Digitisation project of kindred languages continues

The National Library's digitisation project of kindred languages, piloted in 2012, will continue in 2014–2015 with funding from the Kone Foundation. The material involved in the project constitutes the world’s most extensive research resource of Uralic languages. The material that the project produced will be made available for both researchers and the general public through the Fenno-Ugrica collection maintained by the National Library.

The material digitised in the pilot project (2012–2013) includes approximately 17,000 pages of publications in the Mari, Mordva, Ingrian and Veps languages, comprising 156 monographs, most of which are textbooks and dictionaries from the early Soviet era. In addition to the monographs, the digitised material includes close to 25,000 pages of Mari and Mordvin newspapers primarily from the 1920s and 30s. The production system prepared during the pilot phase will be expanded further to exploit research in the Uralic languages and to promote crowdsourcing in them as well.

The purpose of the two-year project is to digitise and publish close to 1,100 monograph and 51 newspaper titles. According to the related plan, this means approximately 88,300 monograph pages and 72,500 newspaper pages. The material to be digitised has been selected together with researchers and is considered useful for research primarily in Finno-Ugrian studies. The project will also render previously inaccessible material public for research use.

Several criteria, defined together with researchers, were employed in the selection of the materials. The key criterion was when the contemporary written language was created and became established. The works were selected for digitisation so that they would not only represent the innovative 1920s accurately, but also reflect the changes in language policy which occurred in the 1930s. Material from the time when the written language was being established is also important for activists seeking to preserve the language today. Neologisms from the 1920s and 1930s as well as texts that use them serve as both source material and source of innovation and inspiration for the developers of the contemporary language. Selecting such works can be considered as supportive of endangered languages and thus to promote linguistic diversity.
To expand the language selection of the pilot phase, a follow-up project will digitise material published in the Permic (Udmurt, Komi, Komi-Permyak), Ob-Ugric (Khanty, Mansi) and Samoyedic (Nenets, Selkup) languages. The extended language selection supports linguistic research conducted both under the auspices of the Language Programme of the Kone Foundation and elsewhere, in Finland and abroad.

For the medium languages (Komi, Udmurt, Erzya, Moksha, and Meadow and Hill Mari), both monograph and newspaper material has been digitised whenever possible. Digitising monograph material which has been translated from Russian into the relevant language supports the goals of the follow-up project. Such parallel titles have been selected primarily from areas of vocabulary which are rarely found in newspapers. For this purpose, the project will also digitise a large number of school books as well as public service leaflets from a total of 27 different disciplines and fields.
The digitisation project for kindred languages is also connected to research in language technologies, since one of the goals of the project can be broadly defined as the improvement of the use methods and usability of digital library and archive materials. In addition to accessibility to Finno-Ugrian material, the project promotes methods which allow raw digitised data to be refined into more usable material. In the digitisation project of kindred languages, these methods mean increasing optical character recognition (OCR) in the digitised material, formatting the material into paragraphs and, above all, developing the OCR editor intended for language correction, thus enabling the correction of errors made in conjunction with digitisation and optical character recognition in an effective manner and by exploiting crowdsourcing.

Jussi-Pekka Hakkarainen

Further information
http://www.nationallibrary.fi/services/digitaalisetkokoelmat/kindred.html
http://blogs.helsinki.fi/fennougrica/
http://fennougrica.kansalliskirjasto.fi/

Contact information
Jussi-Pekka Hakkarainen
Project Manager
PO Box 15
00014 University of Helsinki
+358 50 363 9223
kk-fennougrica@helsinki.fi

Jussi-Pekka Hakkarainen is a Project Manager in Research Library of the National Library of Finland