From crowdsourcing to nichesourcing with Uralic languages

During the Digitisation Project of Kindred Languages (2012–15) the National Library of Finland has digitised and made available approximately 1200 monographs and more than 100 newspaper titles in several Uralic languages. The resulting materials will constitute the largest resource for the Uralic languages in the world. As well as the digitised materials in the Uralic languages, the project will also produce their development tools to support linguistic research and citizen science. The project will allow researchers to gain access to corpora, which they have not been able to study before and to which all users will have open access regardless of their place of residence.

By Jussi-Pekka Hakkarainen, Project Manager, the National Library of Finland's research library
Data enhancement to meet the researchers’ needs

The Digitisation Project of Kindred Languages is linked with the research of language technology. The mission is to improve the usage and usability of digitised content. During the project we have advanced methods that will refine the raw data for further use, especially in the linguistic research.

The machined-encoded text (OCR) contains quite often too many mistakes to be used as such in research. The mistakes in OCR texts must be corrected. For enhancing the OCR texts, the National Library of Finland developed an open source code OCR editor that enabled the editing of machine-encoded text for the benefit of linguistic research.

This tool was necessary to implement, since these rare and peripheral prints often included perished characters, which are sadly neglected by the modern OCR software developers, but belong to the historical context of kindred languages and thus are an essential part of the linguistic heritage.

The material offers a lot, but how to find it?

The majority of the digitised literature was originally published in the 1920s and the 1930s, which was an era when the many Uralic languages were converted into a medium of popular education, enlightenment and dissemination of information pertinent to the developing political agenda of the Soviet state.

The “deluge” of popular literature in the 1920s-1930s suddenly challenged the lexical orthographic norms of the limited ecclesiastical publications from the 1880s. Newspapers were now written in orthographies and in word forms that the locals would understand. Textbooks were written to address the separate needs of both adults and children. New concepts were introduced in the language.

This was the beginning of a renaissance and period of enlightenment. The linguistically oriented population can also find writings to their delight, especially lexical items specific to a given publication, and orthographically documented specifics of phonetics.

From Crowdsourcing to Nichesourcing

The written material from this period is a gold mine, but how to filter the material for the benefit of research? Could crowdsourcing play some role here? How does our library meet the objectives, which appear to be beyond its traditional playground?

The traditional methods of crowdsourcing cannot be implemented here, since the targets in crowdsourcing have often been split into several micro-tasks that do not require any special skills.

This way of crowdsourcing may produce quantitative results, but from the point
Nichesourcing is a specific type of crowdsourcing where tasks are distributed amongst a small crowd of citizen scientists (communities). Although communities provide smaller pools to draw resources, their specific richness in skill is suited for the complex tasks with high-quality product expectations found in nichesourcing.

**Citizen scientists at work**

Communities have purpose, identity and their regular interactions engender social trust and reputation. These communities can correspond to research more precisely. Instead of repetitive and rather trivial tasks, we are trying to utilise the knowledge and skills of citizen scientists to provide qualitative results.

Some selection must be made, since we are not aiming to correct all 200,000 pages which we have digitised, but to give such assignments to citizen scientists that would precisely fill the gaps in linguistic research. A typical task would be editing and collecting the words in such fields of vocabularies, where the researchers do require more information.

**Research and society**

For instance, there is a lack of Hill Mari words in anatomy. We have digitised the books in medicine and we could try to track the words related to human organs by assigning the citizen scientists to edit and collect words with the OCR editor.

From the perspective of nichesourcing, it is essential that altruism plays a central role, when language communities are involved. Our goal with nichesourcing is to reach a certain level of interplay where the language communities would benefit from the results.

For instance, the corrected words in Ingrian will be added to the online dictionary, which is made freely available to the public and the society to benefit from as well. This objective of interplay can be understood as an aspiration to support the endangered languages and the maintenance of lingual diversity, but also as a servant of ‘two masters’, research and society.

The materials are available to both researchers and the general public in the National Library’s Fenno-Ugrica collection. The project is financially supported by the Kone Foundation.