Instant annotations in ELAN corpora of spoken and written Komi-Zyrian, an endangered language of the Barents Sea region (Russia)

Ciprian Gerstenberger\textsuperscript{a}, Niko Partanen\textsuperscript{b}, Michael Rießler\textsuperscript{c}

\textsuperscript{a}Giellatekno – Saami Language Technology, UiT The Arctic University of Norway \hspace{1em} \textsuperscript{b}Department of Uralic Studies, Uni Hamburg \hspace{1em} \textsuperscript{c}Freiburg Institute for Advanced Studies, Uni Freiburg

### Background

- ELAN\textsuperscript{1} is a widespread GUI tool for
  - transcribing and translating field recordings
  - creating further annotations aligned to audio and video
  - searching and analysing the resulting corpus data
- Morphosyntactic annotations are typically done
  - manually in ELAN, or
  - semi-manually in interaction with other tools
- NLP tools for low-resourced \textit{written} languages exist, but they are rarely applied in \textit{spoken} language documentation projects.

### Promising alternative approach

- adapt NLP tools to small spoken languages
- avoid ineffective manual work
- create larger and deeper annotated corpora

### ELAN-FST/CG Integration

Automated workflow for generating morphosyntactic analysis in ELAN

Using available Giellatekno\textsuperscript{2} tools:
- Finite-State-Transducer (FST) for
  - morphological analysis
- Constraint Grammar (CG) for
  - disambiguation and syntactic analysis

**Test case** with structurally uniform data from Komi-Zyrian (160,000 speakers), full syntactic analysis and dependency tagging is in the works.\textsuperscript{3}

Ongoing projects on smaller languages from the Barents region:
- Kildin Saami (500 speakers)
- Skolt Saami (50 speakers)
- Pite Saami (30 speakers)\textsuperscript{4}

### Guiding principles

- orthographic transcription system
- computerized annotation methods

### Size of the different spoken and written corpora for Komi-Zyrian

<table>
<thead>
<tr>
<th>Language</th>
<th>Modality</th>
<th>Recorded speakers</th>
<th>Time span of texts</th>
<th>Tokens in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Komi-Zyrian (Standard)\textsuperscript{5}</td>
<td>written</td>
<td>~2,500</td>
<td>1920–2017</td>
<td>30,000,000</td>
</tr>
<tr>
<td>Komi-Zyrian (Izhva dialect) spoken</td>
<td>~150</td>
<td>1844–2016</td>
<td></td>
<td>200,000</td>
</tr>
<tr>
<td>Komi-Zyrian (Udora dialect) spoken</td>
<td>~50</td>
<td>1902–2013</td>
<td></td>
<td>40,000</td>
</tr>
</tbody>
</table>

### Processing pipeline

1. utterance extraction: Python script
2. tokenization: Perl script
3. morphosyntactic analysis: FST
4. disambiguation: CG
5. ELAN tier building: Python script

### Prospects

Advantages of \textit{rule-based} morphosyntactic modeling for \textit{endangered languages}:
- precise results of \textit{automatic tagging}
- simultaneous creation of both a \textit{tool} and a \textit{morphosyntactic description}
- deployable for new (I)CALL technology for language revitalisation purposes

The Giellatekno open-access infrastructure includes dictionaries and rule-based grammars for several circumpolar (written) languages. It can be used for new (spoken/written) language projects easily.

Our approach challenges current manual practices in endangered language documentation projects.

### References

2. http://giellatekno.uit.no