**HAKKARAINEN — Nichesourcing**

<0 images>

**Nichesourcing the Uralic Languages for the Benefit of Linguistic Research and Lingual Societies**

Hakkarainen, J.-P.

The emerging technologies have recently challenged the libraries to reconsider their role as a mere mediator between the collections, researchers, and wider audiences (Sula, 2013), and libraries, especially the nationwide institutions like national libraries, haven’t always managed to face the challenge (Nygren et al., 2014). In the Digitization Project of Kindred Languages, the National Library of Finland has become a node that connects the partners to interplay and work for shared goals and objectives. In this paper, I will be drawing a picture of the crowdsourcing methods that have been established during the project to support both linguistic research and lingual diversity.

The National Library of Finland has been executing the Digitization Project of Kindred Languages since 2012. The project seeks to digitize and publish approximately 1,200 monograph titles and more than 100 newspapers titles in various, and in some cases endangered Uralic languages. Once the digitization has been completed in 2015, the Fenno-Ugrica online collection will consist of 110,000 monograph pages and around 90,000 newspaper pages to which all users will have open access regardless of their place of residence.

The majority of the digitized literature was originally published in the 1920s and 1930s in the Soviet Union, and it was the genesis and consolidation period of literary languages. This was the era when many Uralic languages were converted into media 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 to 1930s suddenly challenged the lexical orthographic norms of the limited ecclesiastical publications from the 1880s onward. 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 (Rueter, 2013). 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.

The project is financially supported by the Kone Foundation in Helsinki and is part of the Foundation’s Language Programme. One of the key objectives of the Kone Foundation Language Programme is to support a culture of openness and interaction in linguistic research, but also to promote citizen science as a tool for the participation of the language community in research. In addition to sharing this aspiration, our objective within the Language Programme is to make sure that old and new corpora in Uralic languages are made available for the open and interactive use of the academic community as well as the language societies. Wordlists are available in 17 languages, but without tokenization, lemmatization, and so on. This approach was verified with the scholars, and we consider the wordlists as raw data for linguists. Our data is used for creating the morphological analyzers and online dictionaries at the Helsinki and Tromsø Universities, for instance.

In order to reach the targets, we will produce not only the digitized materials but also their development tools for supporting linguistic research and citizen science. The Digitization Project of Kindred Languages is thus linked with the research of language technology. The mission is to improve the usage and usability of digitized content. During the project, we have advanced methods that will refine the raw data for further use, especially in the linguistic research. How does the library meet the objectives, which appears to be beyond its traditional playground?

The written materials from this period are a gold mine, so how could we retrieve these hidden treasures of languages out of the stack that contains more than 200,000 pages of literature in various Uralic languages? The problem is that the machined-encoded text (OCR) contains often too many mistakes to be used as such in research. The mistakes in OCRed texts must be corrected. For enhancing the OCRed 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 did often include already 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 (van Hemel, 2014).

Our crowdsourcing tool application is essentially an editor of Alto XML format. It consists of a back-end for managing users, permissions, and files, communicating through a REST API with a front-end interface—that is, the actual editor for correcting the OCRed text. The enhanced XML files can be retrieved from the Fenno-Ugrica collection for further purposes.

Could the crowd do this work to support the academic research? The challenge in crowdsourcing lies in its nature. The targets in the traditional crowdsourcing have often been split into several microtasks that do not require any special skills from the anonymous people, a faceless crowd. This way of crowdsourcing may produce quantitative results, but from the research’s point of view, there is a danger that the needs of linguists are not necessarily met. Also, the remarkable downside is the lack of shared goal or the social affinity. There is no reward in the traditional methods of crowdsourcing (de Boer et al., 2012). Also, there has been criticism that digital humanities makes the humanities too data-driven and oriented towards quantitative methods, losing the values of critical qualitative methods (Fish, 2012).

And on top of that, the downsides of the traditional crowdsourcing become more imminent when you leave the Anglophone world. Our potential crowd is geographically scattered in Russia. This crowd is linguistically heterogeneous, speaking 17 different languages. In many cases languages are close to extinction or longing for language revitalization, and the native speakers do not always have Internet access, so an open call for crowdsourcing would not have produced appeasing results for linguists. Thus, one has to identify carefully the potential niches to complete the needed tasks.

When using the help of a crowd in a project that is aiming to support both linguistic research and survival of endangered languages, the approach has to be a different one. In nichesourcing, the 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 complex tasks with high-quality product expectations found in nichesourcing. Communities have a purpose and identity, and their regular interaction engenders social trust and reputation. These communities can correspond to research more precisely (de Boer et al., 2012). Instead of repetitive and rather trivial tasks, we are trying to utilize the knowledge and skills of citizen scientists to provide qualitative results.

In nichesourcing, we hand in such assignments 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. For instance, there is lack of Hill Mari words and terminology in anatomy. We have digitized 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 nichesourcing’s perspective, it is essential that altruism play a central role when the language communities are involved. In nichesourcing, our goal 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 an online dictionary, which is made freely available for the public, so the society can benefit, too. 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.

**References**

**De Boer, V., et al.** (2012). Nichesourcing: Harnessing the Power of Crowds of Experts. In *Knowledge Engineering and Knowledge Management*. Lecture Notes in Computing Science 7603. Berlin: Springer, pp. 16–20.

**Fish, S.** (2012)**.** Mind Your P’s and B’s. The Digital Humanities and Interpretation. *New York Times*, 23 January 2012, http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/?\_r=0.

**Nygren, T., Foka, A. and Buckland, P. I.** (2014). The Status Quo of Digital Humanities in Sweden: Past, Present and Future of Digital History*. H-Soz-Kult*, 23 October 2014, http://hsozkult.geschichte.hu-berlin.de/forum/id=2402&type=diskussionen.

**Rueter, J.** (2013). Open-Source Morphological Parsers for Uralic Minority Languages. Presentation at *Finnic Languages, Culture, and Genius* *Loci*, University of Tartu, Tartu, Estonia, 7 March 2013.

**Sula, C. A.** (2013). Digital Humanities and Libraries: A Conceptual Model. *Journal of Library Administration,* **53**(1): 10–26, http://dx.doi.org/10.1080/01930826.2013.756680.

**van Hemel, W.** (2014). OCRUI: Interface for the Correction of OCR Text Material. Presentation at *OCR Webinar*, National Library of Finland, Helsinki, 5 March 2014, http://blogs.helsinki.fi/fennougrica/2014/03/06/webinar-recap/.