

# **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

- The National Library of Finland is implementing the Digitization Project of Kindred Languages in 2012–16. Pilot phase in 2012-2013, the continuation project in 2014-2015 with possible extension to 2016.
- 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.

# Kone Foundation Language Programme

- 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.



Kansalliskirjaston Fenno-Ugrica on suomalais-ugrilaisen julkaisujen digitaalinen kokoelma. Kokoelma koostuu 17 uralilaisella kielellä julkaistuista monografioista ja sanomalehdistä. Digitoituja monografianimekkejä on kokoelmassa tällä hetkellä yli 1100 ja sanomalehtinimekkejä yli 100.

Fenno-Ugrican aineisto on tuotettu Kansalliskirjaston toteuttamassa [Sukukielten digitoitiprojektissa](#), joka on [Koneen Säätiön](#) rahoittaman [Kieliohjelman osahanke](#). Kokoelman aineisto sijaitsee [Venäjän Kansalliskirjastossa](#) Pietarissa, missä aineisto myös digitoitu. Aineistoa koskevia tekijänoikeuksia on selvitetty yhdessä moskovalaisen [National Library Resource](#)n kanssa. Liivinkieliset aineistot on digitoitu [Viron kielen instituutissa](#), Tallinnassa. Fenno-Ugrican aineisto on linkitetty myös [Uralica-portaaliin](#), joka on avoin tietojärjestelmä eri kirjastoissa digitoituille suomalais-ugrilaisille kirjoille, kartoille ja äänitteille.

Kielentutkijoiden käyttöön on Sukukielten digitoitiprojektissa tuotettu myös vapaa lähdekoodin OCR-editori, jonka avulla Fenno-Ugrican teosten konetunnistettua tekstiä voidaan korjata ja muokata. Käyttöoikeuksia Fenno-Ugrican aineistojen muokkaamiseen myönnetään ensisijaisesti suomalais-ugrilaisen kielten tutkijoille ja käyttöoikeuksien hallinnoidaan Kansalliskirjaston Sukukielten digitoitiprojektissa.

Hankkeen edistymistä voi seurata [projektin blogin](#) välityksellä.

Yhteydenotot ja tiedustelut: [kk-fennougrica@helsinki.fi](mailto:kk-fennougrica@helsinki.fi)

## Kokoelmat

- [Eesti Keele Instituut](#) [63]
- [Monografiat](#) [1128]
- [Sanomalehdet](#) [5248]

## Selaa Fenno-Ugricaa

- [Nimekkeet](#)
- [Tekijät](#)
- [Julkaisuaajat](#)
- [Asiasanat](#)
- [Uusimmat](#)
- [Selaa kielen mukaan](#)
- [Sivukartta](#)

## Omat tiedot

- [Kirjautu sisään](#)
- [Rekisteröidy](#)

KONEEN SÄÄTIÖ



## 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  
simplestats.downloads



[View/Open](#)



Name: 04f6aaa2-442a-449 ...

Size: 222.9Kb

Format: Unknown

Description: Alto XML files of ...  
simplestats.downloads



[View/Open](#)

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 izdatel'stva, 1936

Subject: aapiset; inkeroisen kieli; ижорский язык

### This Collection

- [Titles](#)
- [Authors](#)
- [By Issue Date](#)
- [Subjects](#)
- [By Submit Date](#)
- [Browse by languages](#)
- [Communities & Collections](#)

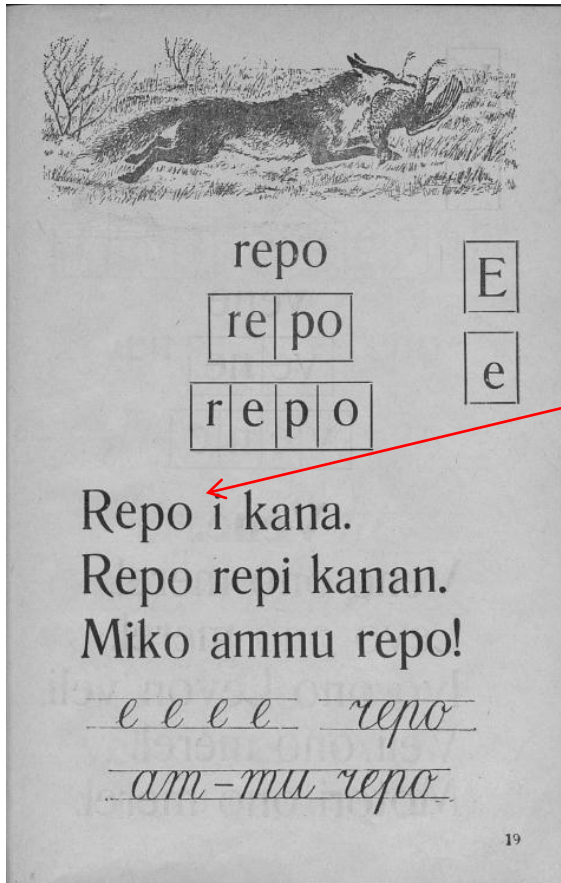
### My Account

- [Login](#)
- [Register](#)

KONEEN SÄÄTIÖ



# Materials and Collection



- <TextLine HPOS="283" VPOS="1461" WIDTH="798" HEIGHT="158">
 <String HPOS="283" VPOS="1461" WIDTH="326" HEIGHT="158" CONTENT="Repo"/>
 <SP HPOS="610" VPOS="1473" WIDTH="62"/>
 <String HPOS="673" VPOS="1473" WIDTH="24" HEIGHT="106" CONTENT="i"/>
 <SP HPOS="698" VPOS="1461" WIDTH="66"/>
 <String HPOS="765" VPOS="1461" WIDTH="316" HEIGHT="122" CONTENT="kana."/>
 </TextLine>
- <TextLine HPOS="281" VPOS="1651" WIDTH="1084" HEIGHT="160">
 <String HPOS="281" VPOS="1651" WIDTH="328" HEIGHT="160" CONTENT="Repo"/>
 <SP HPOS="610" VPOS="1693" WIDTH="62"/>
 <String HPOS="673" VPOS="1663" WIDTH="230" HEIGHT="146" CONTENT="repi"/>
 <SP HPOS="904" VPOS="1651" WIDTH="60"/>
 <String HPOS="965" VPOS="1651" WIDTH="400" HEIGHT="120" CONTENT="kanan."/>
 </TextLine>
- <TextLine HPOS="279" VPOS="1843" WIDTH="1128" HEIGHT="154">
 <String HPOS="279" VPOS="1845" WIDTH="320" HEIGHT="120" CONTENT="Miko"/>
 <SP HPOS="600" VPOS="1885" WIDTH="66"/>
 <String HPOS="667" VPOS="1881" WIDTH="366" HEIGHT="84" CONTENT="ammu"/>
 <SP HPOS="1034" VPOS="1881" WIDTH="66"/>
 <String HPOS="1101" VPOS="1843" WIDTH="306" HEIGHT="154" CONTENT="repo!"/>
 </TextLine>



# Languages of Publications

## Baltic Finns

- Ingrian
- Veps
- Karelian
- [Livonian]

## Permic

- Udmurt
- Komi-Zyrian
- Komi-Permyak

## Mari

- Meadow Mari
- Hill Mari

## Sami

- Skolt

## Samoyedic

- Nenets
- Selkup

## Ob-Ugric

- Khanty
- Mansi

## Mordvinic

- Erzyan
- Moksha
- (Shoksha)

# Languages of Publications

## F Suomalais-ugrilaiset kielet

## S Samoedikielet

### FO Itämerensuomalaiset kielet

- FO1 suomi
- FO2 karjala
- FO3 vepsä
- FO4 inkeroinen
- FO5 viro
- FO6 vatja
- FO7 liivi

### FS Saamelaiskielet

- FS1 Länsisaamelaiset kielet
- FS2 Keskisaamelaiset kielet
- FS3 Itäsaamelaiset kielet

### FU Ugrilaiset kielet

- FU1 unkari
- FU2 mansi / voguli
- FU3 hanti / ostjakki

### FP Permläiset kielet

- FP1 komi(syrjääni)
- FP2 komipernjakki
- FP3 udmurtti / votjakki

### FW Volgalaiset kielet

- FW1 mari / šeremissi
- FW2 mordva

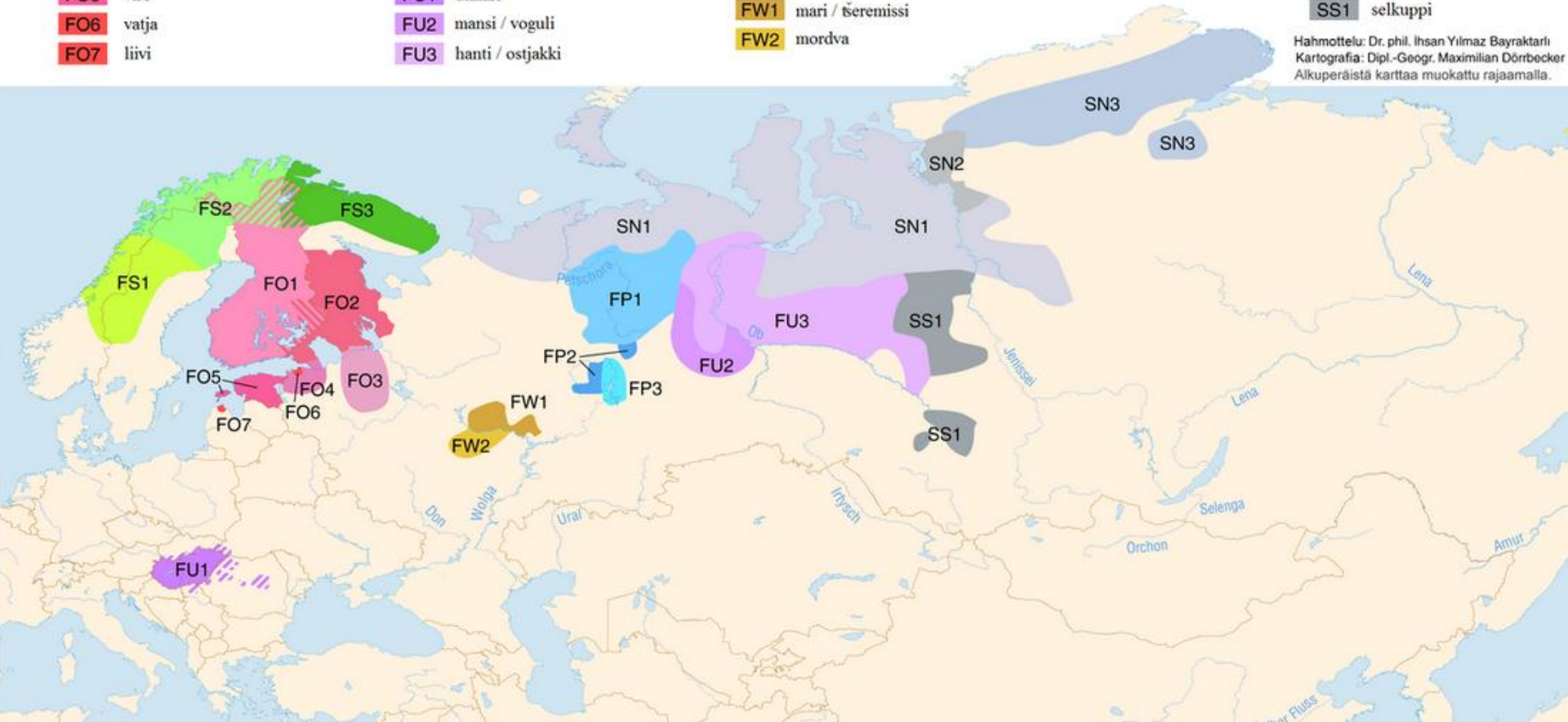
### SN Pohjoissamoedilaiset kielet

- SN1 nenetsi
- SN2 enetsi
- SN3 nganasani

### SS Eteläsamoedilaiset kielet

- SS1 selkuppí

Hahmottelu: Dr. phil. İhsan Yılmaz Bayraktarlı  
Kartografia: Dipl.-Geogr. Maximilian Dörrbecker  
Alkuperäistä karttaa muokattu rajaamalla.



# 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.

# Uralica

- 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.



# KANSALLISKIRJASTO URALICA

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

***...and some more***

***Currently more than 8000 links to the items***

Find...

All Fields



Advanced Search

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 Södertörn](#)
- 24.9. [In Creation of New Partnerships](#)

liivin

✕ Title



Advanced Search

## Search: liivin

Search Results 1 - 8 / 8

Sort

Relevance ▼

Results per page

20 ▼

### Narrow Search

Metadata Publisher:  
University of Tartu Library



Subject: liivi keel



#### Publication Type



#### Author



#### Year of Publication



#### Metadata Publisher

University of Tartu Library

#### Subject

liivi keel

keelenäited (3)

eesti keel (2)

rööptekstid (2)

soome keel (2)

more ...



#### Liivin katko



book

by: Wiik, Kalevi Published 1989

Contents □

Full text

#### Vielä liivin kielen intonatioista



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

## Vielä liivin kielen intonatioista

Kettunen, Lauri

URI: <http://hdl.handle.net/10062/16861>

Date: 1937

### Description:

Separaat väljaandest: Virittäjä, 1937

[Show full item record](#)

### Files in this item

<a href="#">Thumbnail</a>	<b>Name:</b> kettunen_viela.pdf	<a href="#">View/Open</a>
	<b>Size:</b> 506.8Kb	
	<b>Format:</b> PDF	

### This item appears in the following Collection(s)

- [Liivi keel ja kultuur](#) [71]

### Search DSpace

- ☒ Search DSpace  
☐ This Collection

[Advanced Search](#)



### Browse

All of DSpace  
[Communities & Collections](#)  
[By Issue Date](#)  
[Authors](#)  
[Titles](#)  
[Subjects](#)  
This Collection  
[By Issue Date](#)  
[Authors](#)  
[Titles](#)  
[Subjects](#)

### My Account

[Login](#)

### Statistics

[View Usage Statistics](#)



# Project and Linguistic Research

- 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

# OCR Editor

⏪ 🔍 🔍 ⏩ ☰ ★ ✍ ✍ 🔄 A ⌚ ⌂

Save Tag

18 / 86

A. S. Puškin

Sarn kalanikha i kalaizehe polin

1

Eli ukoine mamsinke  
Ani sinizen merenno;  
Eliba kulus hö mahizes pertizes  
Kuume kyme ruuno kuumen vodenke.  
Ukoine verkol kalati kalaizen,  
Mamş hänen kezerzi iceze pyuhaizen.  
Kerdan hän merhe lykäizi verkon,  
Mudanke yhten verkon hän vedi.  
Toizen hän kerdan lykäizi verkon,  
verkon hän meren hiinänke vedi.  
Kuumanden kerdan hän lykäizi verkon,  
kalaizen yhten vedi hän verkol,  
Kalaizen ij prostan, kalaizen kuudaizen.  
Zavodib mol'däs ku kuudaine kalaine,  
Ani ku mehen virkab hän änel:  
„Pästa mindai, ukoine, merhe.  
Kal'hen otkupan icesain andan:  
Mil sinä ofotid, sil minä maksan“?  
Pöl'gastui ukoine, cudha hän langez':  
Kalatab kuume kyme kuumen hän vodenke  
Kuleske ij, mişe pagiziz kala.

A. S. Puškin

Sarn kalanikha i kalaizehe polin

Eti ukoine mamsinke  
Ani sinizen merenno;  
Eliba kulus hö mahizes pertizes  
Kuume kyme ruuno kuumen vodenke.  
Ukoine verkol kalati kalaizen,  
Mamş hänen kezerzi iceze pyuhaizen.  
Kerdan hän merhe lykäizi verkon,  
Mudanke yhten verkon hän vedi.  
Toizen hän kerdan lykäizi verkon,  
verkon hän meren hiinänke vedi.  
Kuumanden kerdan hän lykäizi verkon,  
kalaizen yhten vedi hän verkol,  
Kalaizen ij prostan, kalaizen kuudaizen.  
Zavodib mol'däs ku kuudaine kalaine,  
Ani ku mehen virkab hän änel:  
„Pästa mindai, ukoine, merhe.  
Kal'hen otkupan icesain andan:  
Mil sinä ofotid, sil minä maksan“?  
Pöl'gastui ukoine, cudha hän langez':  
Kalatab kuume kyme kuumen hän vodenke  
Kuleske ij, mişe pagiziz kala.  
Pästi hän kuudaizen kalaizen merhe,  
sanui hän laskvaşti hänele vaihen:  
Syndunke kuudaine kalaine mäne.  
Otkupad sinun minij ij tariz;  
Mäne zo holetta sinizehe merhe,  
meren prostoras holetta guläi.

A a Ä ä Å å B b C c Ç ç D d Æ æ E e F f G g Y y I I Veps

i J j K k L l M m N n O o Ö ö P p R r S s Ş ş T t

v X x Z z Ž ž b Ъ ъ rx lh -

ui-kk.lib.helsinki.fi/editor/#

# 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.



129 korpusta valittuina — 4 848 984 674 sanetta

Hakuhistoria

Yksinkertainen Laajennettu Edistynyt

Etsi repo (substantiivi) Etsi myös alkuosa loppuosa ja samaista pien- ja suuraakkoset

Konkordanssi: osumia sivulla: 25 järjestä korpuksen sisällä: järjestämätön Tilastoja: laske tilastot tämän perusteella: sana

Konkordanssi Tilastoja Sanakuva

Tuloksia: 28 451

Edellinen 1 2 3 4 5 6 7 8 9 10 11 .. 1138 1139 Seuraava Näytä konteksti

FINN TREEBANK 3: JRC ACQUIS

Repo-operaatio (repo operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.

okkia kuin yön yli -talletuksia, irtisanomisehtoisia talletuksia ja sekkiluottoja, eli määräaikaistalletuksia, taista erittelyä sovelletaan kaikkiin rahalaitosten korkotilastoissa edellytettyihin talletuksiin ja lainoihin

Lisäliitteessä 1 oleva indikaattori 5 ja lisäliitteessä 2 oleva indikaattori 11 koskevat

otto riippuu joissakin rahaliittoon osallistuvissa jäsenvaltioissa siitä, minkä sektorin hallussa repot ovat, skään maturiteettierittelyä ei edellytetä kaikkien rahaliittoon osallistuvien jäsenvaltioiden tasolla, koska repojen oletetaan olevan pääasiassa hyvin lyhytaikaisia.

Repoja koskevia rahalaitosten korkoja ei saa kohdistaa yhteen sektoriin, vaan ne liittyvät erittelemättä m

repoja lukuun ottamatta sekä kaikkiin kantatietoja koskeviin lainakorkoihin lisäliitteen 1 mukaisesti.

repo ei jaotella maturiteettien mukaan, sillä ne ovat yleensä erittäin lyhytaikaisia instrumentteja ( niiden operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.

repo " ) tai myy ( " repo " ) arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-

repo " ) arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-operaatiota vakuu-

repo ) osalta käteisen lainan antajana oleva rahaviranomainen ei rekisteröi mitään muutoksia valuutta

repo operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.

repo / repo.cfm? institution = COMM % amp % doc \_ to \_ browse = COM / 2005 / 0115 % amp % refresh ,

repo operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.

repo " ) tai myy ( " repo " ) arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-

repo " ) arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-operaatiota vakuu-

Käänteisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa ( " käänteinen eisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa ( " käänteinen repo " ) tai myy ( " sopimuksen kohteena olevat arvopaperit säilyvät edelleen taseessa; takaisinmyyntisopimusten ( reverse

Repo-operaatio (

[ 12 ] http : / www.cc.cec / home / dgserve / sg / sgvista / i / sgvt2 /

Repo-operaatio (

Käänteisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa ( " käänteinen eisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa ( " käänteinen repo " ) tai myy ( "

## Korpus

FinnTreeBank 3: JRC Acquis

URN: urn:nbn:fi:lb-201406021

Kuvailutiedot

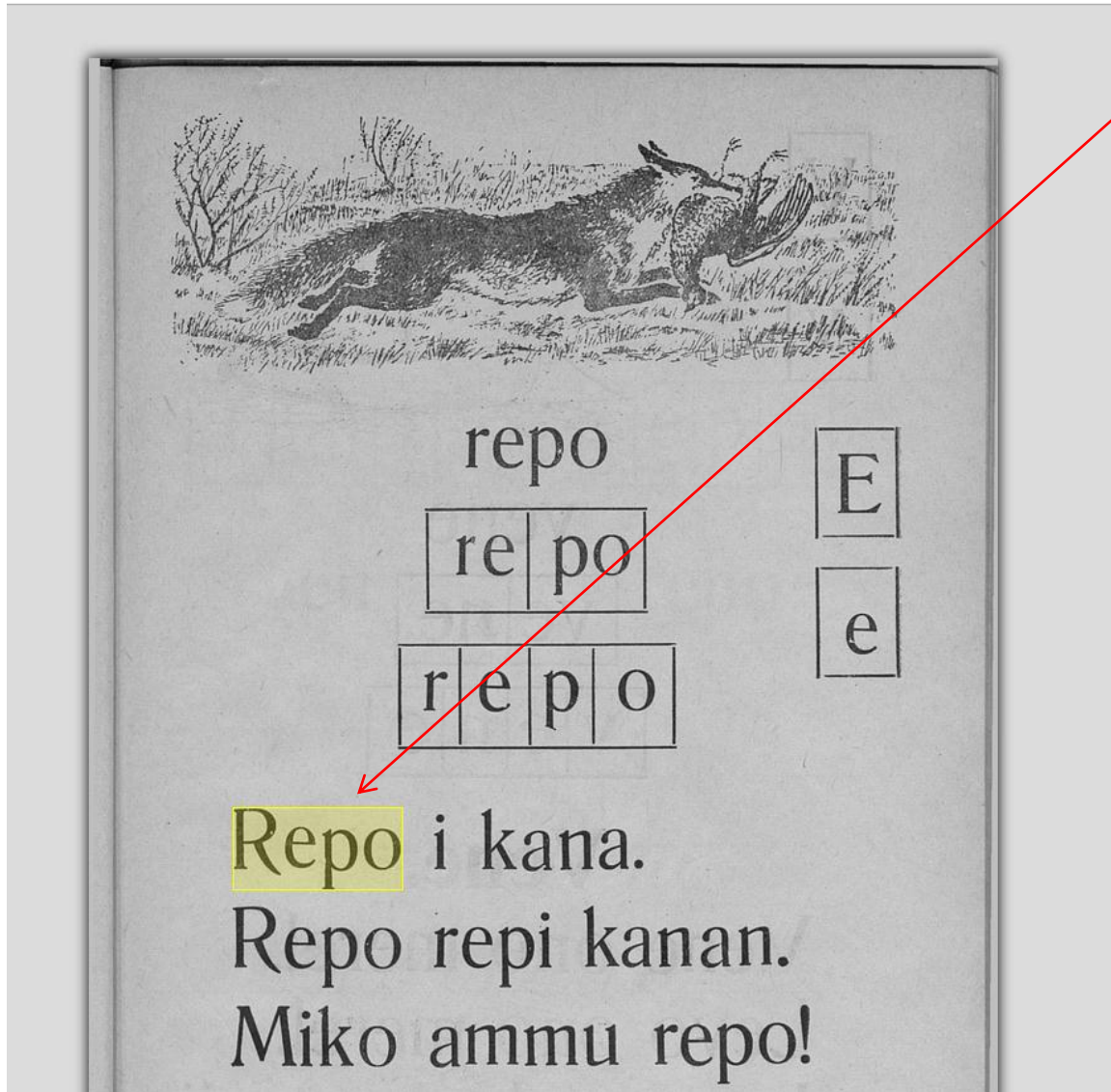
Lisenssi: CC-BY 3.0

## tekstin ominaisuudet

säädöksen otsikko: Euroopan keskuspankin suuntaviivat, annettu 31 päivänä elokuuta 2000, eurojärjestelmän rahapolitiikan välineistä ja menettelyistä (EKP/2000/7) säädöksen tunniste: JRC-ACQUIS 3200000007 Finnish URL: <http://europa.eu/...ELEX:3200000007:fi:HTML> tiedoston nimi: JRC\_Acquis Corpus/fi /2000/jrc3200000007-fi.xml rivinumero: 1022

## sanan ominaisuudet

sanaluokka: substantiivi perusmuoto: repo perusmuoto (yhdyssanarajat): repo morfologinen analyysi: N Nom Sg dependenssisuhde: attribuuitti



repo  
Repo i kana.  
Repo repi kanan.  
Miko ammu repo!  
19

# 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.

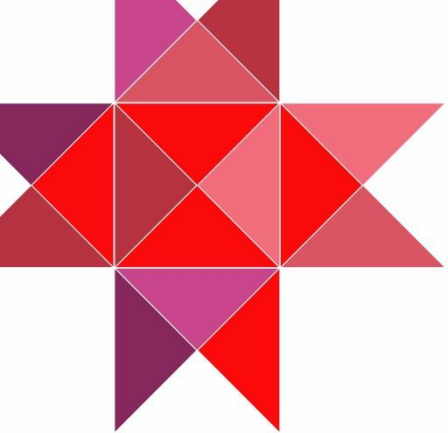
# Some Conclusions

- 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-pekka.hakkarainen@helsinki.fi](mailto:jussi-pekka.hakkarainen@helsinki.fi)

[fennougrica.kansalliskirjasto.fi](http://fennougrica.kansalliskirjasto.fi)

[uralica.kansalliskirjasto.fi](http://uralica.kansalliskirjasto.fi)

[blogs.helsinki.fi/fennougrica](http://blogs.helsinki.fi/fennougrica)