A FINITE-STATE MORPHOLOGICAL TRANSDUCER FOR KYRGYZ

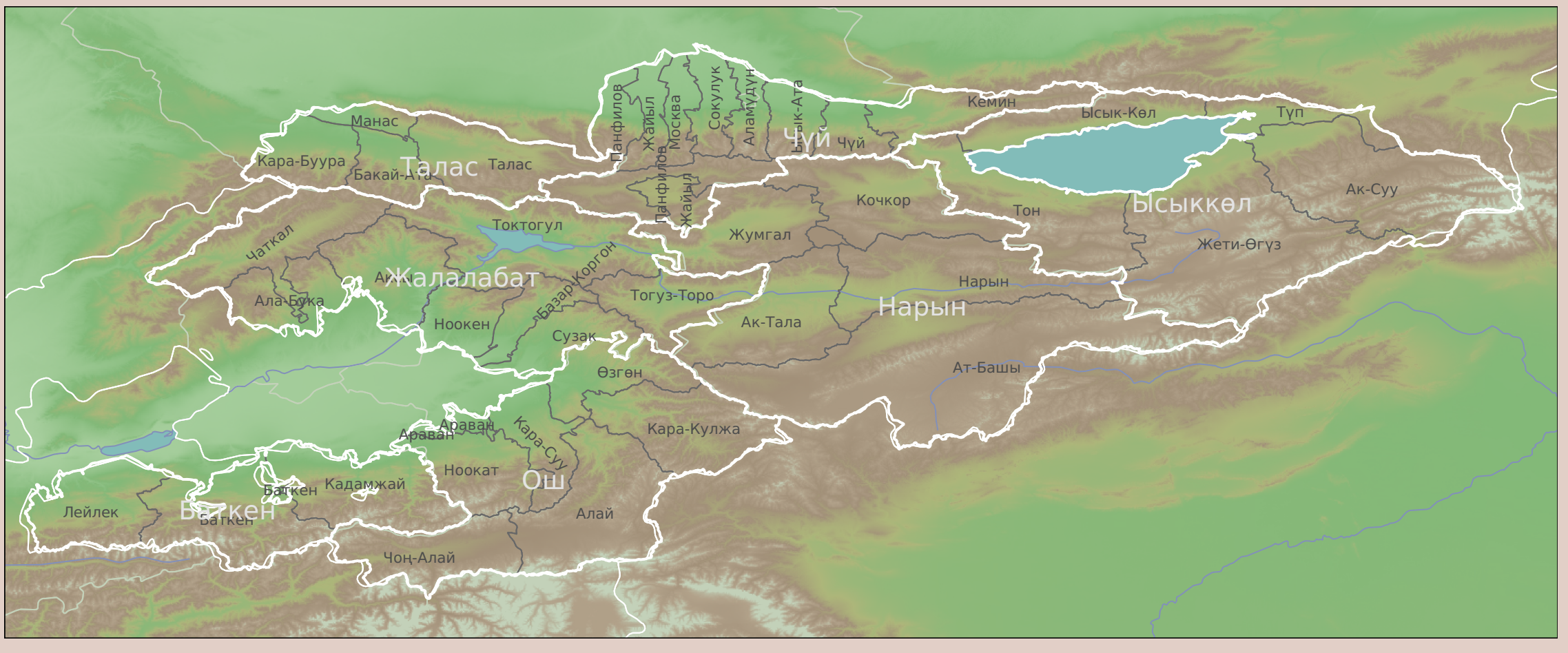
Jonathan North Washington
Indiana University
jonwashi@indiana.edu

Mirlan Ipasov
International Atatürk Alatoo University
mipasov@gmail.com

Francis M. Tyers
Universitat d'Alacant
ftyers@dlsi.ua.es

Kyrgyz

- Turkic language
  - Similar historically to Southern Altay
  - Similar by convergence to Kazakh, Uzbek
- Spoken in
  - Kyrgyzstan as official language
    high levels of bilingualism with Russian
  - China, Tajikistan, Uzbekistan
- Over 3 million native speakers
  (estimate based on data from Ethnologue)



Morphological transducers

Morphological transducers

- Take a surface form, and produce all valid lexical forms
  e.g. 'алдым'
- Take a lexical form, and produce one or more valid surface forms
  e.g., ал<v><tv><ifi><p1><sg>, алд<n><px1sg><nom>

Transducers for Turkic languages

- Turkish (Çöltekin, 2010; Öflazer, 1994)
- Crimean Tatar (Altıntaş, 2001)
- Turkmen (Tantuğ et al., 2006)
- …this is the first transducer for Kyrgyz
- and it's GPL (=free and open)!

Framework: HFST

- Reimplementation of Xerox FST formalisms (lexc and twol)
- Also provides a wrapper around popular FST free/open-source toolkits: SFST, OpenFST, and Foma

Morphotactics

Morphological & orthographical words

- өнүктүрөбүзбү ? '*will we develop [it]?*'
  өнүк<v><tv><caus><aor><p1><pl>+бы<qst>
- келатсаң '*if you come*'
  кел<v><iv><prt_impf>+жат<vaux><gna_cnd><p2><sg>

Irregular [noun + possessive + case] forms

- Some combinations of possessive and case morphemes are distinct (i.e., not formed simply by concatenation):

case	form	1SG	2SG	3SP
nom	—	-(I)м	-(I)ң	-(S)I
acc	-NI	-(I)мдI	-(I)ңдI	**-(S)Iн**
gen	-NIн	-(I)мдIн	-(I)ңдIн	-(S)IнIн
loc	-DA	-(I)мдA	-(I)ңдA	**-(S)IндA**
abl	-DAн	-(I)мдAн, **-(I)мAн**	-(I)ңдAн, **-(I)ңAн**	**-(S)IнAн**
dat	-GA	**-(I)мA**	**-(I)ңA**	**-(S)IнA**

- Trade-off:
  - morphophon. complicateder, morphotactics simpler
  - underlying form used: {S}{I}{n}
  - phonological rules delete {n}, {S} by context

Noun-noun compounds

- one type of N-N compunds: N2 has <px3> and related morphology

```
LEXICON N-INFL-3PX-COMPOUND
%<n%>:%>%{S%}%{I%}%{n%} GEN-POS ;

LEXICON Nouns
аба% ырайы:аба% ырай N-INFL-3PX-COMPOUND ;
                                   ! ”weather”
чакыруу% кагазы:чакыруу% кагаз N-INFL-3PX-
                    COMPOUND ; ! ”invitation”
```

Example output

Gloss

(1) Үстөл жана отургучтардын астын карап жатат, бирок Азамат аякта эмес.
table and chairs' underside looking is, but Azamat there not.
'*[She's] looking under the tables and chairs, but Azamat isn't there.*'

Output

```
^Үстөл/Үстөл<n><nom>$
^жана/жан<v><iv><prc_impf>/жана<adv>/жана<cnjcoo>$
^отургучтардын/отургуч<n><pl><gen>$
^астын/аст<n><px3pl><acc>/аст<n><px3sg><acc>$
^карап/кара<v><iv><gna_perf>/кара<v><iv><prc_real>/кара<v><tv><gna_perf>/кара<v><tv><prc_real>$
^жатат/жат<vaux><aor><p3><pl>/жат<vaux><aor><p3><sg>/жат<vaux><prc_irre>$ (intransitive verb forms removed)
^,/,<cm>$
^бирок/бирок<cnjadv>$
^Азамат/Азамат<np><ant><m><nom>$
^аякта/ал<det><dem>+жак<n><loc>/аяк<n><loc>/аякта<v><tv><imp><p2><sg>$
^эмес/э<cop><neg><p3><pl>/э<cop><neg><p3><sg>$
^./.<sent>$
```

Tagset

<n>	Noun	<p2>	Second person	<px3sg>	3rd person poss. (Singular)
<np>	Proper noun	<p3>	Third person	<px3pl>	3rd person poss. (Plural)
<v>	Verb	<ant>	Anthroponym	<neg>	Negative
<det>	Determiner	<dem>	Demonstrative	<aor>	Aorist
<cnjcoo>	Coord. conjunct.	<m>	Masculine	<imp>	Imperative
<cnjadv>	Adv. conjunct.	<sg>	Singular	<gna_perf>	Verbal adverb (Perfect)
<adv>	Adverb	<pl>	Plural	<prc_impf>	Participle (Imperfect)
<vaux>	Auxiliary verb	<nom>	'Nominative'	<prc_irre>	Participle (Irrealis)
<cop>	Copula	<gen>	Genitive	<prc_real>	Participle (Realis)
<iv>	Intransitive	<acc>	Accusative	<cm>	Comma
<tv>	Transitive	<loc>	Locative		

Morphophonology

Desonorisation

- {N} desonorises to д after a consonant
  алма-{N}{I} → алманы '*apple*–ACC'
  сыр-{N}{I} → сырды '*secret*–ACC'
- {L} desonorises to д after cons. of equal or lower sonority
  сыр-{L}{A}р → сырлар '*secret*–PL'
  кыз-{L}{A}р → кыздар '*girl*–PL'

```
”L Desonorisation”
%{L%}:д <=> :VoicedLowSonCns %>: __ ;

”N Desonorisation”
%{N%}:д <=> :VoicedCns %>: __ ;
```

Lenition

- Turn {y} into a harmonised high vowel when a vowel doesn't follow the following consonant
  мур{y}н → мурун '*nose*'
  мур{y}н+{I}м → мурдум '*my nose*'

```
%{y%}:Vy <=> [ :LastVowel :Cns* :Cns ]/[:0] __
          [ :Cns [ .#. | :Cns ] ]/[ :0 | %>:] ;
 where Vy in ( и ү и и ү ы ы у у ы у у )
LastVowel in ( и ү е э ө я а ё о ы ю у )
             matched ;
```

й+vowel letters

- [а о у] become [я ё ю] after й and й deletes
- й incorporated into the context of many rules
- + separate rules to change the characters
- + a rule to delete the original й

```
”Deletion of й before yoticised vowels”
й:0 <=> __ [ :YotVow ]/[ :0 | %>: ] ;
```

Evaluation

Test corpora

- **Kyrgyz Wikipedia** dump dated 2011-09-23
- All 2010 articles from **Radio Free Europe / Radio Liberty** (RFE/RL)'s Kyrgyz service (azattyk.org)
- both split into 10 equal parts; coverage calculated over each separately; standard deviation of mean calculated

Coverage measures

- **Naïve coverage** - percentage of surface forms in a given corpus receiving ≥ 1 analysis
  (surface forms may have missing analyses)
- **Mean ambiguity** - average number of analyses for each surface form found in analyed corpus

Coverage results (as of r36739)

corpus	tokens	known	cov.	amb.
Wikipedia	329,524	270,668	**82.1%**	**2.35**
RFE/RL	4,112,558	3,614,193	**87.9%**	**2.43**

Precision & recall

- **Precision** (of a form's analyses % correct): **97.32%**
- **Recall** (percentage of analyses provided by the transducer that are correct for a form, by comparing against a gold standard): **94.56%**

Future Work

- case changes for words with one root
  Финландия '*Finland*', финландиялык '*Finnish*'
- vowel harmony with abbreviations
  АКШ [акышы] '*USA*', АКШнын / *АКШтын
- vowel harmony with numbers
  100 [жүз], 100дүн
- compound verbs
- gerunds with mono-syllabic V-final verbs
  (*ж□□ / жеш < же-)
- Disambiguation
- Machine translation between Turkic languages

References