# Finite-State Morphologies & Text Corpora
## as Resources for Improving Morphological Descriptions


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

Tommi Pirinen
Ollscoil Chathair Bhaile Átha Cliath
tommi.pirinen@computing.dcu.ie

Jonathan North Washington
Indiana University
jonwashi@indiana.edu


## Morphological Descriptions

### What are they?

- Core part of grammatical descriptions
- Found in reference grammars and textbooks

### What they describe

- **Morphotactics** — the morphemes of a language and how they can be combined
- **Morphophonology** — the alternations between the various phonological and orthographical forms of each morpheme

### Reasons for being incomplete

- Restrictions of the medium (e.g., number of pages)
- Limitations of introspection and working with native speakers
- **Lack of automatic testing**

## 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>

### Turkic-language transducers we've made

- 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, Tuvan, Khakas, and more!

### Framework: HFST

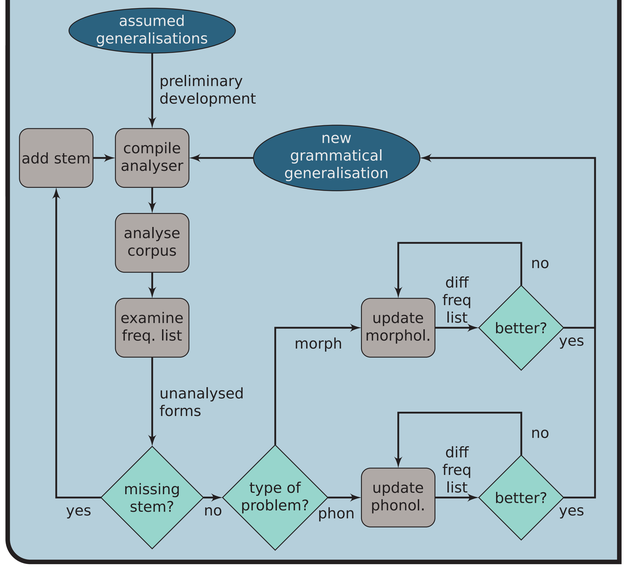
- Reimplements Xerox FST formalisms (`lexc` & `twol`)
- Also provides a wrapper around popular free/open-source FST toolkits: SFST, OpenFST, and Foma

### 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>

## Development cycle




## Example: Tuvan vowel harmony

### Tuvan vowel harmony

- High and low vowels harmonise in backness

  кижи<n><pl><acc>: кижилерни, ном<n><pl><acc>: номнарны
- High vowels also harmonise in roundness

  round: өг<n><acc>: өгнү, ой<n><acc>: ойну

  unround: үе<n><acc>: үени, ай<n><acc>: айны

### Initial implementation (high-vowels)

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

### Unanalysed forms in corpus

- forms of loanwords ансамбль 'ensemble' and рубль 'rouble': the following vowel is always front unround

| | | |
|---|---|---|
| 347 рубльди | 28 рубльден | 10 ансамблинге |
| 99 рубльге | 17 ансамбльдиң | 8 ансамбльди |
| 83 ансамблиниң | 15 ансамбльдер | 1 рубльдерни |
| 60 ансамбли | 14 ансамбль | 1 рублин |
| 54 рубльдиң | 11 ансамблин | 1 рубли |

### Revised generalisation

- After cons. cluster ending in ль, vowel always front unround
- New discovery? Appears not to be previously documented

### Revised implementation

```
"{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:  ]* _ ;
```

## Example: Categorisation

### Unrestricted 0-derivation

- Common overgeneralisation of adjective morphology:
  - all adjectives may act like nouns (i.e. take noun morphology)
  - all adjectives may be used as is as adverbs
- Commonly reported in Turkic language grammars

Somfai Kara (2002, pp. 28-29): Adjectives morphologically do not differ from nouns
All adjectives can be used as adverbs without any morphological changes

### Attested system

- Some adjectives may not be substantivised
- Some adjectives may not be used adverbially
- Some adjectives do not have comparative forms
- ...a range of "adjective classes" in most Turkic languages

### Implementation: proper categorisation

- only correct forms are analysed and generated (Tatar exs.)

| Type | Gloss | <adj>(<comp>) | <adj>(<comp>)<subst> | <adj>(<comp>)<advl> |
|---|---|---|---|---|
| A1 | 'good' | яхшы (яхшырак) | яхшы (яхшырак) | яхшы (яхшырак) |
| A2 | 'old' | иске (искерәк) | иске (искерәк) | — (—) |
| A3 | 'dead' | үле (—) | үле (—) | — (—) |
| A4 | 'basic' | төп (—) | — (—) | — (—) |

- Native speaker intuitions on morph. limitations often more restrictive than range of possible uses found in large text corpora.

## Example: Tuvan velar elision

### Previous descriptions

- Anderson and Harrison (1999, pp. 9, 22-23)

**1.3.1 Velar Elision and Vowel Lengthening**

/g/ > [ɣ] may appear in coda position of a monosyllabic word containing a short vowel: *öγ* 'yurt' *čaγ* 'fat'; it never appears in coda-position in a mono-syllable with a long vowel. When the 3SG POSS suffix is added (or any other vowel-initial suffix), [ɣ] is elided and the vowel of the word undergoes lengthening: *öö* 'yurt-3, *čaa* 'fat-3. The voiceless velar/uvular stop [k/q] patterns differently. First, [k/q] may appear finally in monosyllables containing either short or long vowels: *ök* 'button' *öök* 'glottis' *šaq* 'time'. Further, [k/q] does not elide when 3SG POSS affix is added: *öγü* button-3 *ööγü* glottis-3 *šaγı* time-3. However, in polysylabic words, both /k/ and /g/ pattern together, both undergo elision: *inek* 'cow' *inee* cow-3 *ayaq* 'bowl' *ayaa* bowl-3 *uruγ* 'daughter' *uruu* daughter-3 *biliγ* 'knowledge' *bilii* knowledge-3. Both velar phonemes have analogs in the archiphonemic/morphophonological system, see 2.1.1.4 and 2.2.3.1 below.

Disyllabic stems ending in a velar/uvular sound lose this in possessed forms, and lengthen the preceding vowel. Monosyllabic forms lose a voiced velar but retain a voiceless one (compare 'yurt' and 'glottis')

| (31) | SG | 1POSS | 2POSS | 3POSS | gloss |
|---|---|---|---|---|---|
| | *tavak* | *tavaam* | *tavaaŋ* | *tavaa* | 'plate' |
| | *uruʁ* | *uruum* | *uruuŋ* | *uruu* | 'child' |
| | *balik* | *baliim* | *baliiŋ* | *balii* | 'fish' |
| | *biliγ* | *biliim* | *biliiŋ* | *bilii* | 'understanding' |
| | *belek* | *beleem* | *beleeŋ* | *belee* | 'gift' |
| | *öγ* | *ööm* | *ööŋ* | *öö* | 'yurt' |
| | *ök* | *öγüm* | *öγüŋ* | *öγüŋ* | 'glottis' |

- Исхаков и Пальмбах (1961, pp. 117-118)

Конечный *к* в многосложных, а конечный *г* как в многосложных, так и в односложных словах перед аффиксами принадлежности обычно выпадают, создавая условия для образования долгих гласных (но иногда сохраняются):
*карак* 'глаз' — *караа* (< *карагы*), *инек* 'корова' — *инээ* (< *инеги*), *балык* 'рыба' — *балыы* (< *балыгы*), *бижик* 'письмо' — *бижии* (< *бижиги*), *кудук* 'колодец' — *кудуу* (< *кудугу*), *чүстүк* 'кольцо' — *чүстүү* (< *чүстүгү*), *даг* 'гора' — *даа* (< *дагы*), *суг* 'вода' — *суу* (< *сугу*), *одаг* 'костер' — *одаа* (< *одагы*).

### More complete characterisation

- /k/ ‹к› elides at the end of >1-σ stems intervocalically

  инэк+I ⟶ инээ, өк+I ⟶ өгү
- /g/ ‹г› elides at the end of any stem intervocalically

  өг+I ⟶ өө
- /ŋ/ ‹ң› elides at the end of **some** stems intervocalically

  **соң+I ⟶ соо**, чаң+I ⟶ чаңы, түң+I ⟶ түңү

### Correct forms in the grammars

But not part of the morphophonological descriptions.

(A. & H., p. 35) (*soŋ* >) *soo-* 'behind

(И. & П., p. 447) § 521. *Соң* 'конец'; *соо* < *соң* + аффикс принадлежности 3-го лица: *ажылдың соо* 'конец работы' (ср. *баштың соо*

### Unanalysed forms before implementation

- инээ, өгү, өө, соо, чаңы

| | | | |
|---|---|---|---|
| 1590 соонда | 6 инээниң | 2 өөңерге | 1 өөңнүң |
| 610 те | 6 соончe | 2 өөңер | 1 өөң |
| 153 өөнүң | 6 соонга | 2 өөмну | 1 өөвусче |
| 48 соондан | 3 инээм | 2 өөвустен | 1 өөвустүң |
| 46 өөм | 3 өөнден | 2 өөвуске | 1 өөвустү |
| 20 соон | 3 өөнде | 2 өөвус | 1 өөвусте |
| 18 өө | 3 өөмнүң | 1 инээңни | |
| 17 өөнге | 2 инээн | 1 инээмниң | |
| 9 соондагы | 2 инээ | 1 Инээм | |
| 8 өөн | 2 өөнче | 1 инээвистиң | |

### Initial implementation

```
"Intervocalic voiced velar deletion"
г:0 <=> :Vow/:0* _ [ %>:  :Vow ]/:0* ;

"Intervocalic voiceless velar deletion"
к:0 <=> :Vow/:0* _ [ %>:  :Vow ]/:0* ;
    except
        [ .#. | %- ] [ ( :Cns* ) ( :Vow* ) :Vow ]/:0 _ [ %>: :Vow ]/:0* ;
```

### Unanalysed forms after initial implementation

- инээ, өгү, өө, соо, чаңы

| | | |
|---|---|---|
| 1590 соонда | 48 соондан | 6 соончe |
| 610 те | 20 соон | 6 соонга |

### Second attempt at implementation

```
"Intervocalic voiced velar deletion"
Cx:0 <=> :Vow/:0* _ [ %>:  :Vow ]/:0* ;
    where Cx in ( г ң ) ;

"Intervocalic voiceless velar deletion"
к:0 <=> :Vow/:0* _ [ %>:  :Vow ]/:0* ;
    except
        [ .#. | %- ] [ ( :Cns* ) ( :Vow* ) :Vow ]/:0 _ [ %>: :Vow ]/:0* ;
```

### Unanalysed forms after second implementation

- инээ, өгү, өө, соо, чаңы

| | | |
|---|---|---|
| 610 те | 56 түңү | 5 түңүндeн |
| 203 бажыңынга | 47 чаңы | 4 чаңындан |
| 164 бажыңының | 44 чаңын | 4 чаңында |
| 102 бажыңы | 13 түңүн | 3 түңүнүң |
| 86 бажыңын | 12 чаңының | 2 түңүнгe |

### Final implementation

```
"Intervocalic voiced velar deletion"
Cx:0 <=> :Vow/:0* _ [ %>:  :Vow ]/:0* ;
    except
        :Vow _ [ %{ӕ%}:  :Vow ]/:0* ;
    where Cx in ( г ң ) ;

"Intervocalic voiceless velar deletion"
к:0 <=> :Vow/:0* _ [ %>:  :Vow ]/:0* ;
    except
        [ .#. | %- ] [ ( :Cns* ) ( :Vow* ) :Vow ]/:0 _ [ %>: :Vow ]/:0* ;
```

### Unanalysed forms after final implementation

- инээ, өгү, өө, соо, чаңы

| |
|---|
| 610 те |

- All forms of velar deletion are now analysed correctly!
- Can now move on to other unanalysed forms...

## Ongoing and future work

- Current annotation of a dependency treebank for Kazakh: a similar process that helps identify errors in morphological analyses and automatic disambiguation (constraint grammar)
  - Improved rule-based morphological disambiguation
  - Improved morphological analyser
  - A syntactic dependency treebank
- Future work:
  - Annotate more gold standard text in more Turkic languages
  - Research cross-linguistic techniques for analysing Turkic

## Further information

- Part of **Apertium Turkic** project: `http://wiki.apertium.org/wiki/Apertium_Turkic`
- Transducers available **live** (`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!

## References


Anderson, Gregory David and K. David Harrison (1999). *Tyvan.* Vol. 257. Languages of the World/Materials. München: Lincom Europa.

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.

Somfai Kara, Dávid (2002). *Kazak.* Vol. 417. Languages of the World/Materials. München: Lincom Europa.

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.

Исхаков, Ф. Г. and А. А. Пальмбах (1961). *Грамматика Тувинского языка: фонетика и морфология.* Москва: Издательство восточной литературы.