Developing technologies for low-resource Uralic languages: Case studies on Saami and Komi varieties

Niko Partanen[1], Michael Rießler[2] & Thierry Poibeau[3]

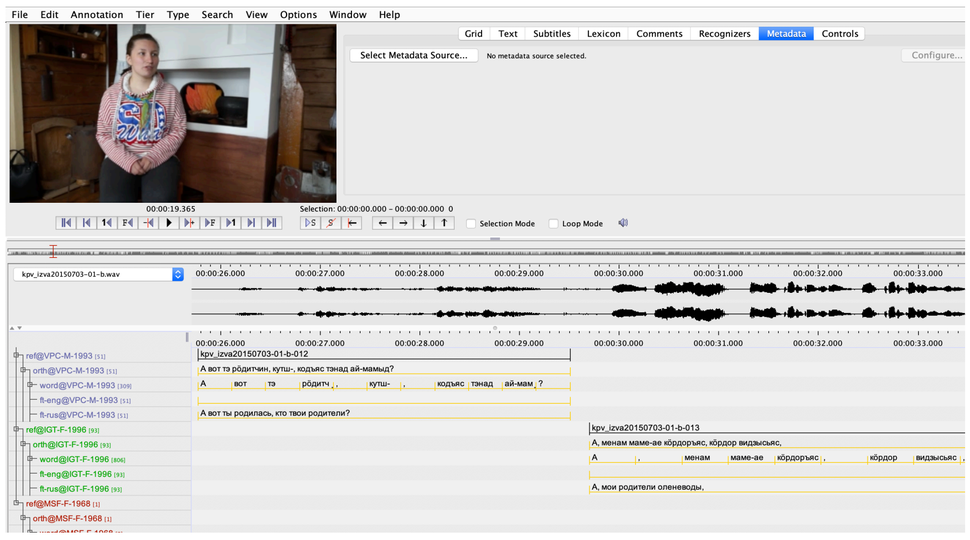
[1]University of Helsinki
[2]University of Eastern Finland
[3]LATTICE (CNRS & ENS / PSL & Université Sorbonne nouvelle / USPC)

Context

- Endangered (or highly endangered) languages in Northern Eurasia
- Spoken (still) by relatively large numbers of people
- Large written corpora exist
  - Different written genres: novels, legislation, news
- Large multimedia corpora are also becoming available
  - E.g. Komi with 350,000 tokens
- Diversity within the languages: dialects and closely related languages
- Several historical and contemporary writing systems
- Need for increased maintenance and revitalization efforts
- Important, and under-described, languages for linguistic research

Language technology for Saami and Komi

- Older rule-based systems at advanced levels in Giellatekno infrastructure
- Recent Universal Dependencies treebanks
  - Northern Saami, Skolt Saami
  - Zyrian Komi, Permian Komi
- Experiments on multilingual dependency parsing
  - Results promising but not yet practical within **corpus building workflows**



Example from recording kpv_izva20150703-01

Language documentation materials

- Data types: Audio, video, associated XML files
- Content: Transcriptions, translations, accurate descriptive metadata
- Several finished projects and ongoing work
  - Spoken Komi corpus (40 hours of transcriptions)
  - Multiple spoken corpora of Saami varieties (+100 hours)

Results

- Successful methods to assist and replace traditional manual annotation practices with NLP
- Research on taking advantage of existing resources on related and contact languages
  - With endangered languages we will continuously work with small languages – it is important to improve the low-resource scenario
  - Multilingual content needs to be addressed
- Universal Dependencies treebanks connected to a language documentation projects outcome

Challenges

- Open Access (for parts of the spoken data)
- Multi-speaker recordings (challenge for automatic processing but also publishing [no good interfaces for overlapping speech])
- Audio quality of fieldwork recordings
- Video is often available but unused in common corpus interfaces
- **Integrating technology** into daily data management practices while working on these languages

Technology to be developed

- Efficient processing of already published data
  - (Often in marginal writing systems)
  - Post-correction, transliteration, alignment
- Better methods for utterance segmentation and speaker diarization
  - Fine-tuning the models with new data
- Video signal currently largely unused – possibilities of pose estimation and face recognition
  - Potential source of better content descriptions
- Attempts to use speech recognition to get preliminary transcriptions (not yet successful)
- Reliable and accurate dependency parsing in low-resource scenarios

Related publications

Blokland, R., **Partanen, N.**, **Rießler, M.**, & Wilbur, J. (2019). Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA, February 26–27, 2019 (Vol. 2, pp. 24-30). University of Colorado.

Partanen, N., Lim, K., **Rießler, M.**, & **Poibeau, T.** (2018). Dependency parsing of code-switching data with cross-lingual feature representations. In Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages (pp. 1-17).

Lim, K., **Partanen, N.**, & **Poibeau, T.** (2018). Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)

Partanen, N., Blokland, R., Lim, K., **Poibeau, T.**, & **Rießler, M.** (2018). The First Komi-Zyrian Universal Dependencies Treebanks. In Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium (pp. 126-132).

Lim, K., & **Poibeau, T.** (2017). A system for multilingual dependency parsing based on bidirectional LSTM feature representations. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (pp. 63-70).

Gerstenberger, C., **Partanen, N.**, **Rießler, M.**, & Wilbur, J. (2017). Utilizing language technology in the documentation of endangered Uralic languages. Northern European Journal of Language Technology: Special Issue on Uralic Language Technology.