Reusing Grammatical Resources for New Languages

Lene Antonsen, Trond Trosterud and Linda Wiechetek

Romsssa Universitehta / University of Tromsø
Giellatekno / Sámi Language Technology
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

Figure: 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.
### Table: Case inventory for the Sámi nouns and pronouns

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>Lule</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominative</td>
<td>nominative</td>
<td>nominative</td>
<td>nominative</td>
</tr>
<tr>
<td>gen-acc</td>
<td>genitive</td>
<td>accusative</td>
<td>genitive</td>
</tr>
<tr>
<td>locative</td>
<td>inessive</td>
<td>elative</td>
<td>inessive</td>
</tr>
<tr>
<td>essive</td>
<td>essive</td>
<td></td>
<td>essive</td>
</tr>
<tr>
<td>comitative</td>
<td>comitative</td>
<td></td>
<td>comitative</td>
</tr>
</tbody>
</table>
## North, Lule and South Sámi - morphosyntactic and syntactic differences

<table>
<thead>
<tr>
<th>level</th>
<th>North</th>
<th>Lule</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>inflection of the negation verb</td>
<td>not for tense</td>
<td>for tense</td>
<td>for tense</td>
</tr>
<tr>
<td>word order</td>
<td>SVO</td>
<td>SOV / SVO</td>
<td>SOV</td>
</tr>
<tr>
<td>copula</td>
<td>full</td>
<td>reduced</td>
<td>omitted</td>
</tr>
<tr>
<td>pro-drop:</td>
<td>1. &amp; 2. person</td>
<td>all persons</td>
<td>1. &amp; 2. person</td>
</tr>
</tbody>
</table>
## Sámi vs. Faroese

<table>
<thead>
<tr>
<th>Similarities</th>
<th>Sámi and Faroese</th>
</tr>
</thead>
<tbody>
<tr>
<td>morphosyntax</td>
<td>medium-sized case system + adpositions, binary tense system finite auxiliaries + infinitives and participles express future and aspect</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Differences</th>
<th>Sámi</th>
<th>Faroese</th>
</tr>
</thead>
<tbody>
<tr>
<td>morphosyntax</td>
<td>no gender/ marginal case agreement relatively free word order pro-drop language postpositions and OV (South Sámi)</td>
<td>extensive case + gender agreement more restricted word order non pro-drop language prepositions, VO, V2</td>
</tr>
<tr>
<td>syntax</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Linguistic similarities and differences between Sámi and Faroese.
### Sámi vs. Greenlandic

#### Similarities

<table>
<thead>
<tr>
<th>Sámi and Greenlandic</th>
<th>Morphosyntax</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>similar case system; suffixes for person + number</td>
<td>dynamic derivation, anteriority morph. expressed</td>
<td>relatively free word order, extensive use of nominals</td>
</tr>
</tbody>
</table>

#### Differences

<table>
<thead>
<tr>
<th>Sámi</th>
<th>Greenlandic</th>
</tr>
</thead>
<tbody>
<tr>
<td>morphosyntax</td>
<td>nom-acc language</td>
</tr>
<tr>
<td>subjective conjugation</td>
<td>objective conjugation</td>
</tr>
<tr>
<td>weak NP-internal agreement</td>
<td>no noun-modifying adj</td>
</tr>
<tr>
<td>SVO</td>
<td>SOV</td>
</tr>
</tbody>
</table>

**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 \texttt{twolc} and \texttt{lexc} (?)
- Constraint Grammar (CG) parsers for disambiguation and syntax
- \texttt{Vislcg3} for the compilation of CG rules (?)
<table>
<thead>
<tr>
<th></th>
<th>sme: Precision</th>
<th>sme: Recall</th>
<th>smj: Precision</th>
<th>smj: Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PoS</td>
<td>0.99</td>
<td>0.99</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>disambiguation</td>
<td>0.93</td>
<td>0.95</td>
<td>0.83</td>
<td>0.94</td>
</tr>
<tr>
<td>syntactic functions</td>
<td>0.93</td>
<td>0.93</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>

`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
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
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. 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
    this is a thought, [that] they most of us have very
    ilt við at góðtaka . . .
    hard with to accept . . .
    ‘This is a thought that most of us have difficulty accepting,
    . . .’
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"
"<navianartorsiuunngitsoq>"
  "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
**Table:** Accuracy (F-score) for dependency analysis

<table>
<thead>
<tr>
<th></th>
<th>sme</th>
<th>smj</th>
<th>sma</th>
<th>fao</th>
<th>kal</th>
</tr>
</thead>
<tbody>
<tr>
<td>grammat funct. / dep.</td>
<td>both</td>
<td>both</td>
<td>both</td>
<td>dep</td>
<td>both</td>
</tr>
<tr>
<td>Sámi base analyser</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>enhanced with</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- lang-spec tags in sets</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.960</td>
<td>0.946</td>
</tr>
<tr>
<td>- rules for lang-spec tags</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.983</td>
<td>0.969</td>
</tr>
<tr>
<td>- lang-spec synt. rules</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.986</td>
<td>0.984</td>
</tr>
</tbody>
</table>

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

- 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 by-products (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!