

Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries

Ryan Johnson, Trond Trosterud and Lene Antonsen,
Centre for Saami Language Technology
<http://giellatekno.uit.no/>



Presentation at NODALIDA 2013, May 22-24, Oslo, Norway

Contents

Introduction

Morphologically Sensitive Dictionaries

Finite State Transducers

Evaluation

Conclusion

Introduction

- ▶ Starting point: Enriching our ICALL learning environment with a dictionary
- ▶ Design requirement:
 - ▶ Handle morphology
 - ▶ No need for downloading and installation

Introduction

- ▶ Starting point: Enriching our ICALL learning environment with a dictionary
- ▶ Design requirement:
 - ▶ Handle morphology
 - ▶ No need for downloading and installation
- ▶ Result: dictionaries for
 - ▶ North and South Saami
 - ▶ and several other Uralic languages

Introduction

- ▶ Starting point: Enriching our ICALL learning environment with a dictionary
- ▶ Design requirement:
 - ▶ Handle morphology
 - ▶ No need for downloading and installation
- ▶ Result: dictionaries for
 - ▶ North and South Saami
 - ▶ and several other Uralic languages
- ▶ ... in a variety of ways

Introduction

- ▶ Starting point: Enriching our ICALL learning environment with a dictionary
- ▶ Design requirement:
 - ▶ Handle morphology
 - ▶ No need for downloading and installation
- ▶ Result: dictionaries for
 - ▶ North and South Saami
 - ▶ and several other Uralic languages
- ▶ ... in a variety of ways

Lemmas are rare in running text

	North Saami	Finnish	Norwegian
Number of words in the text	252 461	45 144	64 994
Number of lemmas in the dictionary	99 071	94 111	38 983
Coverage	7,9 %	10,0 %	30,5 %

(Antonsen et al 2009)

The morphology is rich, and not always concatenative

The morphology is rich, and not always concatenative

- ▶ Inflection:

- ▶ Case: *niidii* → *nieida* girl, *gãatan* → *gãetie* house
- ▶ Person-tense: *eahtsa* → *iehtsedh* love, *bođii* → *boahtit* come

The morphology is rich, and not always concatenative

- ▶ Inflection:
 - ▶ Case: *niidii* → *nieida* girl, *gãatan* → *gãetie* house
 - ▶ Person-tense: *eahtsa* → *iehtsedh* love, *bođii* → *boahtit* come
- ▶ Derivation:
 - ▶ Passivisation: *juhkkojuvvui* → *juhkat* drink, *juohkit* share
 - ▶ Superlative: *nööremes* → *noere* young

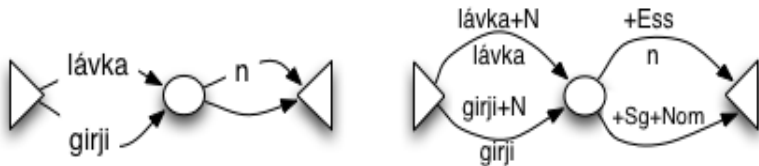
The morphology is rich, and not always concatenative

- ▶ Inflection:
 - ▶ Case: *niidii* → *nieida* girl, *gáatan* → *gáetie* house
 - ▶ Person-tense: *eahtsa* → *iehtsedh* love, *bođii* → *boahtit* come
- ▶ Derivation:
 - ▶ Passivisation: *juhkkojuvvui* → *juhkat* drink, *juohkit* share
 - ▶ Superlative: *nööremes* → *noere* young
- ▶ Compounding:
 - ▶ *bargojoavku* → *bargu* + *joavku* work + group
 - ▶ *bigkemegierkine* → *bigkedh* + *gierkie* building + stone
 - ▶ In a 1.1 mill word North Saami corpus, 26.7 % of the nouns are compounds, or 7.0 % of all words

The morphology is rich, and not always concatenative

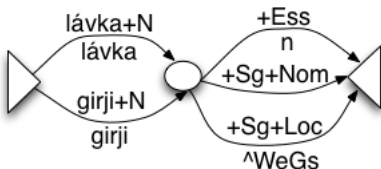
- ▶ Inflection:
 - ▶ Case: *niidii* → *nieida* girl, *gáatan* → *gáetie* house
 - ▶ Person-tense: *eahtsa* → *iehtsedh* love, *bođii* → *boahtit* come
- ▶ Derivation:
 - ▶ Passivisation: *juhkkojuvvui* → *juhkat* drink, *juohkit* share
 - ▶ Superlative: *nööremes* → *noere* young
- ▶ Compounding:
 - ▶ *bargojoavku* → *bargu* + *joavku* work + group
 - ▶ *bigkemegierkine* → *bigkedh* + *gierkie* building + stone
 - ▶ In a 1.1 mill word North Saami corpus, 26.7 % of the nouns are compounds, or 7.0 % of all words
- ▶ Cliticisation:
 - ▶ *bođiige* → *boahtit* + *ge* came + too

Finite state transducers (FST), compiled with lexc



girji+N: gir'ji GOAHTI ;
 lávka+N: láv'ka GOAHTI ;

Morphophonological transducer — twolc



lávka+N+Ess	lávkan
lávka+N+Sg+Nom	lávka
lávka+N+Sg+Loc	lávkkas
girji+N+Ess	girjin
girji+N+Sg+Nom	girji
girji+N+Sg+Loc	girjjis

vk	->	vkk	 	-	Vow*	^WeG	;
rj	->	rjj	 	-	Vow*	^WeG	;

^WeG: báhcahat

girji+N: gir'ji GOAHTI ;
lávka+N: láv'ka GOAHTI ;

Two different approaches to using FSTs with dictionaries

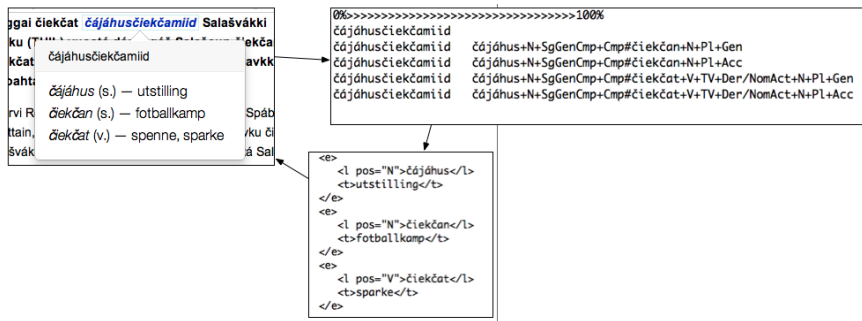
- ▶ generating a list of wordforms pointing to their respective lemma articles
- ▶ analysis via FST at runtime and using outcome for lookup in dictionary

Wordform dictionaries (static)

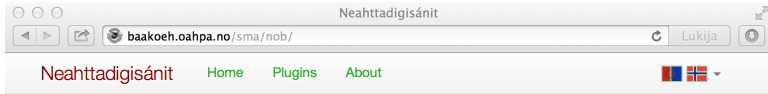
- ▶ Vuosttaš digisánit: 9999 lemma → 110 000 wordforms
- ▶ Voestes digibaakoeh: 10 657 lemma → 85 000 wordforms

Dynamic FST dictionaries

FST-analysis and translation via a web service



The dictionary understands inflected forms



South Sámi → Norwegian

Norwegian → South Sámi

Other dictionaries

South Sámi → Norwegian (⇌ Swap)

iehtsedh (v., i)

1. elske, være glad i

Manne datnem eahtsam.

Jeg elsker deg.

2. **engste seg for noen**

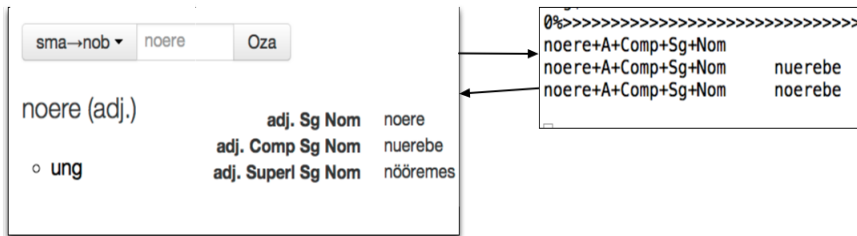
Datneste eahtsam.

Jeg er engstelig for deg.

eahtsa is a possible form of ...

iehtsedh
v. TV Ind Prs Sg3

... and may generate word paradigms via FST



Necessary adjustments

- ▶ Add lexicalised forms to the transducer:
gaskabeaivi ‘noon’ – *gaskabeaivvit* ‘dinner’
gávdni ‘guts’ – *gávnnit* ‘bedclothes’
 → *gaskabeaivvit*+N+Pl, *gávnnit*+N+Pl
- ▶ Resolve homonymy in the analyser:
vuovdi / *vuov***ddi** vs. *vuovdi* / *vuov***di** ‘forest’ vs. ‘salesman’
beassi / *beass***i** vs. *beassi* / *beas***i** ‘birchbark’ vs. ‘nest’
 → *vuovdi*+N+NomAg, *beassi*+N+G3
- ▶ Clean up the generated paradigms when there are variants:
fievrridin – *fievrr***edin** ‘transport’
 → *fievrridit*+v1+V, *fievrridit*+v2+V

- ▶ *http://sanit.oahpa.no*
 - ▶ North Saami ↔ Norwegian, Finnish
- ▶ *http://baakoeh.oahpa.no*
 - ▶ South Saami ↔ Norwegian
- ▶ *http://sanat.oahpa.no*
 - ▶ Livonian ↔ Finnish, Estonian, Latvian
 - ▶ Kven ↔ Norwegian
 - ▶ Ingrian ↔ Finnish
- ▶ *http://valks.oahpa.no*
 - ▶ Erzya ↔ English, Finnish, Russian
 - ▶ Mokša ↔ Finnish
- ▶ *http://muter.oahpa.no*
 - ▶ Eastern Mari ↔ Finnish
 - ▶ Western Mari ↔ Finnish
- ▶ *http://kyv.oahpa.no*
 - ▶ Komi ↔ English, Finnish
- ▶ *http://vada.oahpa.no*
 - ▶ Nenets ↔ English, Finnish

Neahttadigisánit – Implementation

- ▶ Python with Flask framework (<http://flask.pocoo.org>)
- ▶ Twitter Bootstrap
(<http://twitter.github.io/bootstrap/>)
- ▶ JavaScript with jQuery (<http://jquery.com/>)

User interface (also works on a mobile phone)

nordsamisk → norsk (⇌ Swap)

váccát

vázzit (v.) –

1. gå, spasere

Odne mun váccán johtilit bargui, go mus lea hoahppu.
I dag går jeg fort til jobb, fordi jeg har det travelt.

váccát is a possible form of ...

vázzit
v. ind. pres. 2.p.ent.



Reading dictionary: read and alt-click

Neahttadigisánit – Neahttadigisánit Reader

Read with Neahttadigisánit

Neahttadigisánit Home Plugins About

Neahttadigisánit Reader

WARNING!! We have problems with Internet Explorer 8!

Bookmarklet

Drag and drop the following link to your Bookmark toolbar, then click it when you are on a page you wish to read. The service will automatically be installed, and you will see a menu icon on the left-hand side of the page.

Read with Neahttadigisánit

In order to look up a word, hold Alt or Option (⌘) and double click the word. The service will connect to this website and return dictionary entries after a brief pause.

To change the dictionary, click the menu icon.


Includes

- Nenets → English
- Nenets → Finnish

If you are looking for other languages, please see our [other dictionaries](#).

javascrpt:(function() {var e="http://vada.oahpa.no",t="0.0.3",n=document.createElement("link");n.href=e+"/static/css/jquery.neahttadigisanit.css"

Eastern Mari



Википедий
Энциклопедий

А

- Лаштык
- Шка
- Тартыш
- Трлатымаш-влак
- Чокым лаштык
- Полшык
- Модмо вер
- Надыр
- Узгар-влак
 - Тышке кондышо кылвер-влак
 - Ваш кылдалтше трлатымаш-влак

 trondtr [канашымаш](#) [кельштарымаш](#) [эскерымаш](#) [лүмер](#)

Лаштык

[Канашымаш](#)

Лудаш

[Трлаташ](#)Пагален [Ўжына](#)

кажне айдемылан трлаташ поч

Кызыт Википедийште [марий](#) йылме дене возымо **3448** лаштык уло.[Кузе у лаштыкым ышташ](#) [Полыш](#)

Ўжына

Ўжаш (v.) — kutsua

Поро кече! [Марий](#) эрыкан Википедийыш пагален ўжына!Добро пожаловать в Википедию на [марийском](#) языке!Welcome to [Mari](#) Wikipedia!

Йолташ, кумылет уло гын, тыят [Марий](#) Википедийым пойдарен кертат. У теме почеш материалым савыкташ але савыктыме

Сай статья

Эрик Сапаев (4 Уярня
1022 йылмыш Ош, 25

Komi



Википедия
Мездмӧм энциклопедия

[Hjeldside](#)
[Wikarportal](#)
[Tult](#)
[Ste endringar](#)
[Tilfeldig side](#)
[Hjelp](#)
[Gåver](#)

▼ Verktøy
[Lenkjer hit](#)
[Relaterte endringar](#)
[Last opp fil](#)
[Spesialsider](#)
[Utskriftsversjon](#)

Медшӧр лист бок



Тайӧ субдоменсӧ видзӧны, медым вӧчны

Википедия

пытшкын **Коми кыв** вылын гижӧдъяс.

3878 гижӧд

Сямьд кӧ эм сёрнитны комиӧн, сьӧлӧм кӧ пӧсь да **вирӧн** тыр, гиж тэ комиӧн, комиӧн, комиӧн, медым сӧм ми вын! Тайӧ сайт вылын позь гижны унатор. Сӧмын гиж. К **вирӧн** колӧ.

Юриндалысь

вир (n.) — blood

Kven

Kainun institutti – Kvensk institutt oon perustannu Kväänin litteratuuripalkinon. Palkinon **jaethaan** tänä vuona Se ei ole suuri yllätys ette sen annethaan Alf Nilsen-Børsskogile joka oon kirjoittannu ensimmäiset ja tähään as – Børsskog oon ollu niin ko yksinäinen pääskynen kvääninkielisessä litte **jaethaan** esikuvana tulleevaisuusessa. Hänen kirjat oon uniikit ja tärkkee apu kvä paikalisesta kulttuurista, mutta samala rommaaniitten historiat oon maili niitä joitten äitinkieli oon kväänin kieli, Kainun institutti kirjoittaa. **jakkaat (v.)** — dele Børsskog oon julkaissu kolme rommaani ja nelje poeemi- eli diktikirjaa. Palkintosumma oon 50 000 kruunuu. Palkinon **jaethaan** Kainun institutissa 29. maikuuta.

Evaluation

Comparing the dictionaries: North Saami

FST dictionary and wordform (WF) dictionary: 9999 lemmas.
 (A lemma dictionary with 99 000 lemma covered 7,9 %)

	Translation		Partial translation		No translation		100 %
WF dictionary: all words	36 231	81.9 %			8025	18.1 %	44 256
FST dictionary: all words	39 920	90.2 %	549	1.2 %	3787	8.6 %	44 256
WF dictionary: different words	7874	57.8 %			5758	42.2 %	13 632
FST dictionary: different words	10 862	80.0 %	448	3.3 %	2322	17.0 %	13 632

Comparing the dictionaries: South Saami

FST dictionary and wordform dictionary: 10 657 lemmas

	Translation		Partial translation		No translation		100 %
WF dictionary: all words	44 989	72.7 %			15 268	25.3 %	60 037
FST dictionary: all words	53 295	88.8 %	475	0.8 %	6266	10.4 %	60 037
WF dictionary: different words	5015	41.4 %			7100	58.6 %	12 092
FST dictionary: different words	8039	67.0 %	308	2.6 %	3660	30.5 %	12 007

#SoMe dictionary (Social media)

#SoMe dictionary (Social media)

Jua, livzze buorre jus nu ain birgese dan aigge, ahte eamidat maid alo siiddas go manat leat unnit, nu ahte manat harjanit ju unnin leahket doppe ja besset de oahppat. Muhto oluhat darbbasit ...

#SoMe dictionary (Social media)

Jua, livzze buorre jus nu ain birgese dan aigge, ahte eamidat maid alo siiddas go manat leat unnit, nu ahte manat harjanit ju unnin leahket doppe ja besset de oahppat. Muhto oluhat darbbasit ...

Jua, livžže buorre jus nu ain birgeše dan áigge, ahte eamidat maid alo siiddas go mánát leat unnit, nu ahte mánát hárbánit ju unnin leahket doppe ja besset de oahppat. Muhto oluhat dárbbašit ...

#SoMe dictionary (Social media)

Jua, livzze buorre jus nu ain birgese dan aigge, ahte eamidat maid alo siiddas go manat leat unnit, nu ahte manat harjanit ju unnin leahket doppe ja besset de oahppat. Muhto oluhat darbbasit ...

Jua, livžže buorre jus nu ain birgeše dan áigge, ahte eamidat maid alo siiddas go mánát leat unnit, nu ahte mánát hárbánit ju unnin leahket doppe ja besset de oahppat. Muhto oluhat dárbbašit ...

- ▶ From the Facebook group *Ártegis ságat*: 20 % of which is written without a North Saami keyboard, on either a desktop or mobile device.

#SoMe dictionary (Social media)

Jua, livzze buorre jus nu ain birgese dan aigge, ahte eamidat maid alo siiddas go manat leat unnit, nu ahte manat harjanit ju unnin leahket doppe ja besset de oahppat. Muhto oluhat darbbasit ...

Jua, livžže buorre jus nu ain birgeše dan áigge, ahte eamidat maid alo siiddas go mánát leat unnit, nu ahte mánát hárjánit ju unnin leahket doppe ja besset de oahppat. Muhto oluhat dárbašit ...

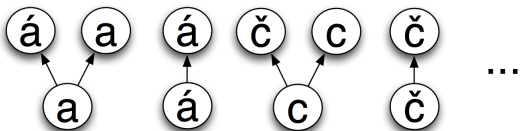
- ▶ From the Facebook group *Ártegis ságat*: 20 % of which is written without a North Saami keyboard, on either a desktop or mobile device.

Our answer:

A North Saami FST dictionary where the text is analysed with an ordinary FST, and in addition with an FST adjusted to the #SoMe writing convention.

#SoMe-FST

If the word in the text contains **c**, then the intended word may also contain **č** ...



Cell phone

Neahttagisáit 

Nordsamisk (#SoMe) ↔ Norsk (⇔ Snu)

Søk

čáhci (s.) –

1. vann, vatn

cazi er en mulig form av ...*čáhci*

subst. ent. akk.

subst. ent. gen.

#SoMe sátnegirji

Results with an ordinary FST and with the #SoMe FST

	Translation		Partial translation		No translation		100 %
Ord. FST all words:	50 263	76.3 %	250	0.4 %	15 364	23.3 %	65 877
#SoMe-FST all words:	54 197	80.6 %	315	0.5 %	12 753	19.0 %	67 265
Ord. FST different words:	8813	50.8 %	224	1.3 %	8326	48.0 %	17 363
#SoMe FST different words:	10 596	59.8 %	286	1.6 %	6825	38.5 %	17 707

Including Nynorsk in a Norwegian → Saami dictionary

Why Nynorsk in a Norwegian → Saami dictionary?

Why Nynorsk in a Norwegian → Saami dictionary?

Lov om stadnamn:

Saker som gjeld gards- og bruksnamn, skal eigar eller festar få tilsendt direkte. Saker som gjeld samiske eller kvenske stadnamn, skal dessutan *sendast* til lokale organisasjonar med særleg tilknytning til saka. I tillegg skal saka kunngjerast i mir aviser som er alminneleg lesbare på staden, eller derast kjend på annan høveleg måte. Saker som gjeld andre namn, skal kunngjerast

Namn _____ nn på norsk, samisk og kvensk.

Med n _____ ld av lov, gjeld reglane i forvaltningsloven kapittel IV, V og VIII ikkje for sak
lova her.

sendast

sende (verb) — såddet

Why Nynorsk in a Norwegian → Saami dictionary?

Lov om stadnamn:

Saker som gjeld gards- og bruksnamn, skal eigar eller festar få tilsendt direkte. Saker som gjeld samiske eller kvenske stadnamn, skal dessutan **sendast** til lokale organisasjonar med særleg tilknytning til saka. I tillegg skal saka kunngjerast i mir aviser som er alminneleg lesbare på staden eller derast kjend på annan høveleg måte. Saker som gjeld andre namn, skal kunngjerast i

sendast

Namn på norsk, samisk og kvensk.

sende (verb) — såddet

Med lov om stadnamn, gjeld reglane i forvaltningsloven kapittel IV, V og VIII ikkje for sak
lova her.

Jon Todal: Samisk språk i Svahken sijte:

8 barnehagebarn og **elevar** som får opplæring i eit lite språk som er i stor fare for å dø ut - er det noko å skrive bok om og kan **elev** vere noko å skrive bokmelding om? Det kan det vere, når røynslene frå dette p
driv an

elevar

elev (subst.) — oahppi

Boka h
Nordisk Samisk Institutt. Handlinga foregår på Austlandet, men innafor denne landsdelen så langt ein kan

Extended Norwegian dictionary

Bokmål - Saami FST dictionary tested on Nynorsk text (Wikipedia)

		FST coverage		Dictionary coverage	
Bokmål text	All Bokmål varieties	703 950	93.36 %	1 644 286	84.50 %
Nynorsk text	Conservative Bokmål	2 191 428	79.34 %	3 504 733	66.96 %
	All Bokmål varieties	1 849 654	82.56 %	3 206 796	69.77 %
	Bokmål with Nynorsk	1 101 116	89.62 %	2 530 995	76.14 %

How did we extend the dictionary?

- ▶ The starting point was a Bokmål - North Saami dictionary

How did we extend the dictionary?

- ▶ The starting point was a Bokmål - North Saami dictionary
- ▶ An FST translates frequent Nynorsk words into Bokmål:
 - ▶ *kva, korleis, kjem, sjølv, ikkje, truleg, betre, ...*

How did we extend the dictionary?

- ▶ The starting point was a Bokmål - North Saami dictionary
- ▶ An FST translates frequent Nynorsk words into Bokmål:
 - ▶ *kva, korleis, kjem, sjølv, ikkje, truleg, betre, ...*
- ▶ The Bokmål FST was then enriched with Nynorsk morphology

How did we extend the dictionary?

- ▶ The starting point was a Bokmål - North Saami dictionary
- ▶ An FST translates frequent Nynorsk words into Bokmål:
 - ▶ *kva, korleis, kjem, sjølv, ikkje, truleg, betre, ...*
- ▶ The Bokmål FST was then enriched with Nynorsk morphology

LEXICON m1 ! dag, hane
+N+Msc+Sg: 0-en ;
+N+Msc+Pl: er-ene ;
+N+Msc+Pl+Nynorsk: ar-ane ;
+N: R ;

One dictionary thus gives rise to many dictionaries

1. Ordinary dictionaries
 - ▶ If the word is lexicalised, then we show only the lexicalised translation
2. Student dictionary
 - ▶ The dictionary also translates each part of compound words
3. #SoMe dictionary
 - ▶ #SoMe dictionary accepts also *acdsz* in place of Saami letters *áčđšž*

FSTs for other Saami languages

1. Pite Saami: $rgg/rrg \leftrightarrow rg$
2. Skolt and Inari Saami: Extensive letter variation
3. Kildin Saami: Non-standard Cyrillic letters + two competing orthographies

DinOrdbok (John Atle Sandbakken)

Engelsk*	50562	Japansk	14553
Svensk*	48824	Tysk*	14171
Spansk*	48698	Slovakisk	13264
Polsk	20904	Tyrkisk*	12915
Fransk*	18846	Kinesisk	12694
Nederlandsk*	18804	Bulgarsk*	11974
Italiensk*	18587	Gresk	11789
Finsk*	18492	Katalansk*	11111
Nynorsk*	17402	Rumensk*	9618
Russisk*	17293	Samisk*	9481
Portugisisk*	17210	Islandsk*	7674
Dansk*	15793	Latinsk	6864

Conclusion 1

The online FST dictionary has many great benefits compared to the offline wordform dictionary:

- ▶ Uses existing FSTs (plus modifications) and XML lexica
- ▶ FSTs handle both linguistic and orthographic variation
- ▶ ... and more morphology:
 - ▶ North Saami 57,8 % → 83.3 % , South S. 41.4 % → 69.6 %
- ▶ The server-side setup is easy, can plug in new FSTs and lexica.

Conclusion 2

Many benefits to users, and easy to use.

- ▶ No installation, just bookmark
- ▶ New updates to lexicon and morphology arrive as they are made
- ▶ Works in all of the most popular desktop and mobile browsers
- ▶ Allows easier access to written material in numerous languages (also majority languages)