2.6 2.2 2.3 2.2 2.3 2.2 1.5 1.5 ws corp
Precision Recall - 9 correct for Kazakh (r50547) Tatar (r50260) Kumyk (r50300) selected & 9 Kazakh	of analyse or a form, by or a form. Corpus Wikipedia News Religion Average News Religion Average Proofed unique Forms 1000	's analyse es provide comparin recomparin recomparing ation recomparing 25.6M, 3.8M, 851K. 159K, 5.2M, 382K. 286K, 227K. random sur	ed by transes against esults Coverage 85.61 ± 92.12 ± 92.49 ± 92.49 ± 89.75 ± 91.25 ± 91.78 ± 91.78 ± 100 (%) face forms forms forms forms (%) 98.61	sducer a gold (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55 1.60 1.60 0.86 1.03 0.94 from nev	Stand Amb 2.4 2.8 2.6 2.6 2.2 2.3 2.2 1.5 1.5 1.5 1.5 41 (%) 57.98
Precision Recall - Second Correct for Corr	of analyse or a form, by or a form, by or a form, by or a form, by or a form. Corpus Wikipedia News Religion Average Wikipedia News Religion Average News Religion Average Proofed unique Forms	's analyse es provide comparin recomparin recomparing ation recomparing 25.6M, 3.8M, 851K. 159K, 5.2M, 382K. 286K, 227K. random sur	85.61 ± 92.12 ± 91.25 ± 91.78 ± 10n (%)	sducer a gold (%) 1.37 2.72 1.66 1.91 2.17 0.07 2.55 1.60 1.60 0.86 1.03 0.94 from nev	2.4 2.8 2.6 2.6 2.2 2.3 2.2 2.3 2.2 1.5 1.5

Number of stems.....

Kazakh

2640

1470

Part of speech

Noun

Verb

Number of stems

Tatar

2795

1143

Kumyk

2568

386

guation, more stems

translation between these languages

• Other Turkic lgs.: Nogay, Bashqort, Uzbek, Chuvash