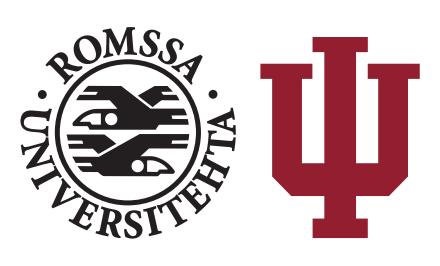
# A FINITE-STATE MORPHOLOGICAL ANALYSER FOR TUVAN



Francis M. Tyers UiT Norgga Árktalaš Universitehta francis.tyers@uit.no

Jonathan North Washington Aziyana Bayyr-ool Indiana University

Сибирское отделение Российской академии наук, azikoa@mail.ru

Aelita Salchak Тываның Күрүне Университеди aelita\_74@mail.ru







## Example

jonwashi@indiana.edu

« Бис кожууннуң соңгукчулары-биле ажылды чорутпастай бээривиске, көргүзүглер баксыраан. » "Figures worsened when we stopped conducting business with the district's constituents."

^Бис/бис<prn><pers><p1><pl>><nom>\$ ^кожууннуң/кожуун<n><gen>\$

^соңгукчулары/соңгукчу<n><pl><px3sp><nom>\$ ^-биле/биле<post>\$

^ажылды/ажыл<n><acc>\$

^чорутпастай/чорут<v><tv><cess><prc impf>\$ ^бээривиске/бер<vaux><ger aor><px1pl><dat>\$

^,/,<cm>\$

`көргүзүглер/көргүзүг<n><pl><nom>\$

^баксыраан/баксыра<v><iv><past><p3><pl>\$

./.<sent>\$

Part of speech	Tag	Stems
Noun	<n></n>	4,226
Proper noun	<np></np>	4,217
Adjective	<adj></adj>	1,603
Verb	<v></v>	1,064
Adverb	<adv></adv>	136
Numeral	<num></num>	85
Conjunction	<cnj*></cnj*>	70
Postposition	<post></post>	28
Pronoun	<pre><prn></prn></pre>	35
Determiner	<det></det>	26
Total		11,490

# Morphological Transducers

Very little work on computational tools

Approximately 300,000 speakers

Agglutinative morphology

#### 

- 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  $\leftrightarrow$  ал<v><tv><ifi><p1><sg>>, алд<n><px1sg><nom>>ΘΓ<n><px3sp><loc> өөнде

#### ..... Framework: HFST.....

- Reimplements Xerox FST formalisms (lexc & twol)
- Also provides a wrapper around popular free/opensource FST toolkits: SFST, OpenFST, and Foma

#### ...... Approach .........

- two-level method (Koskenniemi, 1983)
- morphotactics implemented in lexc
- morphophonology implemented in twol (SPE-style) rules)
- compiled separately; compose-intersected to single transducer
- алдым  $\leftrightarrow$  ал> $\{D\}\{I\}$ >м  $\leftrightarrow$  ал<v><tv><ifi><p1><sg> $aлд>{I}M \longleftrightarrow aлд<n><px1sg><nom>$ алдым  $\Theta\ThetaHДE \longleftrightarrow \Theta\Gamma > \{Z\}\{I\}\{n\} > \{D\}\{A\} \longleftrightarrow \Theta\Gamma < n > \{px3sp> < loc> \}$

dative

рульга

рубльге

:BackVow :Cns\* :Cns :л ь: :Cns\* :RealCns ]/:0\* \_ ;

new

grammatical

generalisation

morph

phon

type of

problem?

update

update

phonol.

[ :BackVow :Cns\* :Cns :ль:0 ]/:[ :0 -ь: ]\* \_ ;

[ :BackVow :Cns\* :Cns :ль:0 ]/:[ :0 -ь: ]\* \_ ;

медальга

ансамбльге

genitive

рульдуң

рубльдиң

медальдың

ансамбльдиң

Finding out new linguistics things

 ${I}$ : Vy <=> : Vx [ :Cns\* :RealCns ]/[ :0 | %- ]\* \_ ;

where Vx in (уиеэ ө а о ы у я ё ю )

Vy in (үиииүыуыуы)

%{I%}:и <=> [ :BackVow :Cns\* :Cns :ль: :Cns\* :RealCns ]/:0\* \_ ;

stem

руль

рубль

"{I} harmony"

matched;

Development cycle

assumed

generalisations

compile

analyser

analyse

corpus

examine

freq. list

missing

stem?

add stem -

initial

development

unanalysed

forms

 $^{\prime\prime}$ {I} always front when intervening Сль $^{\prime\prime}$ 

медаль

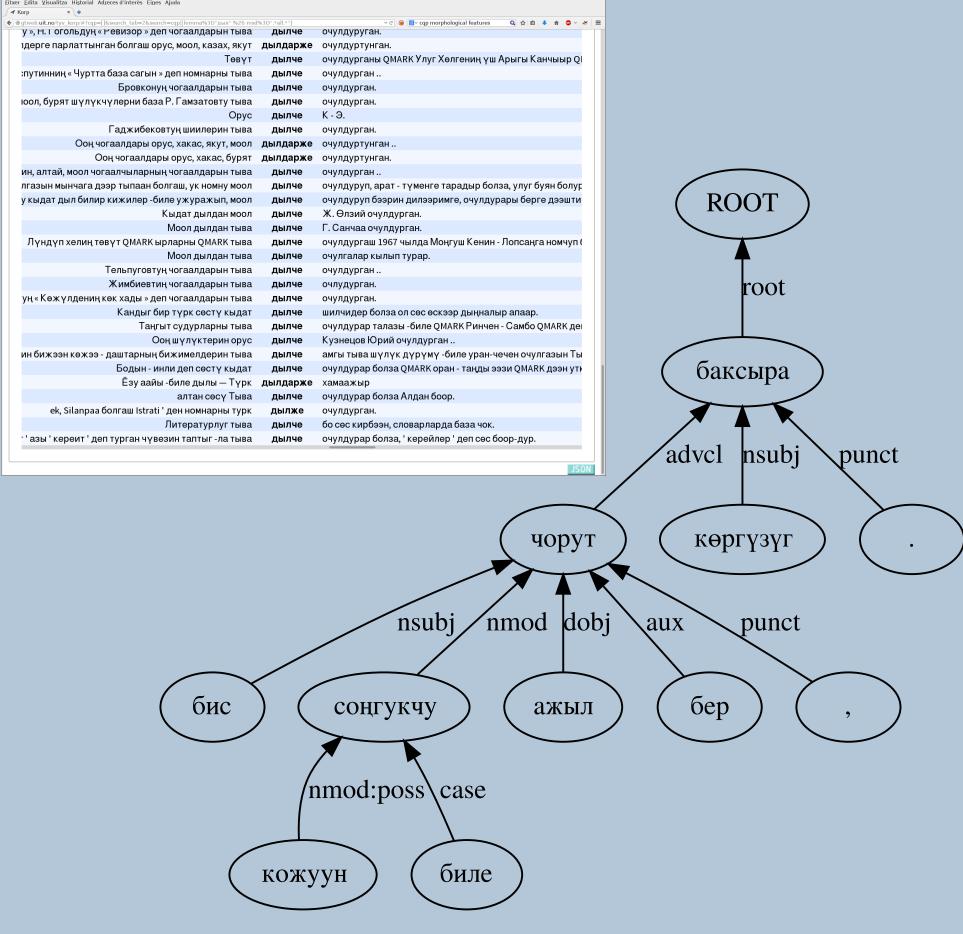
ансамбль

# Ongoing and future work

- ..... our other Turkic-language transducers .....
- **Kyrgyz** (Washington et al., 2012)
- Tatar & Bashkort (Tyers et al., 2012)
- Kazakh (Salimzyanov et al., 2013)
- Kumyk (Washington et al., 2014) ongoing work on Sakha, Khakas, and more!
- Kazakh (Assylbekov et al., 2016, forthcoming)

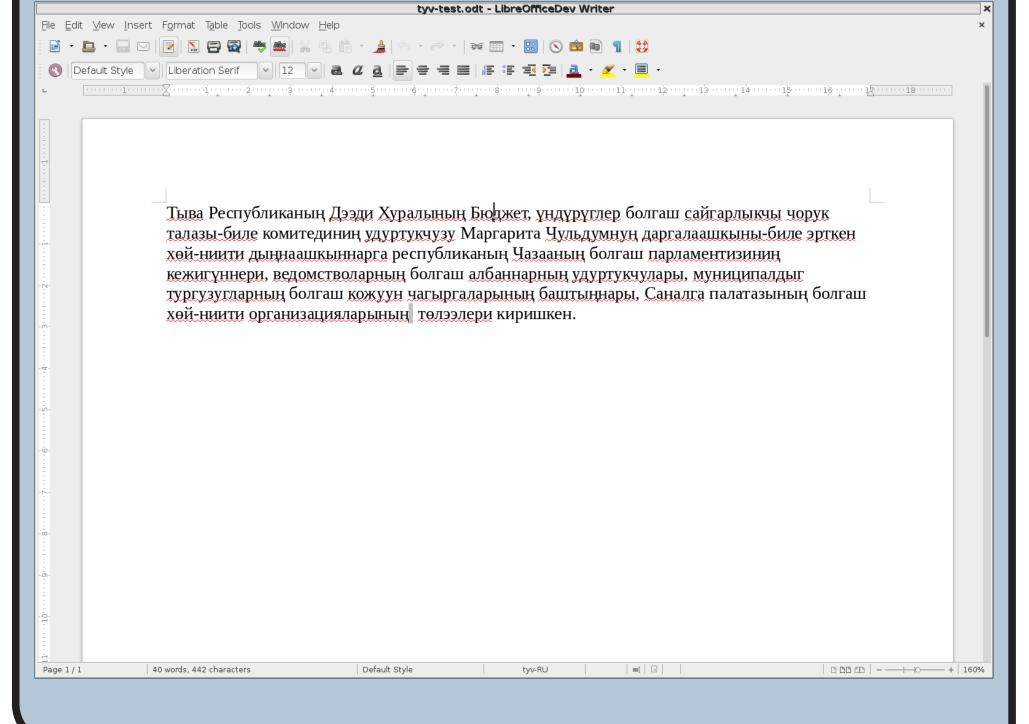
#### ..... Treebanks.....

- Kazakh (Tyers and Washington, 2015)
- work on **Tuvan** ongoing



### 

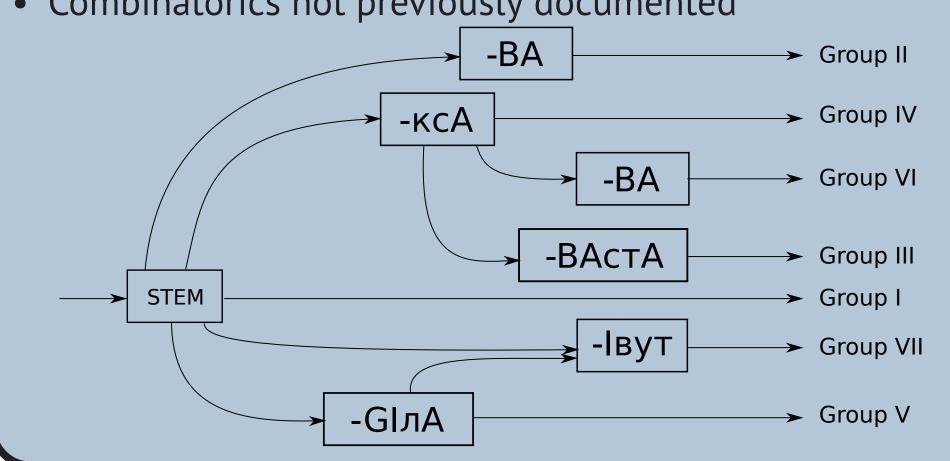
- Dependency parsing (native and crosslingual)
- Increase number of stems
- Make more available to linguists
- Integrate into end-user applications (spellchecking, MT)



## Morphotactics

Lexicon contents /

- Introduction of quasi-derivational verbal morphotactics
- Combinatorics not previously documented



## Evaluation

..... 5-part corpus for naïve coverage ......

Domain	Tokens	Coverage (%)
News	1,539,459	95.73
Religion	746,124	93.84
Literature	297,830	91.96
Encyclopaedic	276,547	90.86
Folklore	27,902	91.57
Average	_	92.79

Precision and recall.....

	count	precision	recall
known tokens	1024	0.99	0.97
all tokens	1425	0.99	0.69

.....Qualitative evaluation.....

error type	count	percentage
Missing stem	364	78.8
Other	65	14.1
Bad morphotactics	19	4.1
Bad phonology	8	1.7
Incorrect categorisation	6	1.3
Total:	462	100

# References

Assylbekov, Zhenisbek, Jonathan North Washington, Francis Tyers, Assulan Nurkas, Aida Sundetova, Aidana Karibayeva, Balzhan Abduali, and Dina Amirova (2016, forthcoming). "A free/open-source hybrid morphological disambiguation tool for Kazakh". In Proceedings of the 17th Annual Conference on Intelligent Text Processing and Computational Linguistics.

Koskenniemi, K. (1983). Two level morphology: a general computational model for wordform recognition and production. Helsinki: Helsingin yliopisto. Salimzyanov, Ilnar, Jonathan North Washington, and Francis Morton Tyers (2013). "A free/open-source

Kazakh-Tatar machine translation system". In Proceedings of Machine Translation Summit XIV. Nice, Tyers, Francis and Jonathan Washington (2015). "Towards a free/open-source universal-dependency treebank for Kazakh". In Proceedings of the 3rd International Conference on Computer Processing

in Turkic Languages (TurkLang 2015). Kazan, Tatarstan. Tyers, Francis, Jonathan North Washington, Ilnar Salimzyan, and Rustam Batalov (2012). "A prototype machine translation system for Tatar and Bashkir based on free/open-source components". In Proceedings of the First Workshop on Language Resources and Technologies for Turkic Languages at the Eighth International Conference on Language Resources and Evaluation (LREC'12). İstanbul,

Washington, Jonathan, Mirlan Ipasov, and Francis Tyers (2012). "A finite-state morphological transducer for Kyrgyz". In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). İstanbul.

Washington, Jonathan North, Ilnar Salimzyanov, and Francis M. Tyers (2014). "Finite-state morphological transducers for three Kypchak languages". In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavík, Iceland.

## Further information



• Part of Apertium Turkic project: http://wiki.apertium.org/wiki/Apertium Turkic

• Transducers available **live** on our website: http://turkic.apertium.org/

• Source code under GPL from Apertium's SVN repo

no

better?

better?

freq list

- Multilingual Turkic RBMT mailing list (>25 subscribers):
- apertium-turkic@lists.sourceforge.net And feel free to contact the authors any time!



Четтир<v><tv><ifi><p1><pl>

http://turkic.apertium.org/

