

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

Ciprian Gerstenberger^a, Niko Partanen^b, Michael Rießler^c

^aGiellatekno – Saami Language Technology, UiT The Arctic University of Norway ^bDepartment of Uralic Studies, Uni Hamburg ^cFreiburg Institute for Advanced Studies, Uni Freiburg

Background

- ▶ ELAN¹ 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 **written** languages exist, but they are rarely applied in **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 Giellatekno² 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.³

Ongoing projects on smaller languages from the Barents region:

- ▶ Kildin Saami (500 speakers)
- ▶ Skolt Saami (500 speakers)
- ▶ Pite Saami (30 speakers)⁴

Guiding principles

- ▶ **orthographic** transcription system
- ▶ **computerized** annotation methods

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

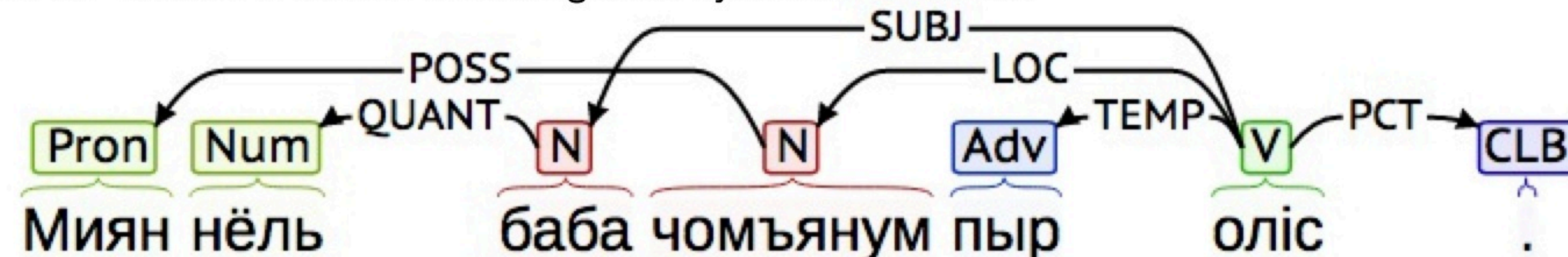
Language	Modality	Recorded speakers/writers	Time span of texts	Tokens in corpus
(<Komi<Permic<Finno-Ugric<Uralic)				
Komi-Zyrian (Standard) ⁵	written	~2,500	1920–2017	30,000,000
Komi-Zyrian (Izhva dialect)	spoken	~150	1844–2016	200,000
Komi-Zyrian (Udora dialect)	spoken	~50	1902–2013	40,000

INPUT: utterance

OUTPUT: word tier, lemma tier, part-of-speech tier, morphosyntactic analysis tier

DISAMBIGUATION

IDEAL ANNOTATION: unambiguous syntactic structure



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 **rule-based** morphosyntactic modelling for **endangered languages**:

- ▶ precise results of **automatic tagging**
- ▶ simultaneous creation of both a **tool** and a **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

- 1 <http://www.mpi.nl/corpus/html/elan/>
- 2 <http://giellatekno.uit.no>
- 3 **R. Blokland, N. Partanen, M. Rießler:** “Language documentation meets language technology” Ongoing project funded by Kone Foundation (2017–2020)
- 4 **C. Gerstenberger, N. Partanen, M. Rießler, J. Wilbur** (2017, in press): Utilizing language technology in the documentation of endangered Uralic languages *Northern European Journal of Language Technology (NEJLT)*
- 5 <http://komicorpora.ru>