Freely accessible minority language resources online

by Jussi-Pekka Hakkarainen

In 2016, the National Library of Finland will carry out the *Digitisation Project of Minority Languages* funded by a grant from the Kone Foundation. In this context, minority languages include the Mordvin, Permic, Mari, Nenets, Yiddish, Sami and Romani languages. The project documents these languages by digitising resources, determining the related copyrights, publicly releasing the digitised resources in the Fenno-Ugrica Collection and extracting data from the resources for research use in third-party systems. This project is a continuation of the *Digitisation Project of Kindred Languages*, conducted in 2012–2015.

Maa *by art collective* Liiketila *combines dance with word art, visual expression and design. The piece is based on language resources digitised in the Digitisation Project for Minority Languages.*
(Photo: Heidi Kotilainen)

Aiming for language documentation

The digitisation of resources, releasing them to the public and storing them reliably can be thought to constitute language documentation. The documentation of languages has been a heatedly debated topic for the past twenty years, as its significance for preserving endangered languages has become recognised. A major shift in language documentation has occurred in that the focus has moved away from linguistic description towards a more comprehensive documentation that aims to preserve the language. Key components of language documentation are 1) storing the resources and generating the related metadata, 2) making the resources transferable, 3) generating added value by annotating, transcribing and linking, 4) archiving and publicly releasing the archived material and 5) mobilising the resources, i.e., making them accessible through third-party systems. These components form the core of the Digitisation Project for Minority Languages.

One of the most important goals of language documentation is to return the documentation to the language community. The material digitised in the project will be published in the National Library's Fenno-Ugrica Collection (fennougrica.kansalliskirjasto.fi). Open resources serve research in Uralic, Germanic and Romani studies in Finland and abroad. The project is releasing a great deal of resources which have been previously inaccessible for research.

Users outside of the academic community are likely to benefit from the resources even more than researchers. They can now freely access documents which describe their own languages and cultures to an extent which is not available through local libraries or archives. The digitised resources can help communities build their collective memory, process their past and improve language skills. The Fenno-Ugrica material can be called a "long-tail" resource, one which can be expected to have a long-term impact on local communities.

A range of resources for research and art

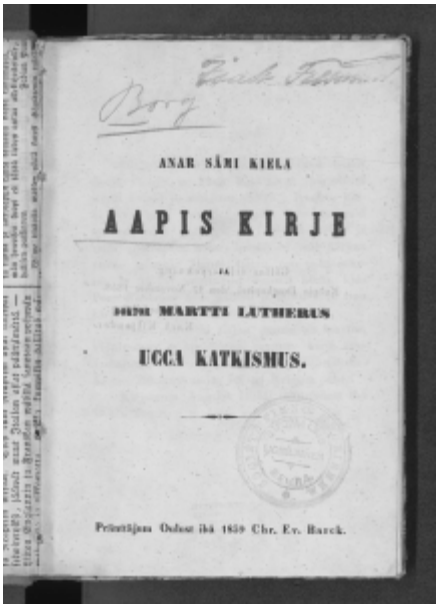
Several criteria, defined in cooperation between the National Library and researchers, were employed in the selection of the materials, the most important being accessibility, previous use in research and public use. In their original format, these resources have been difficult to access. Providing them to researchers online promotes better understanding of the detailed development of written languages and enables language documentation.

The project has focused on printed resources which comprehensively reflect the early stages of the development of the languages.

In addition to research projects, the Digitisation Project of Minority Languages cooperates with the Liiketila art collective, providing them with edited dictionaries in several Uralic languages. These dictionaries will later be used in a performance piece entitled *Maa*. The dictionaries used in the piece will be chosen with Liiketila, and can include both previously edited resources and dictionaries specifically collected for *Maa*.

Digitising resources also serves the international community

The Sami resources primarily consist of material stored in the national libraries of Finland and Russia. The most important part of the online resources is made up of material digitised from the collections of the National Library of Finland, primarily from the Lapponica Collection curated by Jakob Fellmann. The collection is very valuable, as it features resources from a broad spectrum of different areas of Sami research and constitutes a coherent compilation of the oldest Sami literature ever printed. Smaller selections from the resources will be digitised from the Fennica Collection of the National Library of Finland as well as the collections of the National Library of Russia.


ANAR SÂMI KIELA

AAPIS KIRJE

MARTTI LUTHERUS

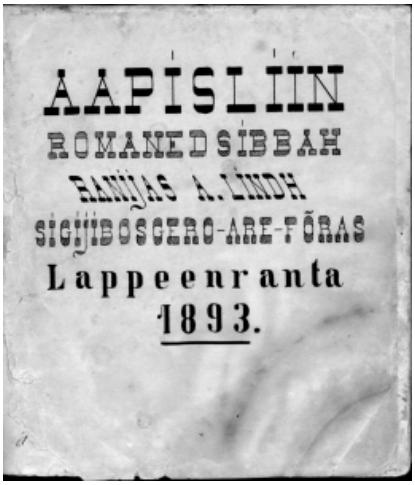
UCCA KATKISMUS.

An ABC book for Inari Sami and Martin Luther's *Small Catechism* from 1857 have been digitised from the National Library's Lapponica Collection.

The oldest Nordic material in the Romani language is characterised by its fragmentary nature. The oldest mentions of the Romani language or its vocabulary have typically been found in unpublished or partially published

manuscripts from the late 19th and early 20th centuries. Due to the scattered material, the project is gathering early manuscripts which are relevant for research from different memory institutions and digitising them. Nordic libraries and archives have little-studied manuscripts that the project seeks to render available for researchers.

There is a clear demand for digitising Romani resources, as such material has never before been digitised to any significant degree. Research in the Romani language and culture is a growing field, characterised by many multidisciplinary connections to language disciplines, general linguistics, cultural and social anthropology, the study of religions, social sciences, education and medicine. Finnish Romani studies focus on the dialectology and grammar of the Romani language as well as the study of migration and Romani history, all of which can be researched with the help of the digitised materials. There are approximately two dozen scholars of Romani studies in Finland, and a few hundred in the world.



Manuscript for a Romani textbook by Adam Lindh from 1897.

The Romani dictionaries compiled by H. A. Reinholm and Arthur Thesleff in the late 19th century constitute the core of the collection of Romani manuscripts. Other unpublished Romani language resources to be digitised in the project include Christfried Ganander's Undersökning om de så kallade tattare from 1780, Adam Lindh's manuscript in Romani from 1897 and Romani dictionaries collected by Paul Ariste in the 1930s and 1940s. In addition, the project will digitise practically all Romani-language material printed in the Soviet Union before World War II.

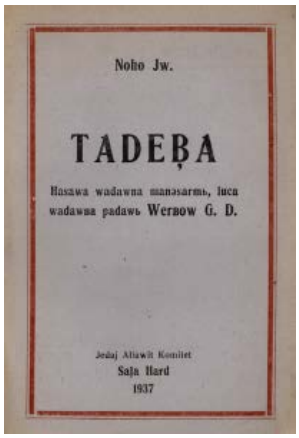
Coverleaf for Romani translation of the fairy tale "Сказка о рыбаке и рыбке" by A.S. Pushkin from 1936.

The Komi-language resources were selected together with FU-Lab, the department of language technology at the Komi Republic Academy of Government Service and Management (KRAGSiU) in Syktyvkar. FU-Lab primarily creates keyboard systems, online dictionaries and proofreading software for Permic languages, but it also generates corpora of literature published in Komi-Zyrian and Komi-Permyak. For FU-Lab, the Digitisation Project of Minority Languages provides material from the 1920s and 1930s which they would otherwise be unable to access locally.

The project will also digitise Yiddish resources from the national libraries of both Finland and Russia. A total of 143 works of literature will be digitised from the National Library of Finland's Hebraica Collection as well as the Yiddish editions of *Pravda* from 1918. The Hebraica Collection consists of legal deposit copies submitted in the Russian Empire. The time period from which the Hebraica Collection originates is considered to be the golden age of Yiddish literature. The most interesting titles in the Hebraica Collection include the hundreds of leaflets which featured short educational, romantic and thrilling stories as well as collections of poems. Most of these works represent popular literature, which has become an increasingly popular topic of surveying and research also in other language areas.

Digitising Nenets resources and adding them to the collection is a natural expansion of the Nenets materials digitised during the Digitisation Project for Kindred Languages. Texts written by or received from Nenets people from the early decades of the Soviet Union are difficult to access in Finland, Russia and elsewhere in the world. They have received little research attention, as they represent neither prose nor folklore. Instead, the texts combine Western and Soviet literary modes with local oral Nenets traditions. Text analysis can reveal the processes underlying rapid linguistic change, the impact of a multilingual and multicultural environment on linguistic expression as well as the impacts of Soviet language and cultural policy on minority languages in general.

The 28 November 1918 issue of the Yiddish-language edition of *Pravda* is stored in the Hebraica Collection of the National Library of Finland.



Tadeḫa - a Nenets-language play about a shaman, written by I. Noho, from 1937

Periodicals and newspapers will be digitised to supplement the collections. The primary focus of the digitisation of newspapers will be on Karelian publications not available in Finland. Digitising these materials and adding them to the collections will also benefit history researchers and local users.

The mission of the Tver Karelian newspaper *Kolhozoin puoleh* included promoting the Karelian language.

Fenno-Ugrica has an international user base

The National Library's Fenno-Ugrica Collection has been freely available to researchers and the general public since June 2013. The collection has become more popular since the end of 2015. By April 2016, material had been accessed from the collection more than 620,000 times. This means that, on average, Fenno-Ugrica resources were accessed 18,000 times a month.

The user base has also become broader geographically – the collection has been accessed from 96 different countries, primarily from Russia (59%) and Finland (24%). The number of new sessions increased by nearly 70% during 2015. It can be said that this collection has become established as an important source of resources published in the Finno-Ugrian languages.

The author is a project manager at the National Library of Finland.

More information:
https://www.kansalliskirjasto.fi/en/projects/digitisation-project-of-minority-languages
http://blogs.helsinki.fi/fennougrica/