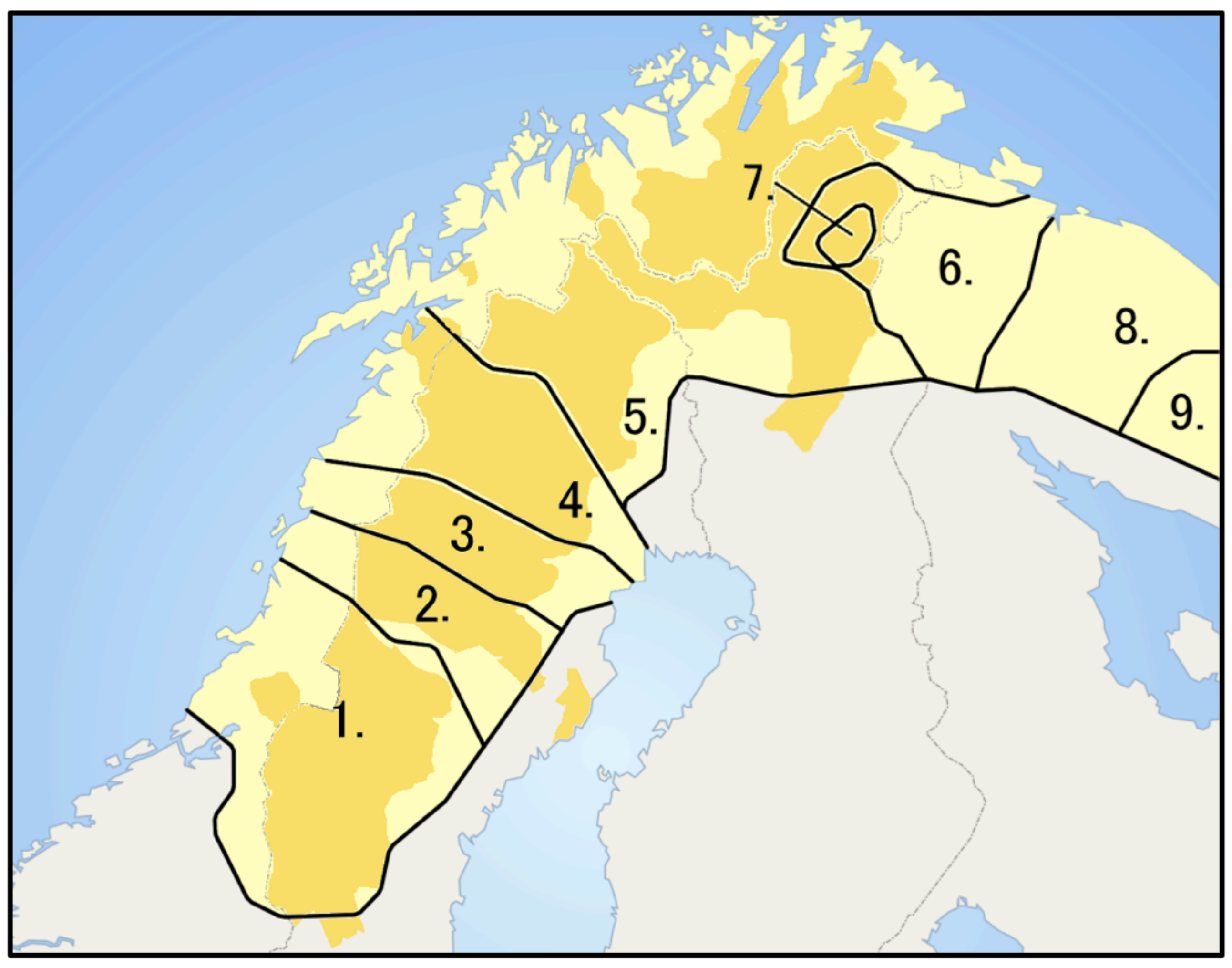
Modelling the Inari Saami morphophonology as a finite state transducer

Lene Antonsen & Trond Trosterud & Marja-Liisa Olthuis & Erika Sarivaara

The Second International Workshop on Computational Linguistics for Uralic Languages, Szeged, January 2016

Abstract

We present a finite-state transducer for Inari Saami, a language with a complex and not too well documented morphophonology. Modelling the grammar as a finite state transducer gives more insight in the morphophonology, and the resulting program will be the foundation of all future Inari Saami language technology applications.



The Saami languages (Inari Saami = No. 7)

Inari Saami at a glance

Inari Saami is an Uralic language with:

- Nine cases for nouns and adjectives
- 3 x 3 person-number inflection for verbs and possessed nouns
- Four moods for the verbs
- Tense system as in the Northern European Sprachbund
- Negation verb inflected for person, but not for tense
- An orthography phonological in nature, giving priority to representing wordform rather than word structure

The FST

- Lexicon: 13,000 nouns, 5,500 verbs and 2000 adjectives
- Inflection: 109 continuation lexica for nouns and 63 for verbs and 102 for adjectives
- Morphophonology: 106 rules with 274 distinct contexts, distributed over phonological phenomena as follows:

vowel centre	consonant centre	stem vowel	stem consonant	suffix	adding hyphen
154	75	27	10	7	1

The publicly available 1.6M word Inari Saami corpus, mainly collected by Ilmari Mattus, has been important when building the analyser.

Modelling morphophonology by combining lexc and twolc

The inflection mechanism is illustrated by two three-syllabic nouns. The genitive is used as lexc-stem because it is less ambiguous with respect to alternations, than the nominative form.

Sg Nom	Sg Gen	lexc-stem	Gloss
eebir	*iäbbár*	`iäbbár^ÁI`	bucket
lyeme	*luámmán*	`luámmá^SVn^ÁE`	cloudberry

The lexc-code
`+N+Sg+Nom:^EWG^RLEN # ;`
triggers these alternations in twolc:

- stem vowel `á:i` and `á:e`
- consonant gradation `bb:b` and `mm:m`
- vowel centre `iä:ee` and vowel centre `uá:ye`

```
"Monophthongisation rule iä:ee 1"
i:e <=> _ ä: Cns:* Vow: Cns [%^ÁI:|%^ÁE:] [%^WG:|%^EWG:] ([%^RLEN:|%^RVSH:]) [%>|.#.] ;

"iä:e and iä:ee rule 2 and Diphthongisation i5ä to ie"
ä:e <=>  i:e _ Cns:+ :i Cns: %^ÁI: [%^WG:|%^EWG:] %^RLEN: [%>|.#.] ;

"From strong to weak, part 1 cc:s, rr:r"
Cy:0 <=>    Cx:  _  Vow:+ (Cns:+) ([%^ÁI:|%^ÁE:]) (%^FCD:) %^EWG: ;
       where Cx in ( b c đ d g k l m n ŋ r p  p r s š v z ž)
             Cy in ( b c đ d g k l m n ŋ r p p4 r s š v z ž)
        matched ;
```

By looking beyond the triggers, one finds a contextual pattern for three-syllabic nouns. A long vowel centre (e.g. *aa*) is connected to the stem vowel change *á:i*-change, and a short vowel centre is connected to the stem vowel change *á:e*. These alternations could thus be triggered by ˆWG, which elsewhere is used for consonant gradation, instead of the special trigger ˆEWG. The diphthong then has to be marked for length in the lexc-stem.

Vocabulary coverage

We evaluate coverage in different frequency cohorts.

Wordform cohort	Coverage
0-1000	100.0 %
1000-2000	96.2 %
2000-3000	94.0 %
3000-4000	92.5 %
4000-5000	88.4 %
5000-6000	87.7 %
6000-7000	87.5 %
7000-8000	86.5 %
8000-9000	83.5 %
9000-10000	80.2 %
Whole corpus	92.0 %

At the present stage, the lexical coverage is best for the core vocabulary, and it drops markedly in the ninth 1000-cohort.

Coverage in running text

Analysing the unrecognised words in a 1.6M word corpus.

Type	Coverage
Recognised	92.0 %

Distribution of the 8 % unrecognised words is as follows:

Type	Coverage
Proper nouns	18.1 %
Finnish citations	13.1 %
Px and V-derivation	12.5 %
Non-words	5.7 %
FST lacunas	53.6 %

Half of the missing list consists of loans, names, and derivations and Px, categories still missing at the time of writing. The rest is misspellings and the long tail of missing words.

Consonant gradation and consonant lengthening

Sg Nom	Sg Gen ˆWG	Sg Ill ˆCSHˆRLEN	Sg Com ˆWGˆCLEN	lexc-stem	gloss
kukká	*kuká*	*kuukán*	*kukkáin*	`ku^RVkká`	'flower'
sukká	*suhá*	*suukán*	*suhháin*	`su^RVkk4á`	'sock'

ˆWG triggers *kk:k* and *kk4:h* alternation, whereas ˆCLEN lengthens the weak consonant in Comitative. ˆCSH shortens the consonant, and ˆRV gives compensatory lengthening of the root vowel.

```
"Consonant shortening and gradation kk:k and kk4:k"
k:0 <=> Vow: _  [k:|k4:] Vow: [%^WG:|%^CSH:] ([%^RLEN:|%^RVSH:]) %> ;

"Gradation for kk4"
k4:h <=> Vow: k:  _  Vow: (Cns:) %^WG: ;

"kk:hh gradation for kk4"
k:h <=> Vow: _ k4:h Vow: (Cns:) %^WG: %^CLEN: ;

"Root vowel u lengthening"
%^RV:u <=> u _ Cns:+ Vow: (Vow:) (Cns:+) Triggers:* %^RLEN: ;
```

TWOLC triggers

Trigger	changing	how	comments
ˆRLEN	vowel centre	lengthening	e.g. *a* to *aa*
ˆRVSH	vowel centre	shortening	e.g. *aa* to *a*
ˆVBACK	vowel centre	quality	e.g. *ä* to *a*
ˆVHIGH	vowel centre	quality	e.g. *á* to *i*
i2	vowel centre	quality	e.g. *iä* to *e*
ˆCLEN	cons.centre	lengthening	e.g. *h* to *hh*
ˆWG	cons.centre	gradation	e.g. *tt* to *đ*
ˆSLEN	stem vowel	lengthening	e.g. *e* to *ee*
ˆSVSH	stem vowel	shortening	e.g. *ee* to *e*
ˆSVLOW	stem vowel	quality	e.g. *u* to *o*
ˆÁI	stem vowel	quality	with ˆEWG: *á* to *i*
ˆÁE	stem vowel	quality	with ˆEWG: *á* to *e*
u2	stem vowel	quality	e.g. *u* to *o* and
	vowel centre	quality	*uá* to *oo*
ˆCSH	cons.centre	shortening	e.g. *tt* to *t* and
	vowel centre	shortening	*aa* to *a*
ˆEWG	cons.centre	gradation	e.g. *tt* to *đ* and
	stem vowel	quality	*á* to *i* or *e*
ˆEA	vowel centre	quality	e.g. *e* to *iä* and
	stem vowel	quality	*i* to *á*
ˆFCD	final consonant	deletion	e.g. delete *t*

Evaluating analysis and generation

We made a gold corpus of 276 random correctly spelled words, ran them through our analyser, and checked each analysis manually. We then ran the list through the analyser (revision 124755), and estimated precision and recall, as explained below.

Evaluation of analysis	%
Precision	91.7 %
Recall	83.4 %

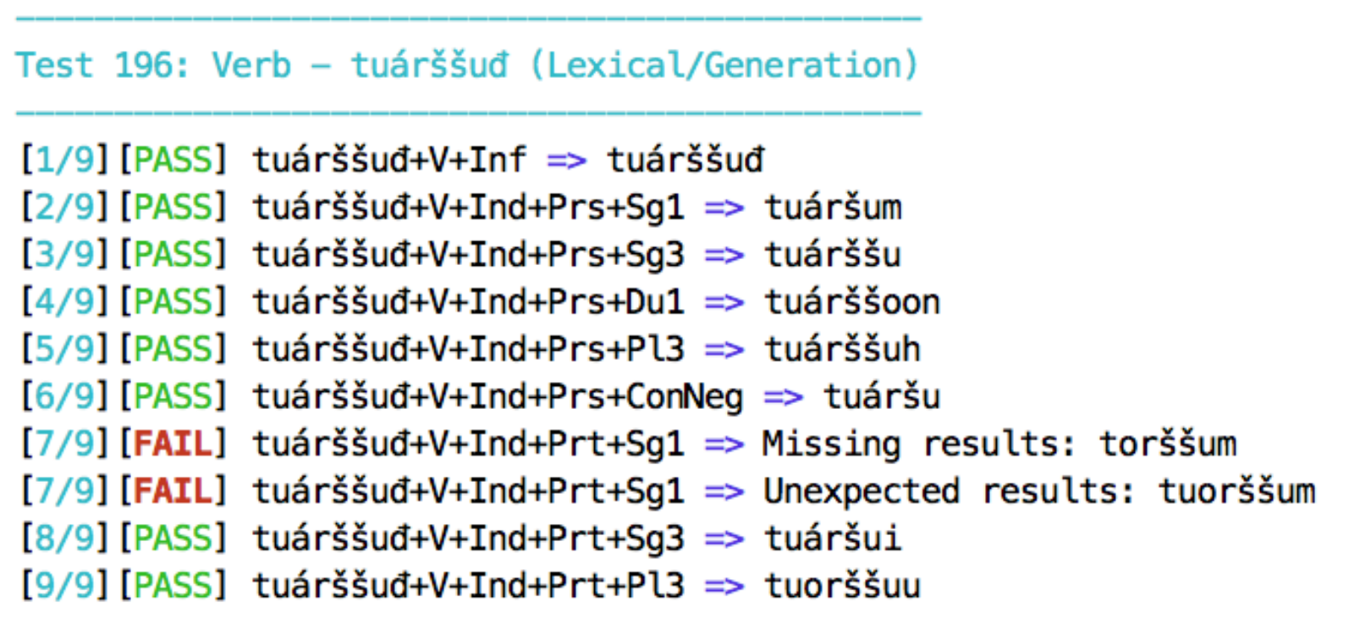
Procedure for estimating precision and recall

Precision the number of analyses which were found in both the output from the morphological analyser and the gold standard, divided by the total number of analyses output by the morphological analyser.

Recall the total number of analyses found in both the output from the morphological analyser and the gold standard, divided by the number of analyses found in the morphological analyser plus the number of analyses found in the gold standard but not in the morphological analyser.

Regression testing

We built a test suite for 663 nouns, 364 verbs and 82 adjectives, 20,836 pairs on the format *lemma + grammatical tags : inflected forms* The test suite tested both analysis and generation, and was an important tool for regression testing during development.



```
------------------------------------------------
Test 196: Verb - tuáršsuđ (Lexical/Generation)
------------------------------------------------
[1/9][PASS] tuárššuđ+V+Inf => tuárššuđ
[2/9][PASS] tuárššuđ+V+Ind+Prs+Sg1 => tuáršum
[3/9][PASS] tuárššuđ+V+Ind+Prs+Sg3 => tuárššu
[4/9][PASS] tuárššuđ+V+Ind+Prs+Du1 => tuárššoon
[5/9][PASS] tuárššuđ+V+Ind+Prs+Pl3 => tuárššuh
[6/9][PASS] tuárššuđ+V+Ind+Prs+ConNeg => tuáršu
[7/9][FAIL] tuárššuđ+V+Ind+Prt+Sg1 => Missing results: torššum
[7/9][FAIL] tuárššuđ+V+Ind+Prt+Sg1 => Unexpected results: tuorššum
[8/9][PASS] tuárššuđ+V+Ind+Prt+Sg3 => tuáršui
[9/9][PASS] tuárššuđ+V+Ind+Prt+Pl3 => tuorššuu
```

Conclusion

The FST is a comprehensive model of Inari Saami grammar. Being rule based, the model offers explicit insight into the morphophonology. The resulting transducer is put into use as a generator for rule-based machine translation from North to Inari Saami.

Beyond that, it may also be used for other purposes, such as corpus analysis (*http://gtweb.uit.no/corp*), e-dictionaries, spell checkers and pedagogical programs. This we leave for the future.

http://giellatekno.uit.no

LaTeX TikZposter