Next to nothing – a cheap South Saami disambiguator

Lene Antonsen and Trond Trosterud
University of Tromsø

Riga, May, 12th 2011

Content

- Introduction
- Saami languages
- Lemmatising
- South Saami as part of a larger Saami analyser
- Analysis
- Results
- Conclusion and future work

The leading idea

For morphologically rich languages, even a very small constraint grammar is able to reliably disambiguate on a POS level

Lemmatising

- What do we mean by lemmatising?

Lemmatising

- What do we mean by lemmatising?
- Deciding whether two wordforms belong to the same lemma or not might be problematic

Lemmatising

- What do we mean by lemmatising?
- Deciding whether two wordforms belong to the same lemma or not might be problematic
  1. We first define the parts of speech of the language by morphosyntactic means

Lemmatising

- What do we mean by lemmatising?
- Deciding whether two wordforms belong to the same lemma or not might be problematic
    1. We first define the parts of speech of the language by morphosyntactic means
    2. Which lexeme a given wordform belongs to will then follow from the overall POS structure

Lemmatising

- What do we mean by lemmatising?
- Deciding whether two wordforms belong to the same lemma or not might be problematic
    1. We first define the parts of speech of the language by morphosyntactic means
    2. Which lexeme a given wordform belongs to will then follow from the overall POS structure
    3. For us, lemmatising means finding the lexeme for each wordform

Lemmatising

- What do we mean by lemmatising?
- Deciding whether two wordforms belong to the same lemma or not might be problematic
  1. We first define the parts of speech of the language by morphosyntactic means
  2. Which lexeme a given wordform belongs to will then follow from the overall POS structure
  3. For us, lemmatising means finding the lexeme for each wordform
- Our results show that even a small constraint grammar may achieve results good enough to be used as a lemmatiser.

Derivations

- In the Saami languages there is much derivation, for all open word classes.

Derivations

- In the Saami languages there is much derivation, for all open word classes.
- In the transducer lexica, many of the derivations are lexicalized.

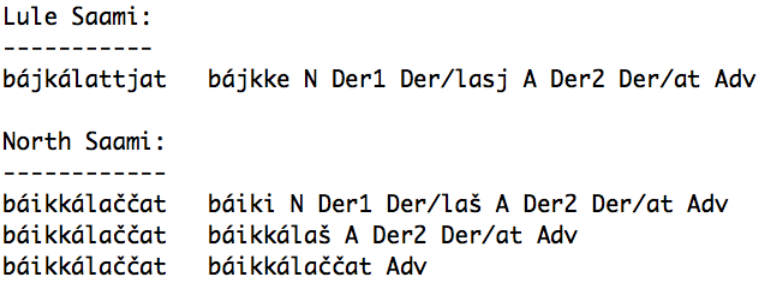
Derivations

- In the Saami languages there is much derivation, for all open word classes.
- In the transducer lexica, many of the derivations are lexicalized.
- In the output from the morphological analyser, there are dynamic analyses, in addition to the eventual lexicalized one

Derivations

- In the Saami languages there is much derivation, for all open word classes.
- In the transducer lexica, many of the derivations are lexicalized.
- In the output from the morphological analyser, there are dynamic analyses, in addition to the eventual lexicalized one
- There are more lexicalisations in the *sme* lexica than in the *sma* and *smj* ones

Derivation and the challenge of lexicalisation 1

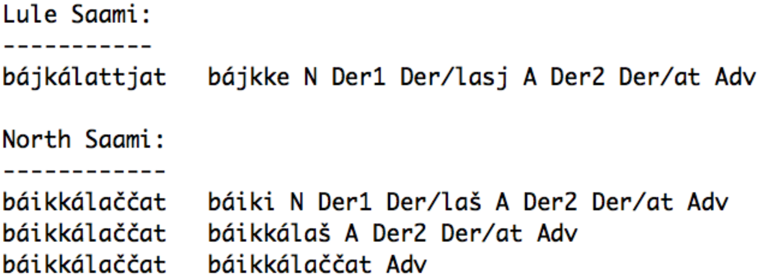


```
Lule Saami:
-----------
bájkálattjat   bájkke N Der1 Der/lasj A Der2 Der/at Adv

North Saami:
------------
báikkálaččat   báiki N Der1 Der/laš A Der2 Der/at Adv
báikkálaččat   báikkálaš A Der2 Der/at Adv
báikkálaččat   báikkálaččat Adv
```

Figure: The morphological analysis of derived words may differ for the *sme* and *smj* analysers.

Derivation and the challenge of lexicalisation 1

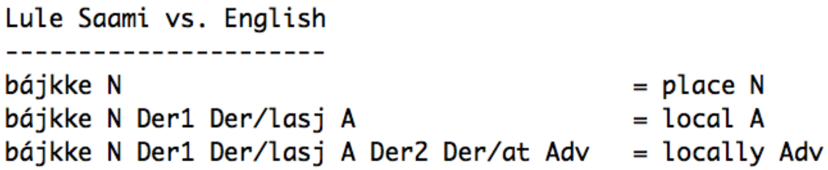


```
Lule Saami:
-----------
bájkálattjat   bájkke N Der1 Der/lasj A Der2 Der/at Adv

North Saami:
------------
báikkálaččat   báiki N Der1 Der/laš A Der2 Der/at Adv
báikkálaččat   báikkálaš A Der2 Der/at Adv
báikkálaččat   báikkálaččat Adv
```

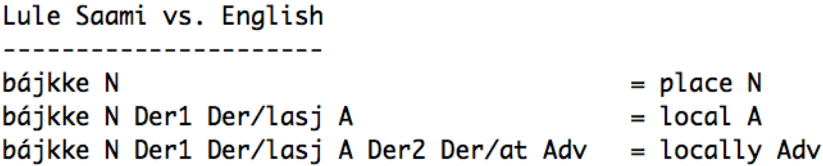
Figure: The morphological analysis of derived words may differ for the *sme* and *smj* analysers.

Derivation and the challenge of lexicalisation 2



```
Lule Saami vs. English
----------------------
bájkke N                                 = place N
bájkke N Der1 Der/lasj A                 = local A
bájkke N Der1 Der/lasj A Der2 Der/at Adv = locally Adv
```

Derivation and the challenge of lexicalisation 2

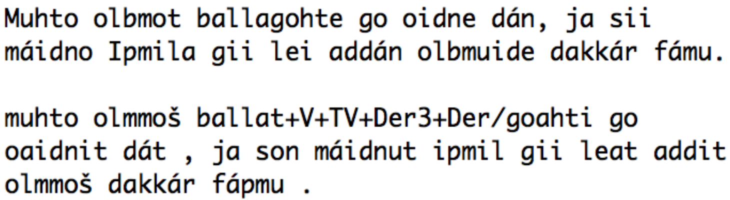


```
Lule Saami vs. English
----------------------
bájkke N                                = place N
bájkke N Der1 Der/lasj A                = local A
bájkke N Der1 Der/lasj A Der2 Der/at Adv = locally Adv
```

Derivation and the challenge of lexicalisation

- Choose the lexicalized reading if there is one
  - word alignment gives:
    noun *bájkke* ‘place’ = *báikkálaččat* ‘locally’

Example of lemmatised text with derivation tags



```
Muhto olbmot ballagohte go oidne dán, ja sii
máidno Ipmila gii lei addán olbmuide dakkár fámu.

muhto olmmoš ballat+V+TV+Der3+Der/goahti go
oaidnit dát , ja son máidnut ipmil gii leat addit
olmmoš dakkár fápmu .
```

Figure: *But people began to be afraid when they saw it, and they prised God which had given the people such a power.*

South Saami as part of a larger Saami analyser

Analysers	**Languages**		
lexicon and morphology	North Saami analyser	Lule Saami analyser	South Saami analyser
disambiguation	North Saami disambiguation	Lule Saami disambiguation	–
syntatic functions	common Saami analyser		
dependency	common Saami analyser		

Table: The common Saami analyser infrastructure. The disambiguation of South Saami is the missing link.

The test corpus

Corpus Bible 52 000 words, administrative text 169 000 words (not unknown to the fst)

Subforms The morphological analyser accepts substandard lemma and inflection forms

Typos For frequent typographical errors we have a correction procedure

Results

Table: Homonymy in South Sami

	Whole corpus	Fully analysed sentences only
Number of words	218.118	92.971
Analyses per thousand words		
Analyses with homonymy	1.625	1.778
Present disambiguation	1.118	1.121
Lemma + PoS disambuguation	1.064	1.065
Lemma + PoS disambuguation without distinguishing closed PoS	1.058	1.059

The CG rule set

The CG consists of 115 rules

Rule coverage

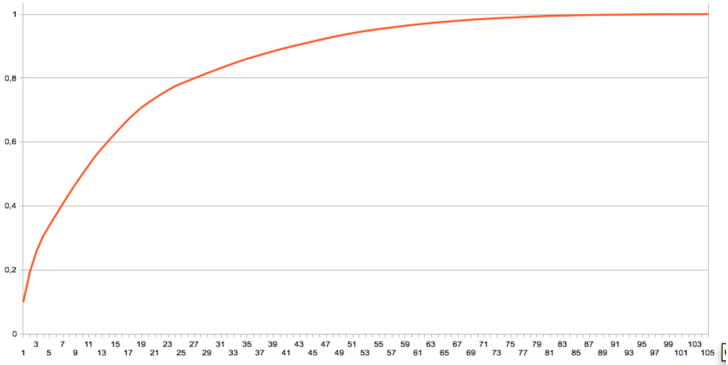


Figure: Cumulative effect of the CG rules

The 10 most efficient CG rules

1. "REMOVE: rm DerN if lexicalised"
2. "REMOVE: rm Prt Neg when Prs"
3. "REMOVE: rm Prop Attr"
4. "REMOVE: rm A Attr"
5. "REMOVE: rm Pron Pers when Pron Dem"
6. "REMOVE: rm Pron Dem"
7. "SELECT: select PrfPrc if copula to the left"
8. "SELECT: select Jupmele as Prop" Jumele = 'God'
9. "REMOVE: rm Px"
10. "REMOVE: rm not CS if Adv"

Remaining homonymies for open POS 1

The remaining homonymies are mainly of the following types:

- The same lemma, but different PoS, eg. *juktie* N ('carcass') vs. *juktie* CS ('so that')
- Different lemmas and different PoS, eg. *vihte* N ('wit') vs. *vihth* Adv ('again')
- Different lemmas, same PoS and inflection eg. *båetedh* V ('to come') vs. *böötedh* V ('to mend, to pay a fine'). These are the really hard ones to disambiguate.
- Different lemma, same PoS, but inflection is different (one of them may be derived from the other), eg. *utniedidh* V ('to held') vs *utnedh* V ('to have, to use')

Remaining homonymies for open POS 2

- The same lemma has one reading as Proper noun and one as common noun – *Saemie* N ('Saami') vs. *saemie* N ('saami')
- There are two orthographic variants of the same lemma, which should have been subsumed under the same lemma, eg. *ussjiedidh* V vs *ussjedidh* V ('think')
- Derivation vs. lexicalisation, eg. *ryöjnesjæjja* N vs *ryöjnesjidh+V+TV+Der1+Der/NomAg+N* ('shepherd')

Cumulative homonymy

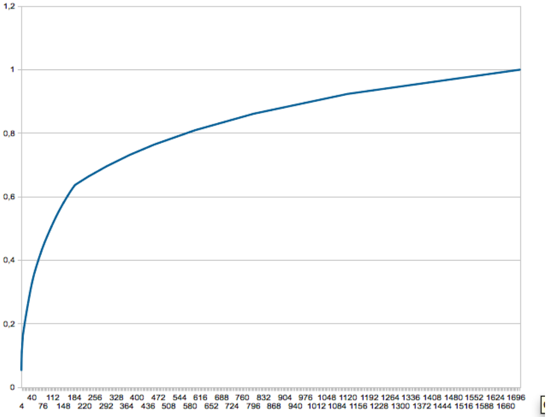


Figure: Cumulative homonymy for wordforms not assigned to a single lemma

Conclusion

- A small-size CG (115 rules) gives an accuracy of 1.118 - 1.058 readings/word
- 1/4 of the rule set removes 80% of the homonymy
- The CG is robust enough to give good disambiguation even with an fst coverage of only 93.5%
- The rule set is a good starting point for a full-fledged disambiguator

Future work

Make a disambiguator for South Saami :-)

GÆJHTOE!