A MediaWiki Environment for Curating Dictionaries by Intercomparison and Community Involvement

Lexicographic resources for endangered languages are often fragmented into several different paper publications that have been digitized by different conventions and during different eras. The severely endangered Skolt Sami is no exception. For one part, a massive open-source dictionary for Skolt Sami has been made available on the MediaWiki based Akusanat (Hämäläinen & Rueter, 2018) and the Giellatekno infrastructure (Moshagen et al., 2013). For the other, valuable resources such as Sammallahti and Moshnikoff materials (Sammallahti & Moshnikoff, 1991) are not formatted in the same structure and their inclusion in the aforementioned systems poses a challenge.

In order to make all three of the lexicographic resources for Skolt Sami available for researchers and non-academic dictionary users alike, we propose a new extension to the MediaWiki based Akusanat dictionary that makes intercomparison of the different materials possible together with a facilitated semantic search functionality. The main goal is to alleviate the workload of dictionary editors when bringing lexicographic data available in the unified system. As community involvement in editing online dictionaries has previously been identified as a viable way of extending the lexicographic data (cf Everson et al., 2019), we propose a simplified interface for non-technical community members to actively participate in the dictionary editing process.

The intercomparison of the different dictionaries is made possible by our external online toolkit which allows the user to import a set of words for comparison with the Akusanat MediaWiki system data. This juxtaposition of the new data with the existing one in the system can then be queried with a powerful semantic search functionality, edited and published in Akusanat.

As MediaWiki is highly customizable, we introduce an extension with an intuitive user interface that allows the user to find and filter lexemes. The user can use the extension to find lexemes by their part of speech, source of the entry, translation languages and so on. Once the user has submitted the desired query, the system returns the matching lexemes along with additional properties if requested. These properties are, for instance, inflection category, semantic tag and assonance. Using these meta-information, researchers can easily gather similar words together.

We implement a simplified interface for involving the community speaking Skolt Sami to improve the quality and accuracy of the information presented in our system. The extension allows users to go through words and their translations in the dictionaries masking all the additional technical information of each entry. Users can then verify the correctness of the translations, suggest edits and provide reasons for their opinion. All input feedback along with who has provided it is stored in the system.

List of references

Everson, R., Honoré, W., & Grimm, S. (2019). An Online Platform for Community-Based Language Description and Documentation. In Proceedings of the Workshop on Computational Methods for Endangered Languages (Vol. 1).

Hämäläinen, M., & Rueter, J. M. (2018). Advances in synchronized XML-MediaWiki dictionary development in the context of endangered Uralic languages. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (pp. 967-978). Ljubljana: Ljubljana University Press.

Moshagen, S. N., Pirinen, T., & Trosterud, T. (2013). Building an open-source development infrastructure for language technology projects. In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) (pp. 343-352).

Sammallahti, P., & Moshnikoff J. (1991). Suomi-koltansaame sanakirja. Lää'dd-sää'm sää'nnǩe'rjj. Girjegiisá Oy. Ohcejohka.