A prototype finite-state morphological analyser for Chukchi

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

[a]Higher School of Economics, Moscow, Russia

[b]Indiana University, Bloomington IN, United States

25.08.2018

Introduction

- Vasya is a 1st year MA student at HSE
- 3-month summer project (thanks Google!)

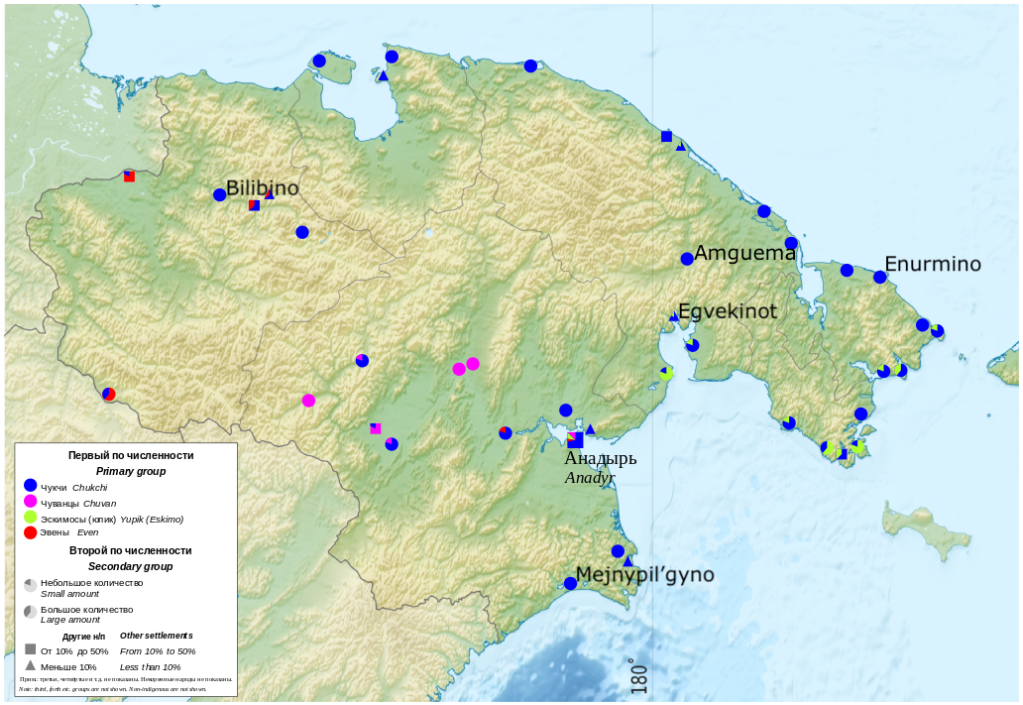
Chukchi

- Highly endangered minority language of the Russian Federation

Chukchi

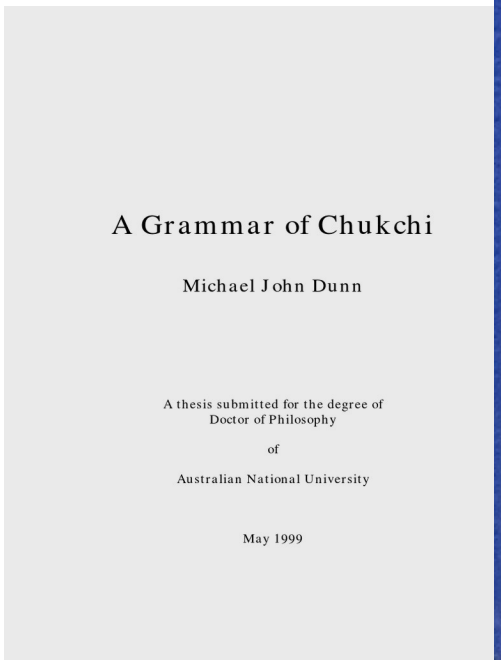
- Highly endangered minority language of the Russian Federation

Chukchi



- Most inhabited settlements: Anadyr, Egvekinot and Amguema

Grammars


A Grammar of Chukchi

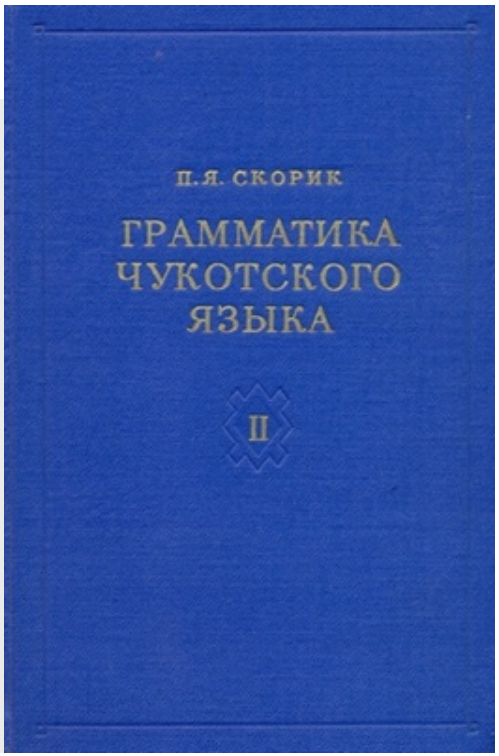
Michael John Dunn

A thesis submitted for the degree of
Doctor of Philosophy

of

Australian National University

May 1999



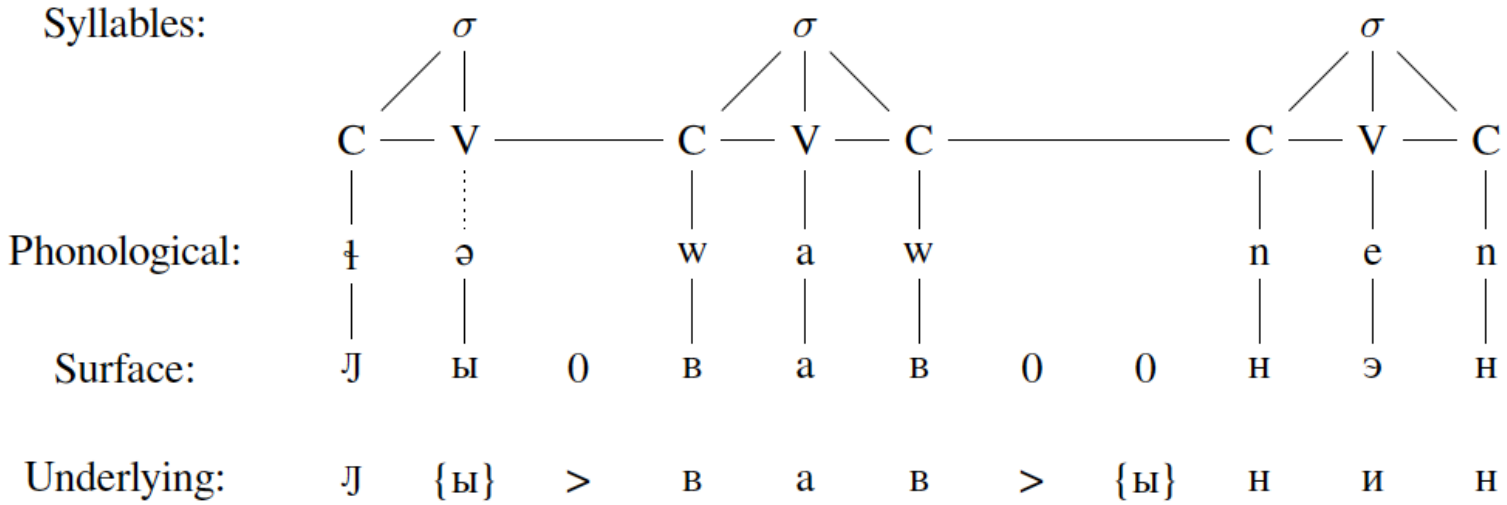
- Soviet-era Academy grammar (Skorik)
- Grammar in English by Michael Dunn

Chukchi

- Ergative—Absolutive
- Circumfixing polypersonal agreement
  - ne-ɬʔu-net
  - 3PL_O.3PL_A-see-3PL_O.3PL_A ‘they see them’
- Compounding
  - omwaamwə-rosɣ-etə
  - Amguema-other.bank-ALL ‘to the other bank of A.’
- Object/adverbial incorporation
  - memәɬ-kәɬeɬʔetәtku-tәk
  - seal-hunt-NEUT.AOR.S_PL2 ‘you are seal hunting’

Chukchi

- Rich circumfixation
- Schwa epenthesis



- Vowel harmony

Vowel harmony

− vowel harmony	*и* [i]	*э* [e]	*у* [u]
+ vowel harmony	*э* [e]	*а* [a]	*о* [o]

- vowels are grouped and only vowels from the same group can occur in one word
- two kinds of *э* [e] that behave differently, but phonetically identical

Vowel harmony

(1)	Recessive	л	ы	л	е	к	э	л	и	+	г	ы	п	ы	= лылекэли- + -гыпы
					↓		↓		↓						
	Dominant	л	ы	л	я	к	а	л	э	+	г	ы	п	ы	= лылякалэгыпы
(2)	Recessive	а	й	в	а	н	+	у							= айван- + -у
								↓							
	Dominant	а	й	в	а	н	+	о							= айвано

- vowels are grouped and only vowels from the same group can occur in one word
- two kinds of э [e] that behave differently, but phonetically identical

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@:@P.VPR.myt@мыты sets a feature
- @R.VPR.myt@<s_pl1>:@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>

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
  ʔe**ʔ**ewaɬ:ə’ə’вал
- 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**лгын
- /j/ and a vowel after are written as one jotised vowel for /a/, /e/, /u/, /o/, but not /i/
  jaɣmaɬ:**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);
- incorporation;
- cross-part-of-speech derivation.

Dictionary

Word class	**Tag**	**Entries**
Verb	<v>	3825
Noun	<n>	2178
Adverb	<adv>	2145
Pronoun	<prn>	927
Conjunctions	<cnjcoo>/<cnjadv>	788
Interjection	<ij>	621
Adjective	<adj>	242
Participles	<ptcp>	176
Postpositions	<post>	26
Total:		10,928

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.

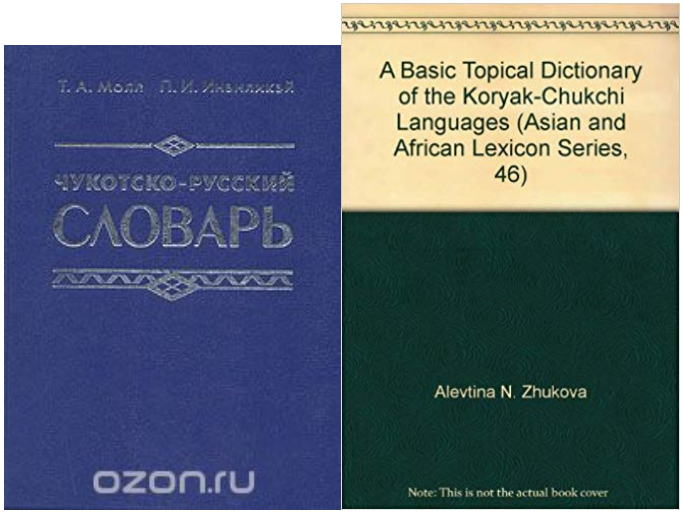
Error analysis

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

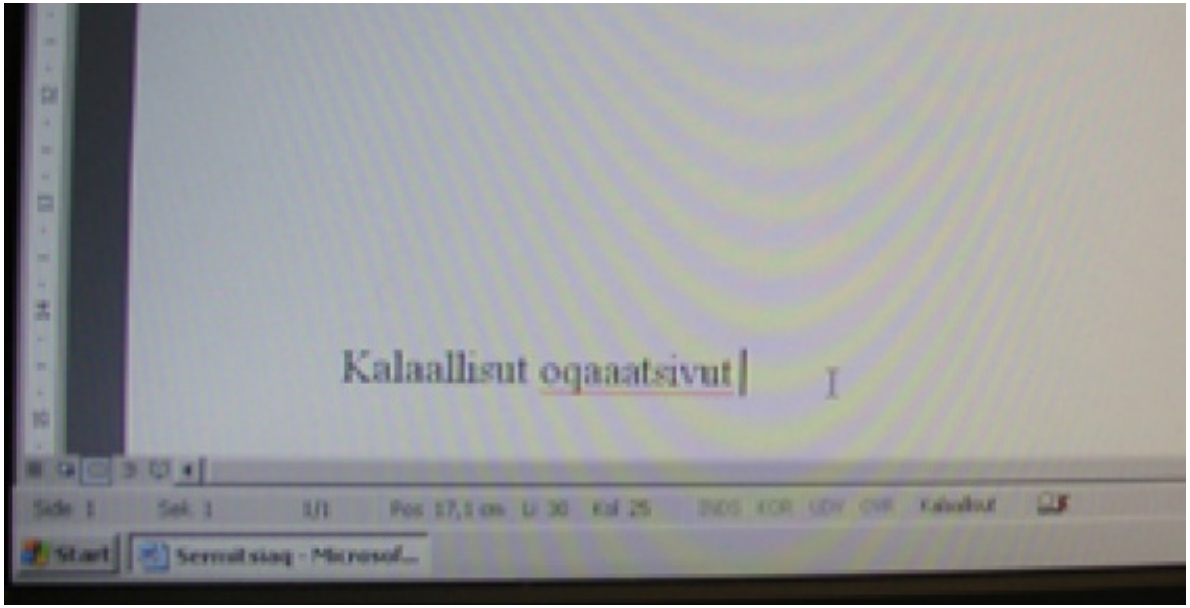


- extend the vocabulary
- guesser module
- spellchecker (involve community)
- Amguema dialect
- treebank

Future work

- extend the vocabulary
- guesser module
- spellchecker (involve community)
- Amguema dialect
- treebank

Future work



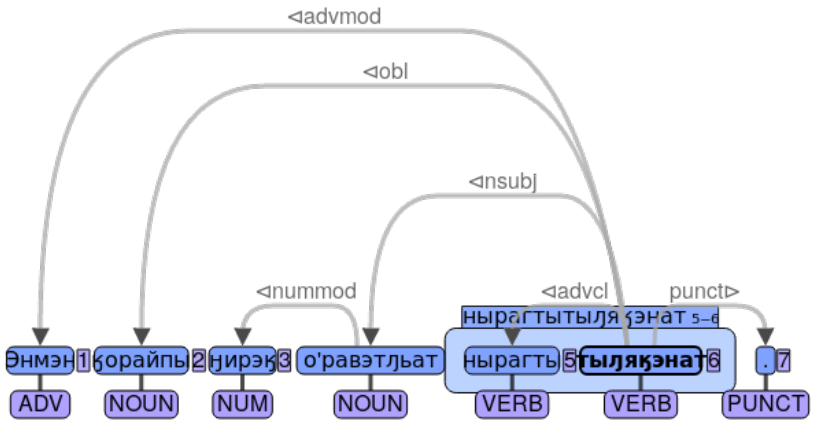
- extend the vocabulary
- guesser module
- spellchecker (involve community)
- Amguema dialect
- treebank

Future work

Amguema dialect has several phonetic differences to the standard, affixes look slightly different, but most importantly, the Amguema dialect is described to have ”lexical affixes” or affixes that bring a verbal meaning ”to take”, ”to make”, ”to catch” etc.

- extend the vocabulary
- guesser module
- spellchecker (involve community)
- Amguema dialect
- treebank

Future work



- extend the vocabulary
- guesser module
- spellchecker (involve community)
- Amguema dialect
- 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)

Вэлынкыӄун!

/weɬənkəqun!/

Comments/suggestions:

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