Nichesourcing the Uralic languages for the benefit of research and societies

Jussi-Pekka Hakkarainen
Project Manager

Emerging Technologies in Academic Libraries 2015
Trondheim, 20.4.2015

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 2012–15.
- 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 corpora 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 objective 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.

Fenno-Ugrica

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

Fenno-Ugrican aineisto on tuotettu Kansalliskirjaston toteuttamassa Sukukielten digitointiprojektissa, 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 Resourcen kanssa. Liivinkieliset aineistot on digitoitu Viron kielen instituutissa, Tallinnassa. Fenno-Ugrican aineisto on linkitetty myös Uralica-portaaliin, joka on avoin tietojärjestelmä eri kirjastoissa digitoiduille suomalais-ugrilaisille kirjoille, kartoille ja äänitteille.

Kielentutkijoiden käyttöön on Sukukielten digitointiprojektissa 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-ugrilaisten kielten tutkijoille ja käyttöoikeuksien hallinnoidaan Kansalliskirjaston Sukukielten digitointiprojektissa.

Hankkeen edistymistä voi seurata projektin blogin välityksellä.

Yhteydenotot ja tiedustelut:kk-fennougrica@helsinki.fi

Kokoelmat

- Eesti Keele Instituut [63]
- Monografiat [1128]
- Sanomalehdet [5248]

Selaa Fenno-Ugricaa

- Nimekkeet
- Tekijät
- Julkaisuajat
- Asiasanat
- Uusimmat
- Selaa kielen mukaan
- Sivukartta

Omat tiedot

- Kirjaudu sisään
- Rekisteröidy

KONEEN SÄÄTIÖ

16 40 KANSALLIS KIRJASTO

Bukvari iƶoroin şkouluja vart

Iljin, N. A.; Junus, V. I.; Ильин, Н. А.; Юнус, В. И.

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

View/Open

ZIP

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

View/Open

Title:	Bukvari iƶoroin şkouluja 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 izdateljstva, 1936
Subject:	aapiset; inkeroisen kieli; ижорский язык

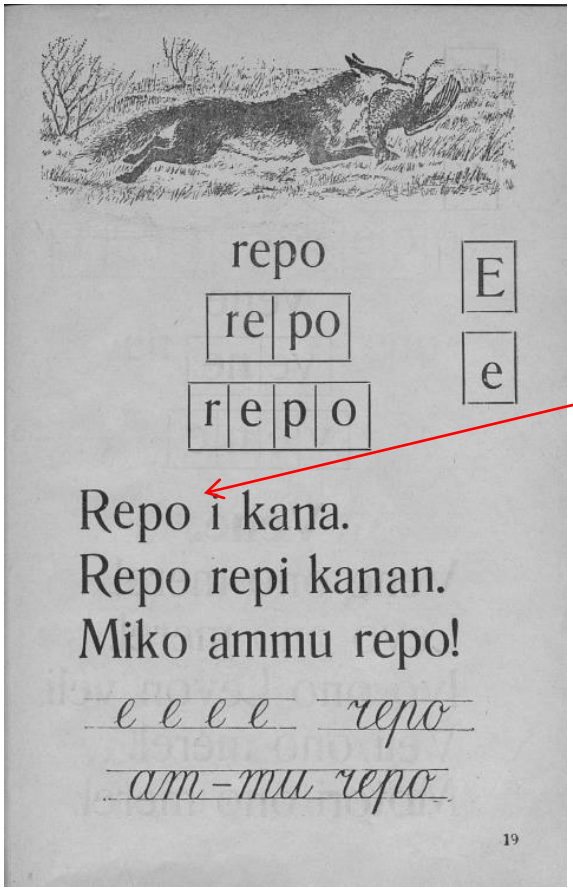
This Collection

- Titles
- Authors
- By Issue Date
- Subjects
- By Submit Date
- Browse by languages
- Communities & Collections

My Account

- Login
- Register

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

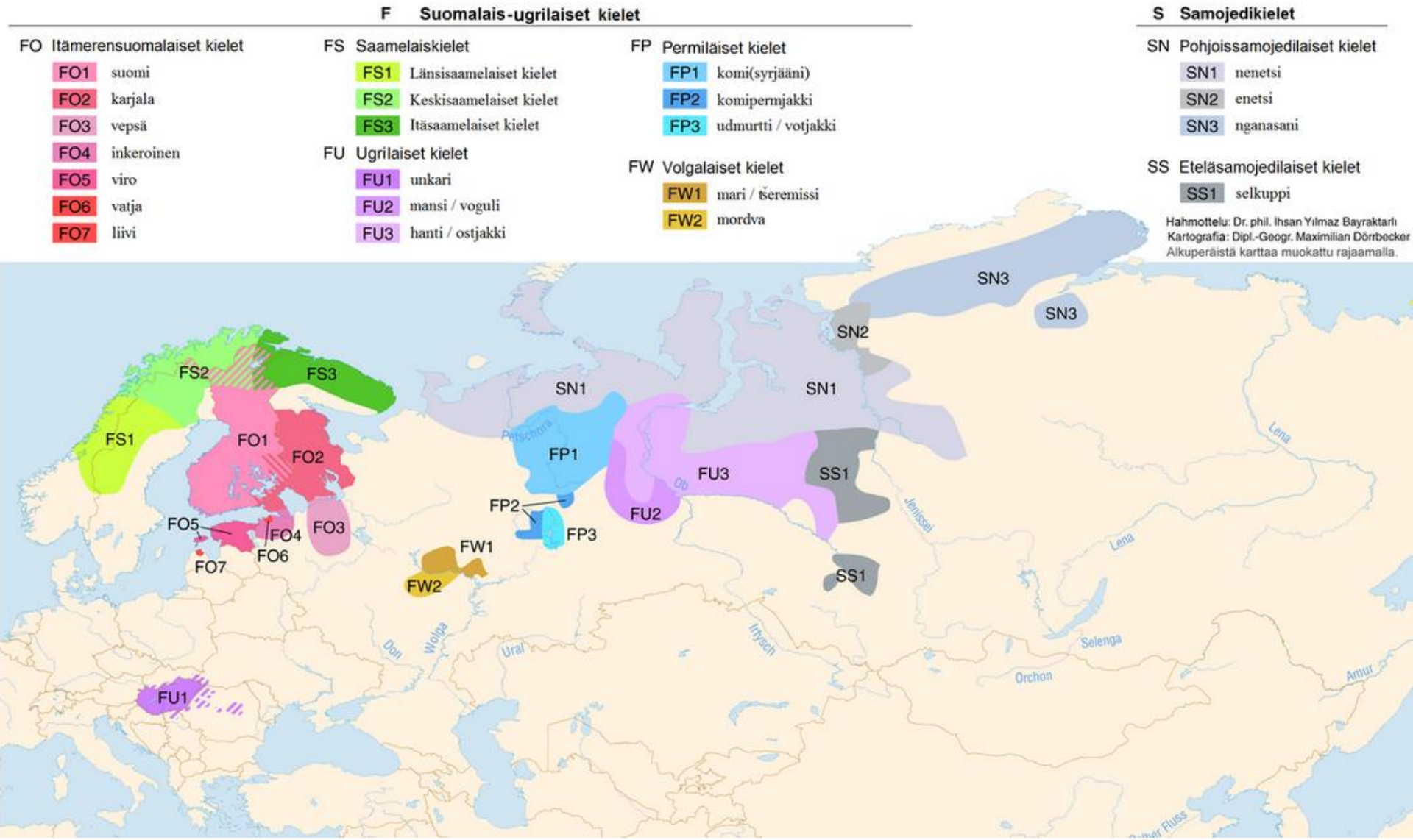
Sami
- Skolt

Samoyedic
- Nenets
- Selkup

Ob-Ugric
- Khanty
- Mansi

Mordvinic
- Erzyan
- Moksha
- (Shoksha)

Languages of Publications



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
  - availablility 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)

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.

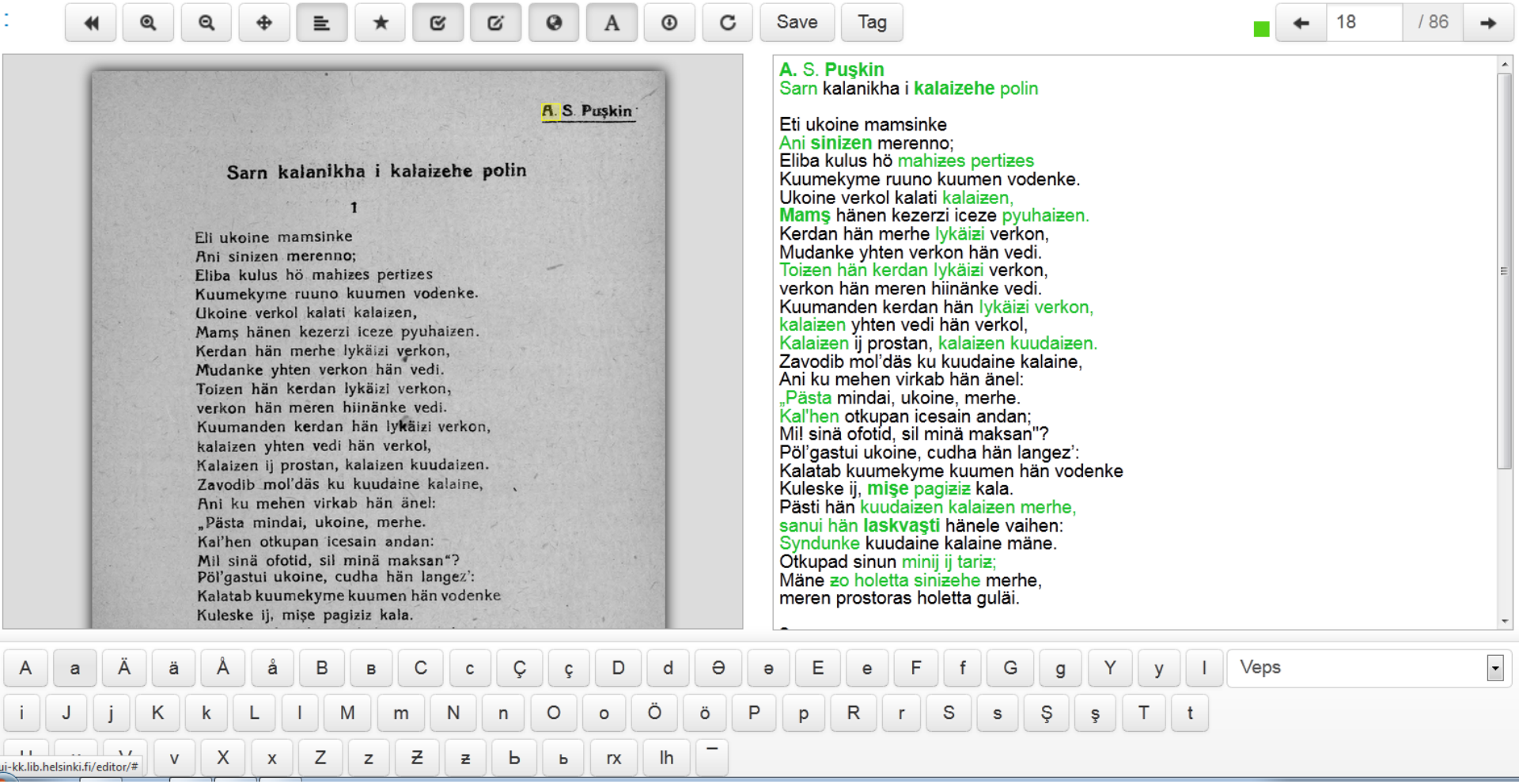
- The machined-encoded text (OCR) contain quite often too many mistakes to be used in research. **The mistakes in OCR'd texts must be corrected.** In order to meet the objective, we have developed an open source code **OCR editor** that enables the editing of erroneous text.

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


Save
Tag
18
/ 86
A. S. Puşkin
Sarn kalanikha i kalaizehe polin
1
Eli ukoine mamsinke
Ani sinizen merenno;
Eliba kulus hö mahizes pertizes
Kuumekyme 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 kuumekyme kuumen hän vodenke
Kuleske ij, mişe pagiziz kala.
A. S. Puşkin
Sarn kalanikha i kalaizehe polin
Eti ukoine mamsinke
Ani sinizen merenno;
Eliba kulus hö mahizes pertizes
Kuumekyme 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;
Mi! sinä ofotid, sil minä maksan"?
Pöl'gastui ukoine, cudha hän langez':
Kalatab kuumekyme 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 в C c Ç ç D d Ə ə E e F f G g Y y 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 Ƶ ƶ ь ь 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.

Citizen Science and 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.

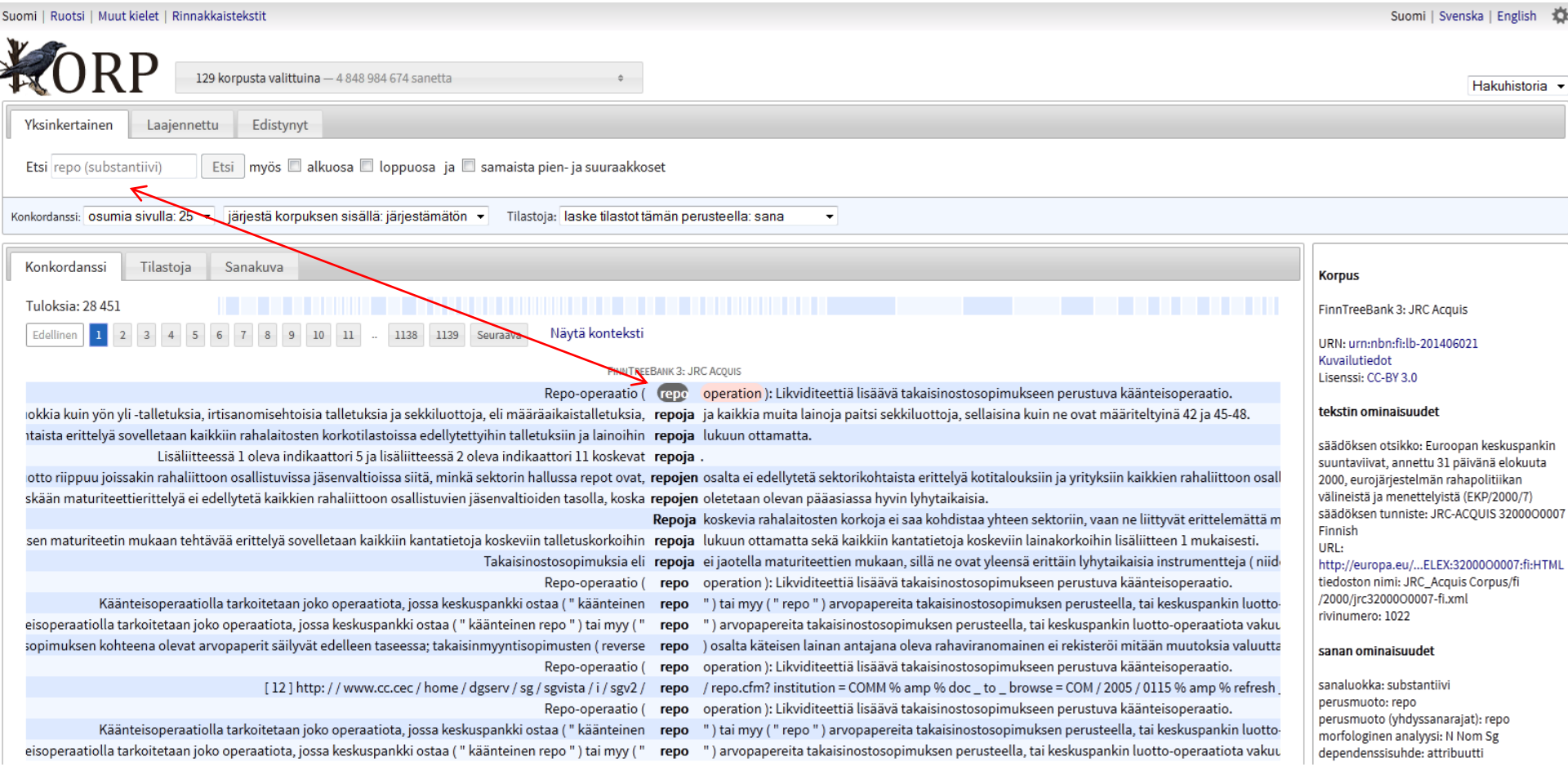
- This objective of interplay can be understood as an aspiration to support the **endangered languages** and the maintenance of **lingual diversity**, but also as a servant of "two masters", the research and the society.

End-Products for End-Users

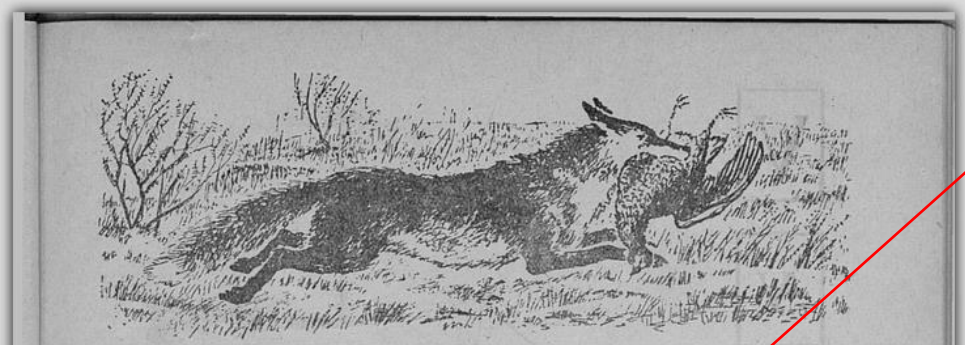
- What to do with the corrected material?
- Library considers the edited material is **raw data** that needs to be released to support the researchers' aspiration, even though the data sets would be incomplete to some extent.
- The distribution of raw data, however, is still a matter of discussion at 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.


Suomi | Ruotsi | Muut kielet | Rinnakkaistekstit
Suomi | Svenska | English
KORP
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
FINNTREEBANK 3: JRC ACQUIS
Repo-operaatio (repo operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.
okkia kuin yön yli -talletuksia, irtisanomisehtoisia talletuksia ja sekkiluottoja, eli määräaikaistalletuksia, repoja ja kaikkia muita lainoja paitsi sekkiluottoja, sellaisina kuin ne ovat määriteltyinä 42 ja 45-48.
taista erittelyä sovelletaan kaikkiin rahalaitosten korkotilastoissa edellytettyihin talletuksiin ja lainoihin repoja lukuun ottamatta.
Lisäliitteessä 1 oleva indikaattori 5 ja lisäliitteessä 2 oleva indikaattori 11 koskevat repoja .
otto riippuu joissakin rahaliittoon osallistuvissa jäsenvaltioissa siitä, minkä sektorin hallussa repot ovat, repojen osalta ei edellytetä sektorikohtaista erittelyä kotitalouksiin ja yrityksiin kaikkien rahaliittoon osall
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
sen maturiteetin mukaan tehtävää erittelyä sovelletaan kaikkiin kantatietoja koskeviin talletuskorkoihin repoja lukuun ottamatta sekä kaikkiin kantatietoja koskeviin lainakorkoihin lisäliitteen 1 mukaisesti.
Takaisinostosopimuksia eli repoja ei jaotella maturiteettien mukaan, sillä ne ovat yleensä erittäin lyhytaikaisia instrumentteja (niid
Repo-operaatio (repo operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.
Käänteisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa (" käänteinen repo ") tai myy (" repo ") arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-
eisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa (" käänteinen repo ") tai myy (" repo ") arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-operaatiota vakuu
sopimuksen kohteena olevat arvopaperit säilyvät edelleen taseessa; takaisinmyyntisopimusten (reverse repo) osalta käteisen lainan antajana oleva rahaviranomainen ei rekisteröi mitään muutoksia valuutta
Repo-operaatio (repo operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.
[12] http: / / www.cc.cec / home / dgserv / sg / sgvista / i / sgv2 / repo / repo.cfm? institution = COMM % amp % doc _ to _ browse = COM / 2005 / 0115 % amp % refresh
Repo-operaatio (repo operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.
Käänteisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa (" käänteinen repo ") tai myy (" repo ") arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-
eisoperaatiolla tarkoitetaan joko operaatiota, jossa keskuspankki ostaa (" käänteinen repo ") tai myy (" repo ") arvopapereita takaisinostosopimuksen perusteella, tai keskuspankin luotto-operaatiota vakuu
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 32000O0007 Finnish
URL: http://europa.eu/...ELEX:32000O0007:fi:HTML
tiedoston nimi: JRC_Acquis Corpus/fi /2000/jrc32000O0007-fi.xml
rivinumero: 1022
sanan ominaisuudet
sanaluokka: substantiivi
perusmuoto: repo
perusmuoto (yhdyssanarajat): repo
morfologinen analyysi: N Nom Sg
dependenssisuhde: attribuutti

Save Tag



repo

re po

r e p o

E

e

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

repo

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

19

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.
- 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.

Some Conclusions

- 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.
- 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**.

Some Conclusions

- **Huge impacts** on society and research are expected.
- Do we really need to know what the impact will be and how valuable that is? **And is that important at all?**

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.

Contact Details

jussi-pekka.hakkarainen@helsinki.fi

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