# A finite-state morphological analyser for Chukchi[1]

Vasilisa Andriyanets[a], Francis M. Tyers[a,b]

[a]Higher School of Economics, Moscow, Russia

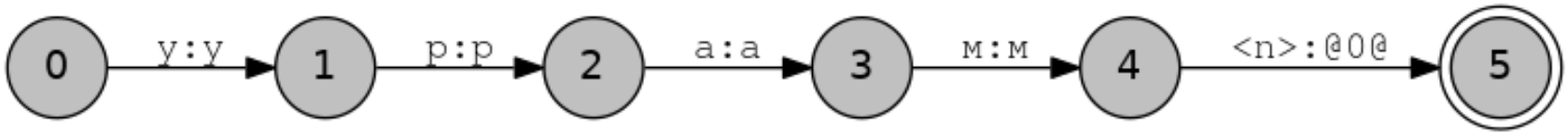
[b]Indiana University, Bloomington IN, United States

05.12.2018

[1]The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017 — 2018 (grant №17-05-0043) and by the Russian Academic Excellence Project «5-100»

# HFST

Helsinki Finite-State Transducer



# Methodology

- lexc + twol
- lexc for morphology, twol for phonology and morphophonology (and some derivational circumfixes)
- Similar toolkits have been used for other agglutinative and polysynthetic languages, such as:
  - Navajo (Hulden and Bischoff, 2008)
  - the Dene languages (Arppe et al., 2017)
  - Quechua (Rios, 2016)
  - Arapaho (Kazeminejad et al., 2017)

# Lexc

- Flag diacritics for inflectional morphology
- @FLAGTYPE.FEATURE.VALUE@
- @P.VPR.myt@1PL.S/A-:@P.VPR.myt@мыты sets a feature
- @R.VPR.myt@-1PL.S/O:@R.VPR.myt@ымык requires the feature to be set, making the myt_myk circumfix

# Twol for morphotactic constraints

- Easier for describing productive derivational morphology than flag diacritics

| | |
|---|---|
| Surface form: | ӈинэнимэтги |
| Morphotactic form: | ӈ{ы}>ин{Æ}>{R}{ы}>им{Æ}т>и |
| Lexical form: | [+caus]имэтык<v><tv>[+caus]<caus><neut><intn><a_sg2><o_sg1> |

2Ṡ/A.SUBJ-INV-CAUS-имэт-IRR-2/3SG.S

# Twol for morphotactic constraints

An example rule for this transducer looks like this:

```
"Causative"
%[%+caus%]:0 <=>          _ :* %[%+caus%]:0 ;
        %[%+caus%]:0 :* _ ;
```

- Remove all strings that do not have matching causative tags

# Twol

- vowel harmony
- vowel deletion
- affix allomorphy
- vowel reduction
- archiphonemic allomorphy
- orthography issues
- epenthesis

# Twol

```
"Vowel deletion for dominant"
! а г н о т в а:0 >:0 у:о
V:0 <=> .#. _ %>: :Vow ;
        :Cons _ (:0) (:0) %>: :Vow ;
        except
             :Cons _ (:0) (:0) %>: :V %>: ;
              %>: _ (:0) (:0) %>: :V ;
        where
             V in ( я э о ё а ) ;
```

# Orthography issues

- word-initial and post-vowel glottal stop is written as ’ *after* the next vowel
  **ʔ**e**ʔ**ewaɬ:ə**’**ə**’**вал̣
- ъ and ь signs stand for both glottal stop and /j/ in some contexts, just like in Russian
  ar**ʔ**apat:ap**ъ**апат
  racqew**j**an:расқэв**ъ**ян

Why orthography?

# Orthography issues

- after ԓ all vowels are written as jotised, and the glottal stop is ь (soft sign); after ч /e/ is written like e and glottal stop as ь
  wopqaneɬ**oo**ɬɣən:вопқанэлԓ**ёо**ԓгын
- /j/ and a vowel after are written as one jotised vowel for /a/, /e/, /u/, /o/, but not /i/
  **ja**ɣmaɬ:**я**гмаԓ
  teɣ**ji**ŋ:тэг**йи**ӈ

Why orthography?

# Results

So far, the analyser accounts for:

- the vast majority of morphophonology and orthography issues;
- nominal, pronominal, adjectival inflection and derivation; uninflected parts of speech;
- verbal inflection;
- verbal derivation (though substantially cut down to allow for incorporation);
- noun-to-verb, adjective-to-verb, adjective-to-noun incorporation;
- cross-part-of-speech derivation.

# Results

| Corpus | Tokens | Coverage | Mean ambiguity |
|---|---|---|---|
| Fairy tales (1) | 26,109 | 76.6 | 1.43 |
| Fairy tales (2) | 45,654 | 62.2 | – |
| Fiction (1) | 29,148 | 58.8 | – |
| Fiction (2) | 23,352 | 53.1 | – |
| Periodicals | 38,552 | 53.7 | – |
| Total: | 162,815 | 60.9 | 1.43 |

The texts for the corpora are publically available, provided to us by the linguists who work with Chukchi.

# Error analysis

| Category | Frequency | Percentage (%) |
|---|---|---|
| Missing stem | 82 | 75.2 |
| Missing morphotactics | 15 | 13.7 |
| Incorporation | 7 | 6.4 |
| Missing phonology | 2 | 1.8 |
| Typographical error | 3 | 2.7 |
| **Total:** | 109 | 100 |

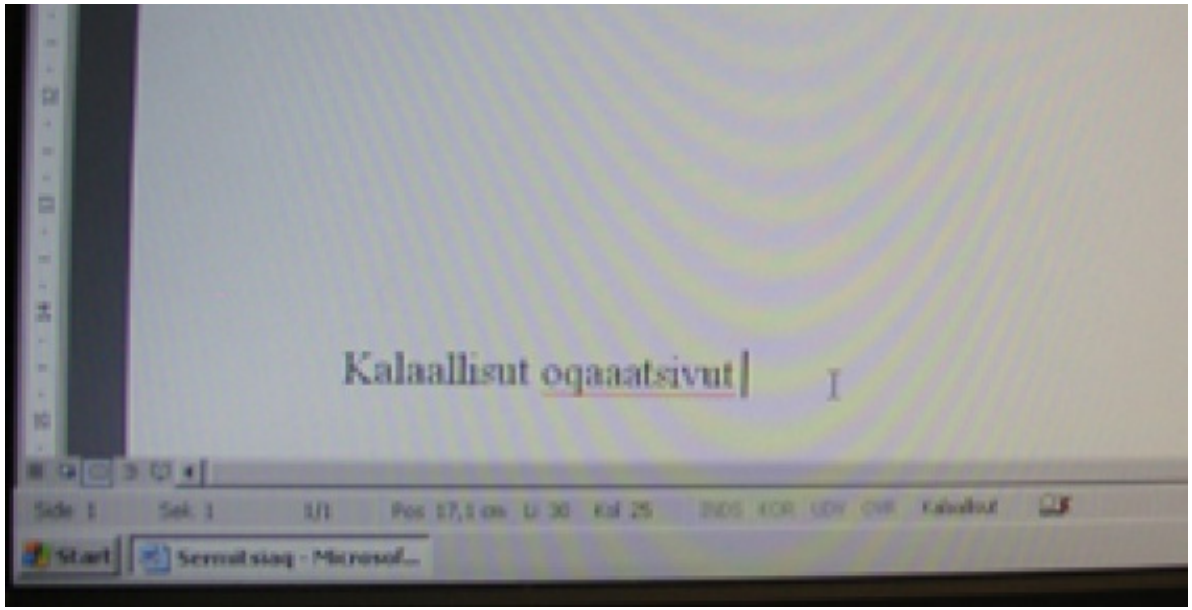
We took a sample of 100 random unique unanalysed forms

- most of the unparsed forms were unanalysed because of the lack of stems

# Future work

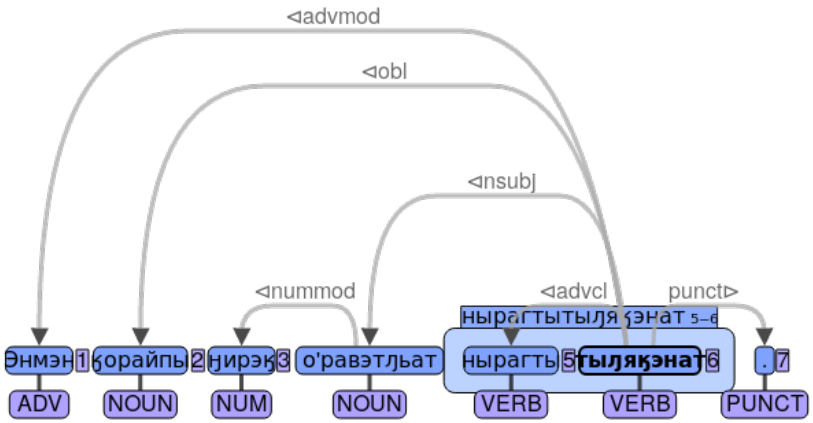
- guesser module
- spellchecker (involve community)
- treebank

# Future work




- guesser module
- spellchecker (involve community)
- treebank

# Future work




- guesser module
- spellchecker (involve community)
- treebank

# Concluding remarks

- First morphological analyser for Chukchi
- Reasonable performance but suffers from lack of complete stem lexicon
- Freely available:
  `https://github.com/BasilisAndr/chkchn` (GPL)

# Вэԯынкыӄун!
/weɬənkəqun!/

Comments/suggestions:

- Василиса Андриянец <blindedbysunshine@gmail.com>,
- Francis M. Tyers <ftyers@prompsit.com>