Digitization Project of Kindred Languages
Materials, methods and tools for researchers

Jussi-Pekka Hakkarainen
Project Manager

Soome-ugri keelte andmebaasid ja e-leksikograafia
Eesti Keele Instituut
Tallinn, 19.11.2014
Introduction

- An overview of the Digitization Project of Kindred Languages.
- Kone Foundation Language Programme and our role.
- Services: Fenno-Ugrica, Uralica
- Co-operation with the scholars
- Tools and methods for enhancing the data
- Impact on research and society
Overview of the Project


- Within the project we will digitize materials in 17 Uralic languages as well as develop tools to support linguistic research and citizen science.

- Through this project, researchers will gain access to new materials [sic!] which they have not been able to study before and to which all users will have open access regardless of their place of residence.
The project is financially supported by the Kone Foundation and is part of its **Language Programme**.

The main objective of the Language Programme is to **advance the documentation** of small Finno-Ugrian languages, the Finnish language, and minority languages in Finland.

Our target within the Language Programme is to make sure that **the new corpora** in Uralic languages are made available for the open and interactive use of both the academic and the language communities.
Materials and Collection

- The project seeks to digitize and publish around 1200 monograph titles and more than 100 newspapers titles in various Uralic languages.

- The digitization will be completed in early 2015, and the online collection, Fenno-Ugrica, will consist of 110,000 monograph pages and 90,000 newspaper pages.

- The majority of the digitized materials belong to the collections of the National Library of Russia in Saint Petersburg and the copyrights are sorted in cooperation with the National Library Resource in Moscow.
Kansalliskirjastoon Fenno-Ugrica on suomalais-ugrilaisen julkaisujen digitaalinen kokoelma. Kokoelma koostuu 17 uralilaisella kielellä julkaistuista monografiosta ja sanomalehdistä. Digi-toijaa monografiainmekkeitä on kokoelmassa tällä hetkellä yli 1100 ja sanomalehtinimekkeitä yli 100.


Kielentutkimoiden käyttöön on Sukukielen digitoimiprojektissa tuotettu myös vapaa lähdekoodin OCR-editori, jonka avulla Fenno-Ugrican teosten konetunnistettava teksti voidaan korjata ja muokata. Käyttöoikeuksia Fenno-Ugrican aineistojen muokkaamiseen myönnetään ensisijaisesti suomalais-ugrilaisen kielen tutkijoille ja käyttöoikeuksien hallinnoidaan Kansalliskirjaston Sukukielen digitoimiprojektissa.

Hankkeen edistymistä voi seurata projektin blogin välyksellä.

Yhteydenotot ja tiedostelut: kk-fennougrica@helsinki.fi

**Kokoelmat**
- Eesti Keele Instituut [63]
- Monografiat [1128]
- Sanomalehdet [5248]
Bukvari izoroin škouluja vart
Iljin, N. A.; Junus, V. I.; Ильин, Н. А.; Юнус, В. И.

The permanent address of the publication is http://urn.fi/URN:NBN:fi-fe2013123010160

Name: bx000010952.pdf
Size: 52.57Mb
Format: PDF
Description: User copy PDF

developer/downloads

Name: 04f6aaa2-442a-449 ...
Size: 222.9Kb
Format: Unknown
Description: Alto XML files of ...
developer/downloads

Title: Bukvari izoroin škouluja vart
Alternative title: Букварь для ижорских школ
Author: Iljin, N. A.; Junus, V. I.; Ильин, Н. А.; Юнус, В. И.
Published: Moskova / Leningrad : Riikin ucebno-pedagogiceskoi izdatelstva, 1936
Subject: aapiset; inkeroin kieli; ижорский язык
Materials and Collection

Repo i kana. Repo repi kanaan. Miko ammu repo!

\[16\overline{40}\] THE NATIONAL LIBRARY OF FINLAND – Research Library
## Languages of Publications

<table>
<thead>
<tr>
<th>Baltic Finns</th>
<th>Mari</th>
<th>Ob-Ugric</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Ingrian</td>
<td>• Meadow Mari</td>
<td>• Khanty</td>
</tr>
<tr>
<td>• Veps</td>
<td>• Hill Mari</td>
<td>• Mansi</td>
</tr>
<tr>
<td>• Karelian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• [Livonian]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Permic</th>
<th>Sami</th>
<th>Mordvinic</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Udmurt</td>
<td>• Skolt</td>
<td>• Erzyan</td>
</tr>
<tr>
<td>• Komi-Zyrian</td>
<td></td>
<td>• Moksha</td>
</tr>
<tr>
<td>• Komi-Permyak</td>
<td></td>
<td>• (Shoksha)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Samoyedic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Nenets</td>
<td></td>
</tr>
<tr>
<td>• Selkup</td>
<td></td>
</tr>
</tbody>
</table>
Languages of Publications
Selection Criteria of Material

- The selection of the materials has been made in co-operation with the researchers and we used several criteria upon the selection of material:
  - genesis and consolidation period of literary languages
  - availability of material in Finnish libraries
  - online access to Russian collections
  - locality – the languages of peripheries is more tempting
  - cost efficiency – loads of parallel titles (translations)
Uralica

- **Uralica** – a search service of the digitized books in Uralic languages

- The Uralica is developed by the National Library of Finland and funded by the Finnish Ministry of Education and Cultures. Uralica is a publicly accessible and open infrastructure information system that **provides access to the digital collections** of the languages of the Uralic language family of **funds of various libraries**.

- Currently more than 9000 links to items from eight institutions, including the **University of Tartu Library**. Five more institutions to join in 2014.
The digitization of materials in Finno-Ugrian languages in the national libraries operating in Russia has grown significantly in past few years. Many libraries have established digital collections and given the public either the full or restricted access to the materials.

The dissemination and harmonization of information has been overlooked. No control how much and which materials are digitized.

This complicates the accessibility of these materials and the systematic planning of digitization projects and may result in redundant work, a highly inefficient use of resources.
Uralica – a search service of the digitized books in Uralic languages
...and some more
Currently more than 8000 links to the items

Not sure what to search for? You can browse the entire collection.

The Uralica is developed by the National Library of Finland and funded by the Finnish Ministry of Education and Cultures. Uralica is a publicly accessible and open infrastructure information system that provides access to the digital collections of books, maps and recordings in the languages of the Uralic language family of funds of various libraries.

News from blog
24.10. National Library of Russia joins Uralica
21.10. Consider It Done! Monograph Frenzy at Fenno-Ugrica
9.10. Time to Decide for Ob-Ugric Languages: How Many Written Varieties Does a Language Need?
7.10. Next Stop: Bibliotheca Baltica at Sodertom
24.9. In Creation of New Partnerships
Search: liivin

Search Results 1 - 8 / 8

Narrow Search

Liivin katko
- book
by: Wrik, Kalevi Published 1989
- Contents
- Full text

Viela liivin kielenintonistoa
- book
by: Kettunen, Lauri Published 1937
- Contents
- Full text

Tutkimus liivin nuorista lainasanoista
- book
by: Posti, Lauri Published 1974
- Contents
- Full text

Liivin kielen näytteitä
- book
by: Suhonen, Seppo Published 1975
- Contents
- Full text

Subject
liivi keel
keelenäited (3)
eeesti keel (2)
rooptekstid (2)
soome keel (2)
mor...
Vielä liivin kielen intonatioista
Kettunen, Lauri

URI: http://hdl.handle.net/10062/16361
Date: 1937

Description:
Separaat väljaandest: Virittäjä, 1937

Show full item record

Files in this Item

<table>
<thead>
<tr>
<th>Thumbnail</th>
<th>Name: kettunen_vielai.pdf</th>
<th>View/Open</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>506.8Kb</td>
<td>PDF</td>
</tr>
</tbody>
</table>

This item appears in the following Collection(s)

- Liivi keel ja kultuur [71]
The Digitization Project of Kindred Languages is also linked with language technology. The one of the key objectives is to improve the usage and usability of digitized content. During the project we are advancing methods that will refine the raw data for further use.

The machined-encoded text (OCR) contain quite often too many mistakes to be used in research. The mistakes in OCR’d texts must be corrected. In order to meet the objective, we have developed an open source code OCR editor that enables the editing of erroneous text.
OCR Editor

- The editor is an interactive web application, enabling many people to contribute simultaneously and revise the OCR text of source materials in the system.

- The project has multiple goals:
  - make sure the transfer of source material into the digital age does not in anyway take away any of the quality of the original
  - make it easier to study the material by availability and dissemination (i.e. internet); "editor as reader" or distributor
  - make automated corpora or word lists to improve the editor itself and other tools
Sarn kalanikha i kalaizeh polin

A. S. Puškin

Sarn kalanikha i kalaizeh polin

1

Eli ukoine mamsinct
Ani sinizen merenno;
Eliba kulus hö mahizes pertizes
Kuumekyme ruuno kuumen vodenke.
Ukoine verkol kalati kalaizen,
Mams hänen kezeri icoze pyuhaizén.
Kerdan hän merhe lykäzi verkon,
Mudanke yhten verkon hän vedí.
Toizen hän kerdan lykäzi verkon,
verkon hän meren hiinänke vedí.
Kuumañden kerdan hän lykäzi verkon,
kalaizen yhten vedí hän verkol,
Kalaizen ij prostan, kalaizen kuudaizen.
Zavodib mol’das ku kuudaine kalaine,
Ani ku mehen virkab hän ānel;
"Pāsta mindai, ukoine, merhe.
Kal’hen otkupan icesain andan;
Mil sinä ofotid, sil minä maksan’â?"
Pōlgastui ukoine, cudha hän lange;
Kalažab kuumekyme kuumen hän vodenke
Kuleske ij, miše pagaiize kala.
Pāstī hun kuudaizen kalaizen merhe,
sanui hän laskvästi hännele valhen;
Syndunke kuudaine kaleine måne.
Otkupad sinun minij jì tarâ;
Måne zo holetta sinizehe merhe,
meren prostoras holetta gulâ.
Crowdsourcing the Finno-Ugrian material

- We have estimated that the Fenno-Ugrica collection will contain around 200,000 pages of editable text.

- The researchers cannot spend so much time with the material that they could retrieve a satisfactory amount of edited words, so the aid of a helping hand is truly needed.

Could crowdsourcing be used here to gain results?

- (Besides, the Kone Foundation required this from us)
Citizen Science and Crowdsourcing

- **Citizen Science** = Interactive research that includes the participation of researchers, students and any interested citizens. It is based on the work of trustworthy volunteers, who help in observation, measuring and calculation work. Citizen science is a way of obtaining new material and carrying out large-scale proofing.

- **Crowdsourcing** = Interactive research can also benefit from crowdsourcing i.e. collaborating with an indeterminate group to carry out development in research. For instance, by crowdsourcing one can solve problems that computers cannot yet solve.
Downsides of Crowdsourcing

- The targets have often been split into several microtasks that do not require any special skills from the anonymous people.

- This way of crowdsourcing may produce quantitative results, but from the research’s point of view, there is a danger that the tasks are too hard to handle by the faceless crowd and the needs of linguistic research are not necessarily met.

- The remarkable downside is the lack of shared goal or social affinity. There is no reward in traditional methods of crowdsourcing.
Nichesourcing and Language Communities

- **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.

- These communities can correspond to research more precisely. Instead of repetitive and rather trivial tasks, we are trying to utilize the knowledge and skills of citizen scientists to provide qualitative results.
Nichesourcing and Language Communities

- Some selection must be made, since we are not aiming to correct all 200,000 pages which we have digitized, but 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/pages in such fields of vocabularies, where the researchers do require more information.

- There’s a lack of Hill Mari words in anatomy. We have digitized the books in medicine and we could try to track the vocabulary of human organs by editing and collecting **the related words** with the OCR editor.
Nichesourcing and Language Communities

- Some selection must be made, since we are not aiming to correct all 200,000 pages which we have digitized, but 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/pages in such fields of vocabularies, where the researchers do require more information.

- There’s a lack of Hill Mari words in anatomy. We have digitized the books in medicine and we could try to track the vocabulary of human organs by editing and collecting **the related words** with the OCR editor.
Nichesourcing and Language Communities

- When the language communities involve, it is essential that the altruism plays a central role.

- Upon the nichesourcing, our goal is to reach a certain level of interplay, where the language communities would benefit on the results. For instance, the corrected words in Ingrian will be added onto the online dictionary, which is made freely available for the public.

- 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”, the research and the society.
End-Products for End-Users

- What to do with the corrected material?
- Library considers the edited material is **raw data** that needs to be released to support the researchers’ aspiration, even though the data sets would be incomplete to some extent.
- The distribution of raw data, however, is still a matter of discussion at the Library and we have no policy how to make the raw data available.
- The raw data hub, data.kansalliskirjasto.fi could be a solution.
End-Products for End-Users

- We will create the corpora ourselves and release the data for other operators in Fenno-Ugrica as wordlists.

- No resources or in-house knowledge for the linguistic work.

- Material will be available also in Korp, which is the concordance search tool of the Swedish Language Bank.
Repo-operaatio (repo operation) liikviditeettiä lisäävää takaisinostosopimuksen perustuva käänkisoperaatio.

Repojoen osalta ei ole edellyttää sektorikohtaisia erilaisia kopiahallinta- ja yritysyrityksiä kaikkien rahaliitoon asennetut aineet.

Repojoen mukaan valmistelun koko on saatu kasvaa, mutta se on lainatusten perusteella luonnollisesti rajoitettu.

Käänteisoperaatioilla tarkoitetaan joko operatiota, jossa keskuspankki ostaa ("käänteinen eioperaatio") tai myy ("käänteinen repo") arvopapereita takaisinostosopimuksen perustuva käänkisoperaatio.

Käänteisoperaatioilla tarkoitetaan joko operatiota, jossa keskuspankki ostaa ("käänteinen repo") tai myy ("käänteinen ei-operaatio") arvopapereita takaisinostosopimuksen perustuva käänkisoperaatio.
Repo i kana.
Repo repi kalan.
Miko ammu repo!
Impact on Research and Society

- **Huge impacts** on society and research are expected, but how to measure the societal impact?

- Do we really need to know what the impact will be and how valuable it is?

  **Get beyond the number games!**

- Once the digital resources and tools for enriching the data will be used, the change will take place and **a wider set of opportunities** will be available to different communities, like native-speakers and academic.
Some Conclusions

- The **Fenno-Ugrica collection** and its materials are only one part of the work, albeit important due to their rare use in research.

- The co-operation with Russian libraries has intensified in several fields due to digitization and **Uralica**. New contacts to Komi, Mari-El, Mordovia, Udmurtia, Murmansk, Tver and Russian State Library have been made and big Russian libraries are following the trends too.

- The researchers of Uralic studies are leaning more towards us. Does the library fill the vacuum in Finland?
Some Conclusions

- National Library of Finland has went beyond the traditional framework of libraries in post-production, crowdsourcing and data releases.

- The machine-encoded texts do contain errors that need to be removed in order to match them with the researchers’ needs. The correction of the words will be done with the help of OCR editor and the tasks are distributed to the crowd.

- Instead of releasing tasks to the faceless crowd, we interplay with the language communities for the research’s and society’s mutual benefit.
Contact Details

jussi-pekk.hakkarainen@helsinki.fi

fennougrica.kansalliskirjasto.fi
uralica.kansalliskirjasto.fi
blogs.helsinki.fi/fennougrica