Reusing Grammatical Resources for New Languages

Lene Antonsen, Trond Trosterud and Linda Wiechetek

Romssa Universitehta / University of Tromsø
Giellatekno / Sámi Language Technology

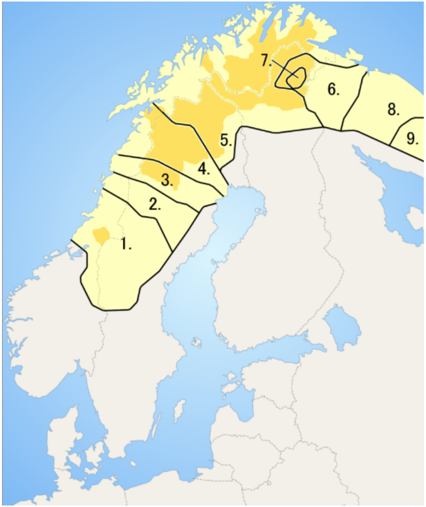
May 20, 2010

Introduction

- reuse of the hand-written North Sámi grammar for other languages (South and Lule Sámi, Faroese, Greenlandic)
- **We argue that:**
  - machine-readable grammars become more portable at higher levels of analysis (e.g. dependency)
  - lower levels: smaller modules can be reused
- we gain: new tools + linguistic insights (writing concise grammars also for languages with few speakers)

LANGUAGES

Sámi language area



- 1. South Sami
- 2. Ume Sami
- 3. Pite Sami
- 4. Lule Sami
- 5. North Sami
- 6. Skolt Sami
- 7. Inari Sami
- 8. Kildin Sami
- 9. Ter Sami

Darkened area represents municipalities that recognize Sami as an official language.

Figure: Sámi language area

North, Lule and South Sámi

North	Lule	South
nominative	nominative	nominative
gen-acc	genitive accusative	genitive accusative
locative	inessive elative	inessive elative
essive	essive	essive
comitative	comitative	comitative

Table: Case inventory for the Sámi nouns and pronouns

North, Lule and South Sámi - morphosyntactic and syntactic differences

level	**North**	**Lule**	**South**
inflection of the negation verb	not for tense	for tense	for tense
word order	SVO	SOV / SVO	SOV
copula	full	reduced	omitted
pro-drop:	1.& 2. person	all persons	1.& 2. person

Sámi vs. Faroese

Similarities	**Sámi and Faroese**	
morphosyntax	medium-sized case system + adpositions, binary tense system	
	finite auxiliaries + infinitives and participles	
	express future and aspect	
Differences	**Sámi**	**Faroese**
morphosyntax	no gender/ marginal case agreement	extensive case + gender agreement
syntax	relatively free word order	more restricted word order
	pro-drop language	non pro-drop language
	postpositions and OV (South Sámi)	prepositions, VO, V2

Table: Linguistic similarities and differences between Sámi and Faroese.

Sámi vs. Greenlandic

Similarities	**Sámi and Greenlandic**	
morphosyntax	similar case system; suffixes for person + number	
	dynamic derivation, anteriority morph. expressed	
	no gender	
syntax	relatively free word order, extensive use of nominals	
Differences	**Sámi**	**Greenlandic**
morphosyntax	nom-acc language	ergative language
	subjective conjugation	objective conjugation
	weak NP-internal agreement	no noun-modifying adj
syntax	SVO	SOV

Table: Similarities and differences between Sámi and Greenlandic

TECHNICAL BACKGROUND

Linguistic framework: Advantages of Dependency Grammar

- nodes are not ordered in a linear fashion
- → suitable for languages with a fairly free word order
- word-based
- → easily applicable to the Constraint Grammar analyser (which also performs word-based analysis)

Technical background

- morphological analysers implemented with finite-state transducers
- compiled with the Xerox compilers `twolc` and `lexc` (Beesley & Karttunen 2003)
- Constraint Grammar (CG) parsers for disambiguation and syntax
- Vislcg3 for the compilation of CG rules (VISL-group 2008)

Precision and recall for the North and Lule Sámi analysers

	sme:	**sme:**	**smj:**	**smj:**
	Precision	**Recall**	**Precision**	**Recall**
PoS	0.99	0.99	0.94	0.97
disambiguation	0.93	0.95	0.83	0.94
syntactic functions	0.93	0.93	0.86	0.86

sme = North Sámi
smj = Lule Sámi

REUSING GRAMMAR

Reusing grammar at lower levels

- morphophonology: rules for the same morphophonological processes with small adaptations (e.g. rule for consonant gradation)
- lexicon: international loanwords, place names
- disambiguation rules: e.g. verb disambiguation rules, rules for sentence and clause boundary detection

Reusing grammar at higher levels: Syntax

- common module shared by all Sámi languages for most syntactic function labels
- lemmata in sets are language specific
- language tags (<sme>, <smj>, <sma>) trigger language-specific exceptions
  - e.g. different cases for different Sámi languages for the habitive construction (North Sámi: locative, Lule Sámi: inessive, South Sámi: genitive)

Reusing grammar at the top level: Dependency Grammar

- lemma and tag sets that denote clause boundaries for the dependencies between clauses
- rules for subordinate clauses functioning as an object or adverbial
- rules for coordination
- same Constraint Grammar module for all 3 Sámi languages

UNRELATED LANGUAGES

Bootstrapping Faroese: adaptations

1. adding Faroese lemmata to existing clause boundary sets + adding new syntactic tags → accuracy: 0.960
2. adding a rule for dependency for infinitive markers + coordination of indirect objects → accuracy: 0.983
3. 11 language-specific rules taking care of subordinate clauses, optional omission of subjunctions *sum*, *ið* introducing subordinate clauses → accuracy: 0.986

Bootstrapping Faroese: adaptations

1. adding Faroese lemmata to existing clause boundary sets + adding new syntactic tags → accuracy: 0.960
2. adding a rule for dependency for infinitive markers + coordination of indirect objects → accuracy: 0.983
3. 11 language-specific rules taking care of subordinate clauses, optional omission of subjunctions *sum*, *ið* introducing subordinate clauses → accuracy: 0.986

(1)

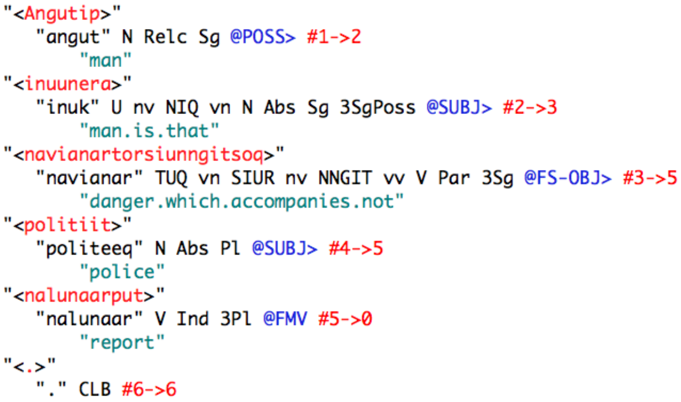
Hetta er ein tanki, [sum] tey flestu av okkum hava sera ilt við
this is a thought, [that] they most of us have very hard with to accept .

'This is a thought that most of us have difficulty accepting, . . .'

Bootstrapping Greenlandic

1. 40 new syntactic tags in the common disambiguation file (no equivalent in Sámi)
2. adding dependency rules for the new syntactic tags

Example: Bootstrapping Greenlandic



```
"<Angutip>"
   "angut" N Relc Sg @POSS> #1->2
       "man"
"<inuunera>"
   "inuk" U nv NIQ vn N Abs Sg 3SgPoss @SUBJ> #2->3
       "man.is.that"
"<navianartorsiunngitsoq>"
   "navianar" TUQ vn SIUR nv NNGIT vv V Par 3Sg @FS-OBJ> #3->5
       "danger.which.accompanies.not"
"<politiit>"
   "politeeq" N Abs Pl @SUBJ> #4->5
       "police"
"<nalunaarput>"
   "nalunaar" V Ind 3Pl @FMV #5->0
       "report"
"<.>"
   "." CLB #6->6
```

Figure: ‘The police report that the man is out of immediate danger.’

Evaluation

- gold standard corpora: 100 sentences per language (30 bible, 30 fiction, 40 newspaper)
- good results for related languages, but also fairly good results for lesser and un-related languages

Results

	sme	**smj**	**sma**	**fao**		**kal**	
grammat funct. / dep.	both	both	both	dep	both	dep	both
Sámi base analyser	0.99	0.99	0.99	-	-	-	-
enhanced with							
- lang-spec tags in sets	-	-	-	0.960	0.946	0.803	0.801
- rules for lang-spec tags	-	-	-	0.983	0.969	0.931	0.928
- lang-spec synt. rules	-	-	-	0.986	0.984	-	-

Table: Accuracy (F-score) for dependency analysis

sme = North Sámi
smj = Lule Sámi
sma = South Sámi
fao = Faroese
kal = Greenlandic

Conclusion

- large potential for reusing grammatical resources
- the higher up in the analysis (dependency) the more can be reused
- good results due to information encoded in the syntactic tag set (function and direction of the head)
- linguistic methods produce a lot of useful biproducts (e.g. verification of the reference grammar, a new contrastive grammar)
- linguistic methods can work language-independently
- for both statistical and linguistic approaches the potential for saving time lies in the reuse of infrastructure and insight

Future work

- rewriting the North Sámi rules to be truly language-independent, and making this accessible to other languages
- rewriting language-specific tag sets in a more modular way in order to make the maintenance of the language-independent file easier
- researching contrastive grammars
- making robust deep-syntactic parsers accessible for a wide range of languages

Many thanks to . . .

- Per Langgård (Greenlandic gold standard)
- Maja Lisa Kappfjell (South Sámi gold standard)
- Zakaris Svabo Hansen and Judithe Denbæk (Faroese and Greenlandic gold standard)

GRAZZI! GIITU!

Bibliography

Beesley, Kenneth R. & Lauri Karttunen (2003), *Finite State Morphology*, CSLI publications in Computational Linguistics, USA.

Karlsson, Fred (2006), *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, Berlin.

VISL-group (2008), Constraint grammar. `http://beta.visl.sdu.dk/constraint_grammar.html`.