Niko Partanen
Department of Finnish, Finno-Ugrian and Scandinavian Studies
University of Helsinki

Finno-Ugrian Society

Digitization of text collections by the Finno-Ugrian Society

The Finno-Ugrian Society has a rich 140-year history of publishing research and materials on Uralic and Altaic languages and cultures. These publications encompass scholarly research, text collections and dictionaries.

Digital processing of transcribed texts in Uralic languages has historically presented formidable challenges. However, advancements seen over the last decades in Unicode coverage and text recognition technology have made it possible to handle even highly detailed phonetic transcriptions relatively easily (see, for example, Partanen et al. 2022: 375). Consequently, the Finno-Ugrian Society has initiated the digitization of its text collections and plans to publish these materials electronically. This follows the work already finished earlier with the digitized journals, and digitization of the dictionaries is expected to follow.

The main rationale for this order of proceeding has been that the resulting data is increasingly complex and requires further and further processing. When research articles are digitized, providing a PDF file online meets most of the demands very adequately. With the text collections we are looking more toward electronic corpora that contain the original data, but in a widely enhanced form. The dictionaries, on the other hand, would demand extensive reformatting from word entries into structured data (see Partanen & Rueter 2025). Good results have also been reported in adapting dialectal data to contemporary tools of language technology (Rueter et al. 2022). Much of this work needs to be done after digitization, and can be seen as later research, whereas some parts already need to be taken care of during the digitization phase: defining which task belongs to which work phase is crucial, as this also delineates the actors involved and resources needed at different steps.

Providing text collections as corpora requires accurately recognized and proofread texts with metadata that is restructured and harmonized. Enhancing the transcription with phonemic transcription and contemporary orthography layers is also central for the general usability and further processing of this data. Providing the original transcription as-is in a digital format would be the bare minimum, but both modern research use and community access call for higher standard. With a contemporary orthography layer, the materials can be easily connected to various historical corpora that will be eventually created, and if the materials are licensed openly, researchers can fruitfully build upon each other's work. Creating the orthography layer either partly or fully automatically also appears to be a realistic goal (Partanen 2024).

In recent years, copyright associations in Finland have started to agree to contracts where older works are made openly available. This is a good first step, but the practice doesn't involve re-lisencing, which would be important for new corpora that are created on the basis of original publications. The Finno-Ugrian Society has started to negotiate new licensing with the original authors, which is a more flexible but also very resource-intensive approach. At the same time, we have also seen initiatives of the Indigenous people represented in these texts. In instances where the rights are particularly challenging to renegotiate, usually due to the copyright holders being impossible to reach, serious consideration must be given to how the authorship is understood and weighted in relation to the people's right to their own cultural heritage. This may also involve more a nuanced distinction between the authorship of different components, i.e. the authors of the text and translations.

References

Partanen, N., Blokland, R., Rießler, M., & Rueter, J. (2022). Transforming Archived Resources with Language Technology: From Manuscripts to Language Documentation. In *The 6th Digital Humanities in the Nordic and Baltic Countries 2022 Conference, Uppsala, Sweden, March 15-1*, 2022. (Vol. 3232). CEUR-WS, pp. 370–380.

Partanen, N. (2024). Using Large Language Models to Transliterate Endangered Uralic Languages. In Hämäläinen, M., Pirinen, F., Macias, M., Crespo Avila, M.

(eds.) *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*. Association for Computational Linguistics, Helsinki, Finland, pp. 81–88.

Rueter, J. M., & Partanen, N. T. (2022). Improving lexical coverage of the Komi morphological analyser: The challenge of dialectal variation. In *Электронная письменность народов Российской Федерации-2021 & IWCLUL 2021. Материалы Международной научно-практической конференции 23–24 сентября 2021 г.* Коми республиканская академия государственной службы и управления (ГОУ ВО КРАГСиУ), pp. 196–207.

Rueter, J., & Partanen, N. (2025). Restructuring and visualising dialect dictionary data: Report on Erzya and Moksha materials. In Hämäläinen M., Öhman, E., Bizzoni, Y., Miyagawa, S., & Alnajjar, K (eds.). *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. Association for Computational Linguistics, Albuquerque, USA, pp. 41−47.