ABSTRACT

for Symposium B.12:
Building diachronic corpora for minority languages – facing methodological challenges at CIFU-XIV, Tartu, 18-23 August 2025

Incorporating past, present and future
into the Pite Saami documentation corpus

Joshua Wilbur • University of Tartu

Pite Saami is a critically endangered Uralic language spoken by a handful of individuals from Arjeplog municipality in Sweden and adjacent areas of Norway. Over the last 15 years, I have been compiling an annotated documentation corpus using the ELAN annotation software (Sloetjes & Wittenburg 2008). The corpus is clearly "opportunistic" in the variety of genres, modes and dates of the contents. Although it predominantly consists of my own multimedia recordings (both elicited and spontaneous speech) of arguably the last generations of native speakers, it also includes random older archived recordings (up to 100 years old), heritage transcribed texts dating back to 1893, as well as very recent contemporary media arising from on-going revitalization efforts, including television, radio and children's literature; with these diverse Pite Saami texts in mind, the corpus covers the "distant" past, the present state of the language, and should be set to incorporate future developments of Pite Saami. Automatic lemmatization and glossing of the corpus (e.g., Gerstenberger et al. 2017) is even supported using Natural Language Processing pipelines developed in collaboration with colleagues involved in endangered language documentation and with Giellatekno at UiT in Tromsø.

In this talk, I will present a brief overview of the creation of the Pite Saami documentation corpus, including how I have attempted to incorporate all of these various texts (from heritage texts up to contemporary media), especially concerning technical, ethical and legal aspects. I will pay special attention to the CARE principles for indigenous data governance (as proposed by Carroll et al. 2021),[1] and provide an analysis of the extent to which these both should and can realistically be implemented for Pite Saami as a minor and a critically endangered Uralic language.

While the Scandinavian political context which frames the Pite Saami situation is different to those of most other minority Uralic languages, this analysis will hopefully con-tribute to the discussion of CARE as a viable framework for minor Uralic languages in general, and especially concerning corpus building.

[1] Inspired by the FAIR data principles (Wilkinson et al. 2016), CARE stipulates that using indigenous data requires supporting the indigenous community, reflecting its values and ethics, and actively involving the community in determining how the data is used.

Bibliography

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, & Shelley Stall (2021). “Operationalizing the CARE and FAIR Principles for Indigenous data futures”. In: *Scientific Data* 8.1, p. 108.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler, & Joshua Wilbur (2017). “Instant annotations. Applying NLP methods to the annotation of spoken language doc-umentation corpora”. In: *International Workshop on Computational Linguistics for Uralic languages (IWCLUL 2017)*. Ed. by Tommi A. Pirinen, Michael Rießler, Trond Trosterud, & Francis M. Tyers. St. Petersburg: Association for Computational Lin-guistics, pp. 25–36.

Sloetjes, Han & Peter Wittenburg (2008). “Annotation by Category: ELAN and ISO DCR”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C ’t Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, & Barend Mons (2016). “The FAIR Guiding Principles for scientific data management and steward-ship”. In: *Scientific Data* 3.