A finite-state morphological analyser for Tuvan

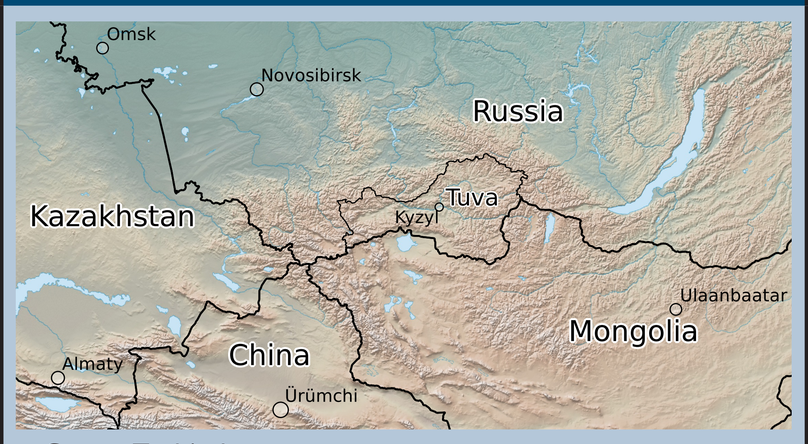
Francis M. Tyers
UiT Norgga Árktalaš Universitehta
francis.tyers@uit.no

Jonathan North Washington
Indiana University
jonwashi@indiana.edu

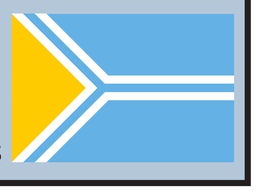
Aziyana Bayyr-ool
Сибирское отделение Российской академии наук, azikoa@mail.ru

Aelita Salchak
Тываның Күргүне Университеди
aelita_74@mail.ru

Tuvan



- Sayan Turkic language
- Spoken in the Tuva Republic
- Approximately 300,000 speakers
- Agglutinative morphology
- Very little work on computational tools



Example

« 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>$
```

Lexicon contents ↗

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

Morphological Transducers

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>
  өөнде ↔ өг<n><px3sp><loc>

Framework: HFST

- Reimplements Xerox FST formalisms (`lexc` & `twol`)
- Also provides a wrapper around popular free/open-source 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 ↔ ал>{D}{I}>м ↔ ал<v><tv><ifi><p1><sg>
  алдым ↔ алд>{I}м ↔ алд<n><px1sg><nom>
  өөнде ↔ өг>{z}{I}{n}>{D}{A} ↔ өг<n><px3sp><loc>

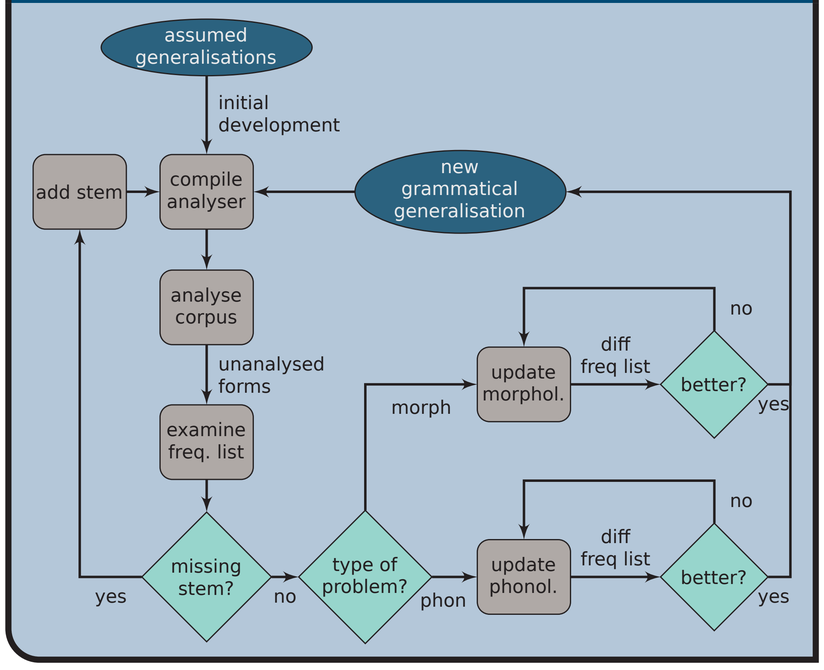
What should we put here

stem	V	C	dative	genitive
медаль	а	ль	медальга	медальдың
ансамбль	а	бль	ансамбльге	ансамбльдиң
руль	у	ль	рульга	рульдуң
рубль	у	бль	рубльге	рубльдиң

```
"{I} harmony"
%{I%}:Vy <=> :Vx [ :Cns* :RealCns ]/[ :0 | %- ]* _ ;
        except
            [ :BackVow :Cns* :Cns :л ь:  :Cns* :RealCns ]/:0* _ ;
            [ :BackVow :Cns* :Cns :л ь:0 ]/:[ :0 - ь:  ]* _ ;
        where Vx in ( ү и е э ө а о ы у я ё ю )
              Vy in ( ү и и и ү ы у ы у ы у у )
        matched ;

"{I} always front when intervening Сль"
%{I%}:и <=> [ :BackVow :Cns* :Cns :л ь:  :Cns* :RealCns ]/:0* _ ;
            [ :BackVow :Cns* :Cns :л ь:0 ]/:[ :0 - ь:  ]* _ ;
```

Development cycle



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!

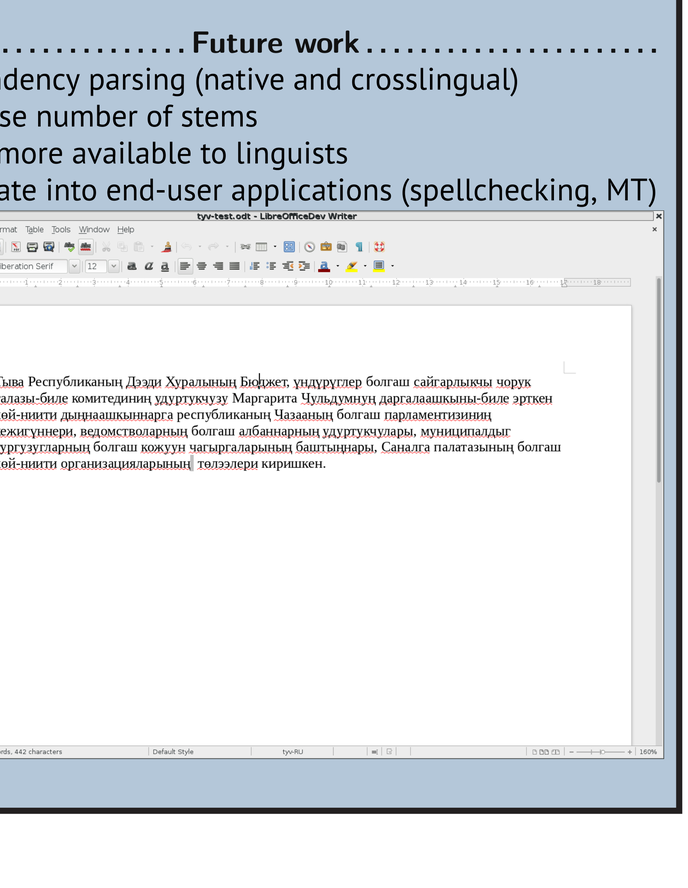
Morphological disambiguation

- **Kazakh** (Assylbekov et al., 2016, forthcoming)

Treebanks

- **Kazakh** (Tyers and Washington, 2015)

- work on **Tuvan** ongoing

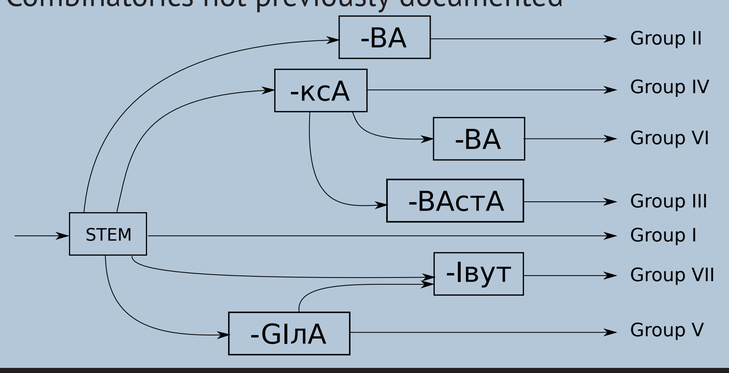


Future work

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

Morphotactics

- 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, France.

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, Turkey.

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

Четтирдивис!

четтир<v><tv><ifi><p1><pl>

`http://turkic.apertium.org/`