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

- Part One
  - An overview of the Digitization Project of Kindred Languages.
  - Kone Foundation Language Programme and our role within it.
  - Fenno-Ugrica collection and collection criteria

- Interlude

- Part Two
  - Co-operation with the scholars
  - Tools and methods (nichesourcing) 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** (2012-2016).
- 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.

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

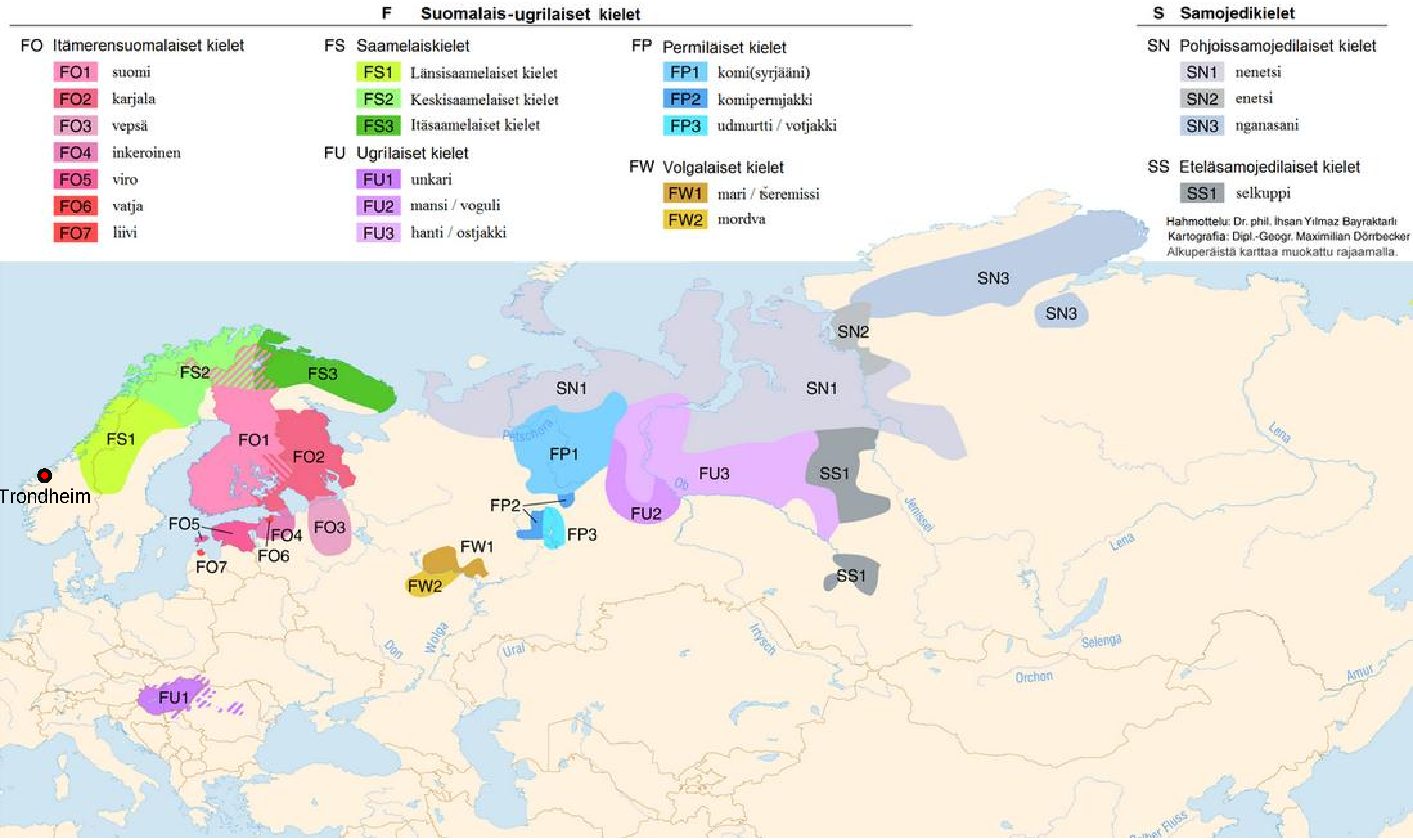
Ob-Ugric

- Khanty
- Mansi

Mordvinic

- Erzyan
- Moksha
- (Shoksha)

Languages of Publications



Selection Criteria of Material

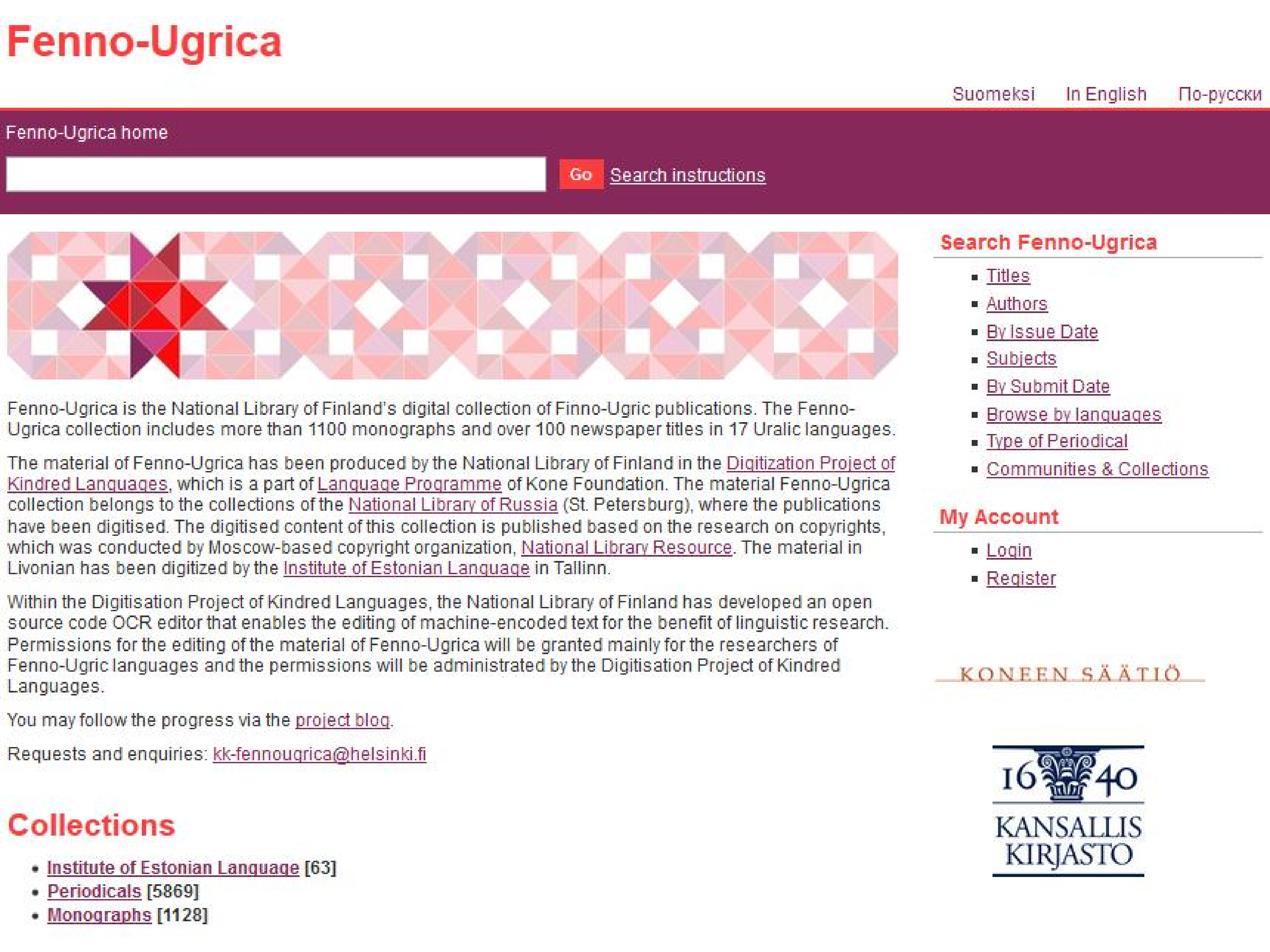
- After the revolution, the Uralic languages were converted into a medium of **popular education**, **enlightenment** and **dissemination** of information pertinent to the developing political agenda of the Soviet state. The deluge of popular literature in 1920s-1930s suddenly challenged **the lexical orthographic norms** of the limited ecclesiastical publications from the 1880s.

- Newspapers were written **in orthographies** and in word forms that the locals would understand. Textbooks were written to address the separate needs of both the adults and children. New concepts were introduced in the language. This was the beginning of a **renaissance** and **period of enlightenment.**

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 and institutions
  - online access to collections in Russia
  - locality – the languages of peripheries is more tempting
  - cost efficiency – loads of parallel titles (translations)
  - **No-one else would digitize and publish this material!**



Fenno-Ugrica is the National Library of Finland's digital collection of Finno-Ugric publications. The Fenno-Ugrica collection includes more than 1100 monographs and over 100 newspaper titles in 17 Uralic languages.

The material of Fenno-Ugrica has been produced by the National Library of Finland in the Digitization Project of Kindred Languages, which is a part of Language Programme of Kone Foundation. The material Fenno-Ugrica collection belongs to the collections of the National Library of Russia (St. Petersburg), where the publications have been digitised. The digitised content of this collection is published based on the research on copyrights, which was conducted by Moscow-based copyright organization, National Library Resource. The material in Livonian has been digitized by the Institute of Estonian Language in Tallinn.

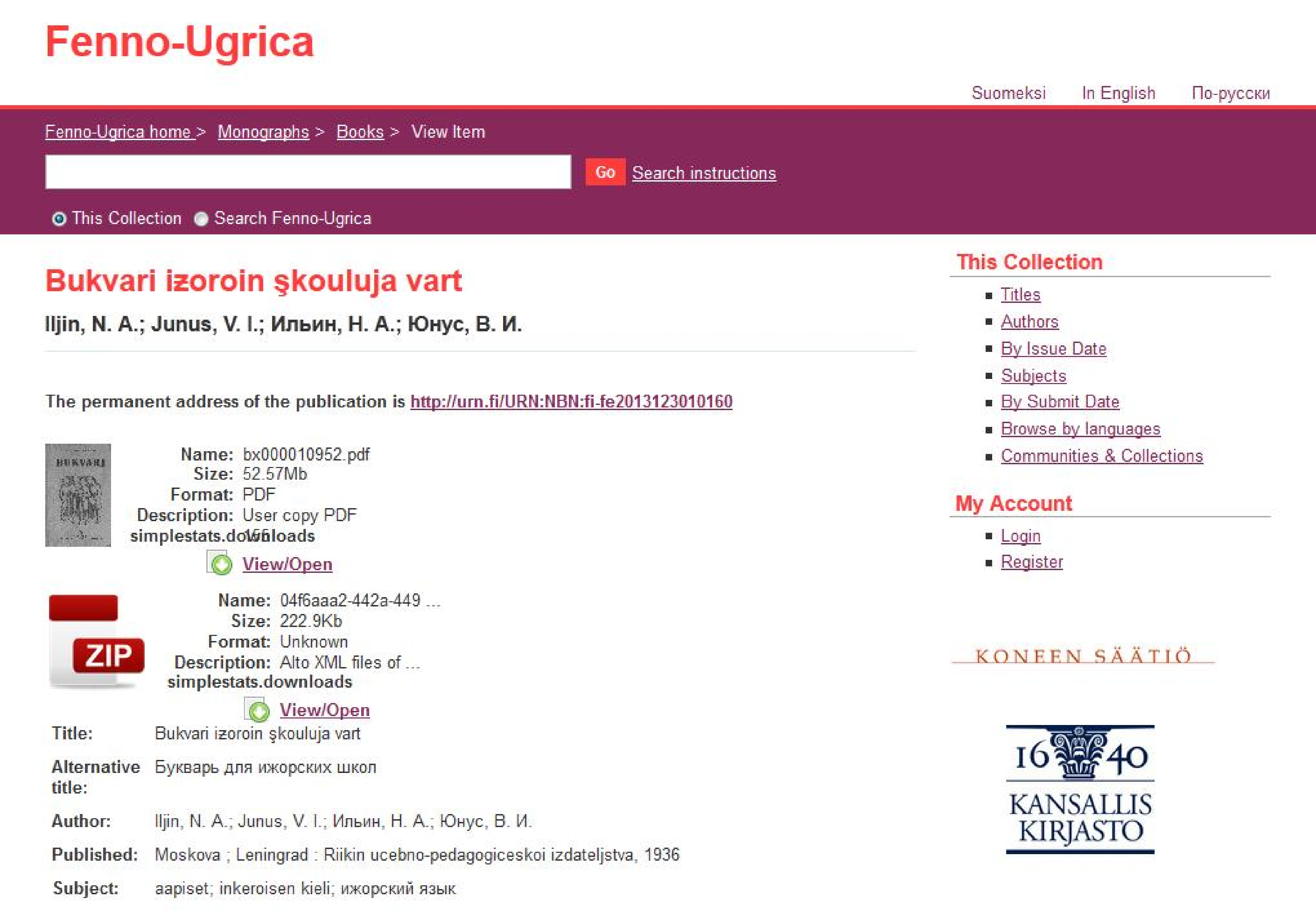
Within the Digitisation Project of Kindred Languages, the National Library of Finland has developed an open source code OCR editor that enables the editing of machine-encoded text for the benefit of linguistic research. Permissions for the editing of the material of Fenno-Ugrica will be granted mainly for the researchers of Fenno-Ugric languages and the permissions will be administrated by the Digitisation Project of Kindred Languages.

You may follow the progress via the project blog.

Requests and enquiries: kk-fennougrica@helsinki.fi

Collections

- Institute of Estonian Language [63]
- Periodicals [5869]
- Monographs [1128]



Fenno-Ugrica

Suomeksi In English По-русски

Fenno-Ugrica home > Monographs > Books > View Item

Go Search instructions

● This Collection ○ Search Fenno-Ugrica

Bukvari izoroin şkouluja 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 155

View/Open

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

View/Open

Title:	Bukvari izoroin 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

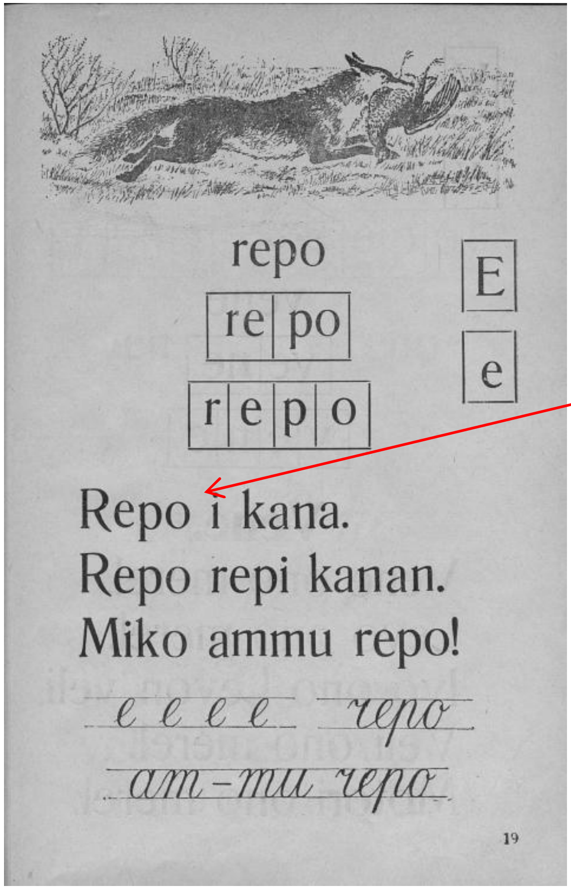
My Account

- Login
- Register

KONEEN SÄÄTIÖ

1640 KANSALLIS KIRJASTO

Materials and Collection



repo

re | po

r | e | p | o

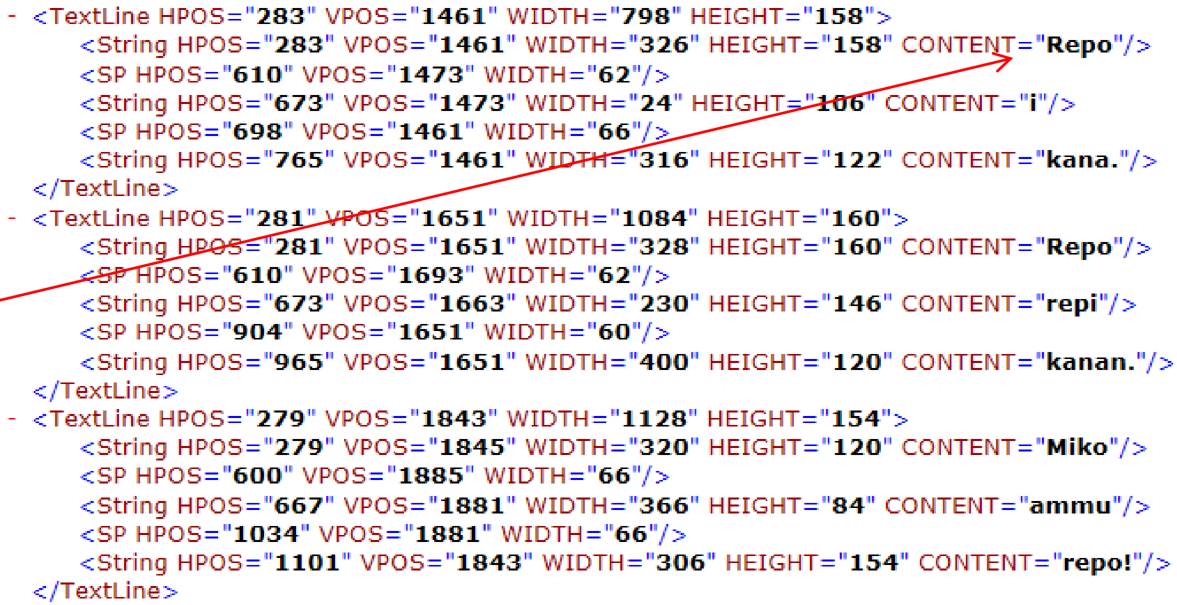
E

e

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

e e e e repo
am-mu repo

19

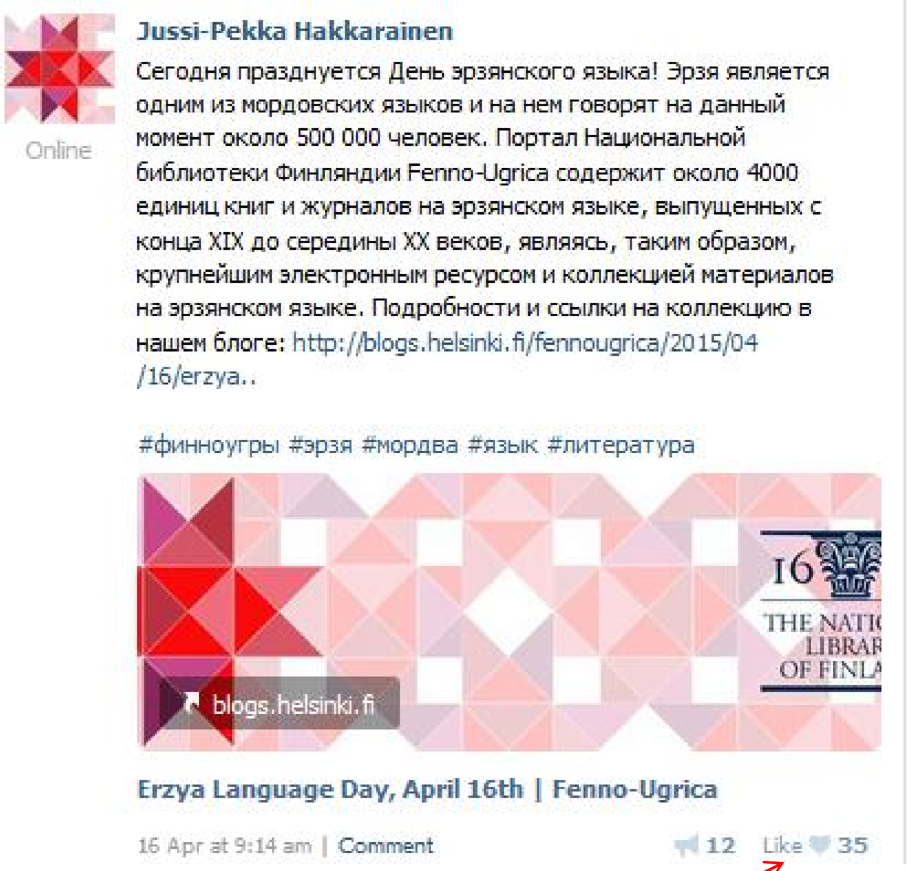


```
- <TextLine HPOS="283" VPOS="1461" WIDTH="798" HEIGHT="158">
      <String HPOS="283" VPOS="1461" WIDTH="326" HEIGHT="158" CONTENT="Repo"/>
      <SP HPOS="610" VPOS="1473" WIDTH="62"/>
      <String HPOS="673" VPOS="1473" WIDTH="24" HEIGHT="106" CONTENT="i"/>
      <SP HPOS="698" VPOS="1461" WIDTH="66"/>
      <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>
```

The Use of Fenno-Ugrica

- When thinking of the possible user groups for Fenno-Ugrica, the most of them are located in Russia.

  - Communication and marketing
  - Accessible user interface
  - Activitity in social media
    - Facebook
    - Twitter
    - Project Blog
    - Vkontakte (in Russian)



Jussi-Pekka Hakkarainen

Online

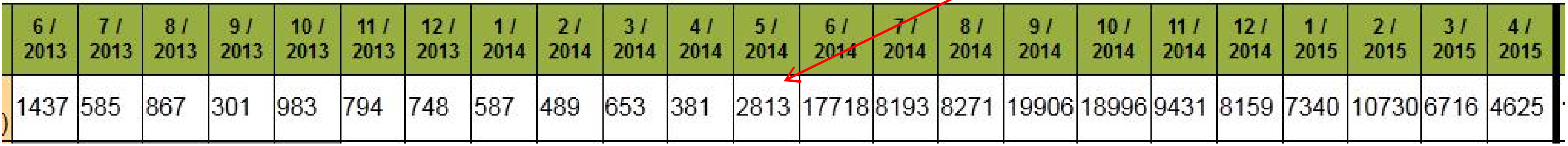
Сегодня празднуется День эрзянского языка! Эрзя является одним из мордовских языков и на нем говорят на данный момент около 500 000 человек. Портал Национальной библиотеки Финляндии Fenno-Ugrica содержит около 4000 единиц книг и журналов на эрзянском языке, выпущенных с конца XIX до середины XX веков, являясь, таким образом, крупнейшим электронным ресурсом и коллекцией материалов на эрзянском языке. Подробности и ссылки на коллекцию в нашем блоге: http://blogs.helsinki.fi/fennougrica/2015/04/16/erzya..

#финноугры #эрзя #мордва #язык #литература

Erzya Language Day, April 16th | Fenno-Ugrica

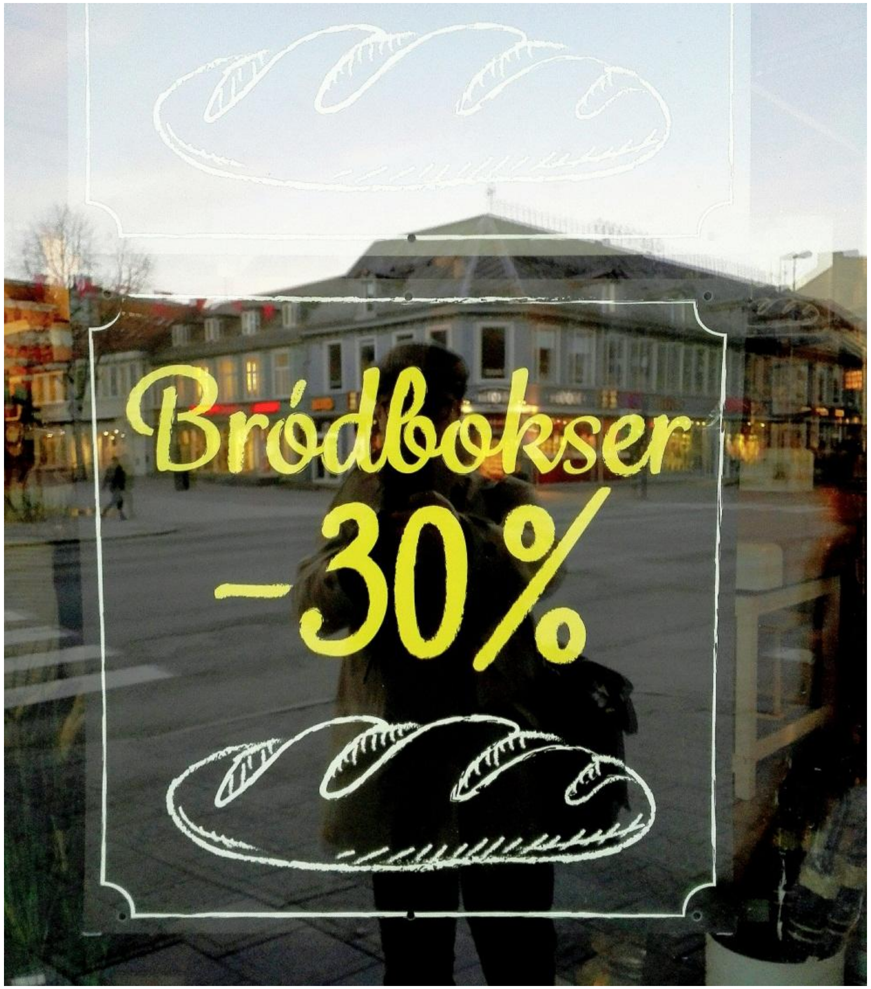
16 Apr at 9:14 am | Comment 12 Like 35

Monthly download statistics



	6 / 2013	7 / 2013	8 / 2013	9 / 2013	10 / 2013	11 / 2013	12 / 2013	1 / 2014	2 / 2014	3 / 2014	4 / 2014	5 / 2014	6 / 2014	7 / 2014	8 / 2014	9 / 2014	10 / 2014	11 / 2014	12 / 2014	1 / 2015	2 / 2015	3 / 2015	4 / 2015
)	1437	585	867	301	983	794	748	587	489	653	381	2813	17718	8193	8271	19906	18996	9431	8159	7340	10730	6716	4625

Interlude

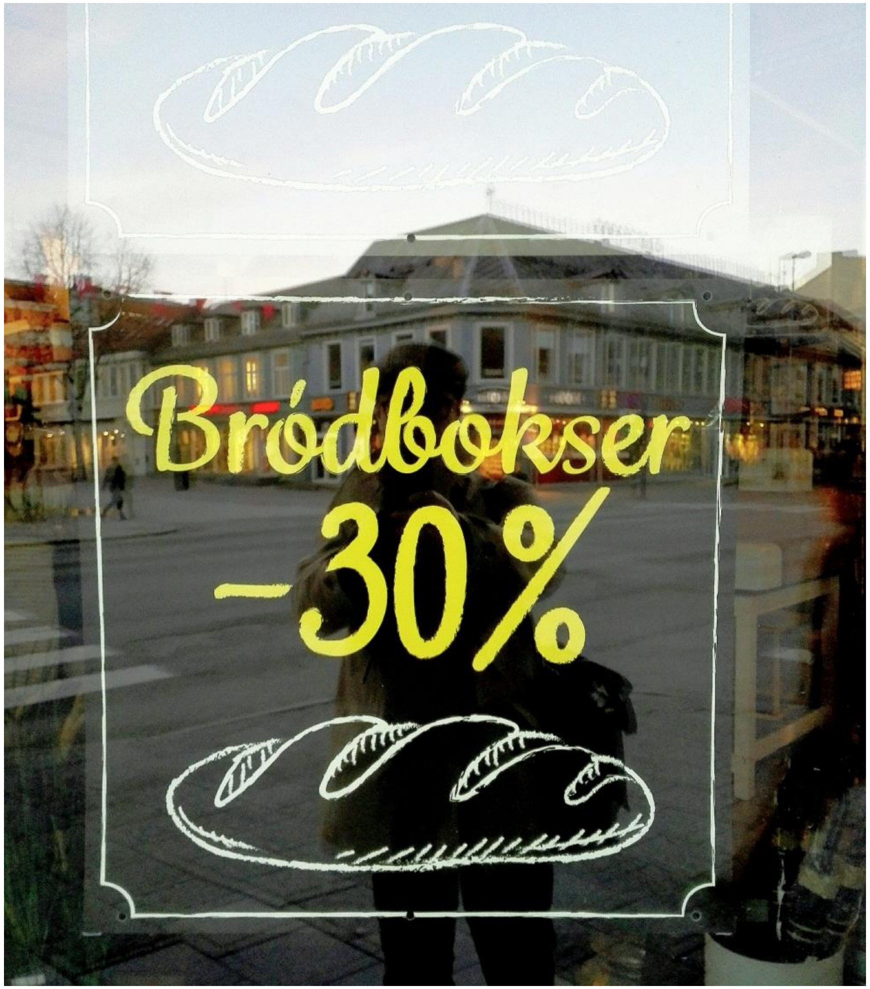


Trondheim,
April 19th, 2015

Bródbokser?
(Bread bins/boxes)

Interlude

- Nynorsk or Bokmål? Old orthography?
- Should it be **brødbokser?**
- What has been the primary intention here?
- What is the **correct** transliteration? With acute or with stroke, or…



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.

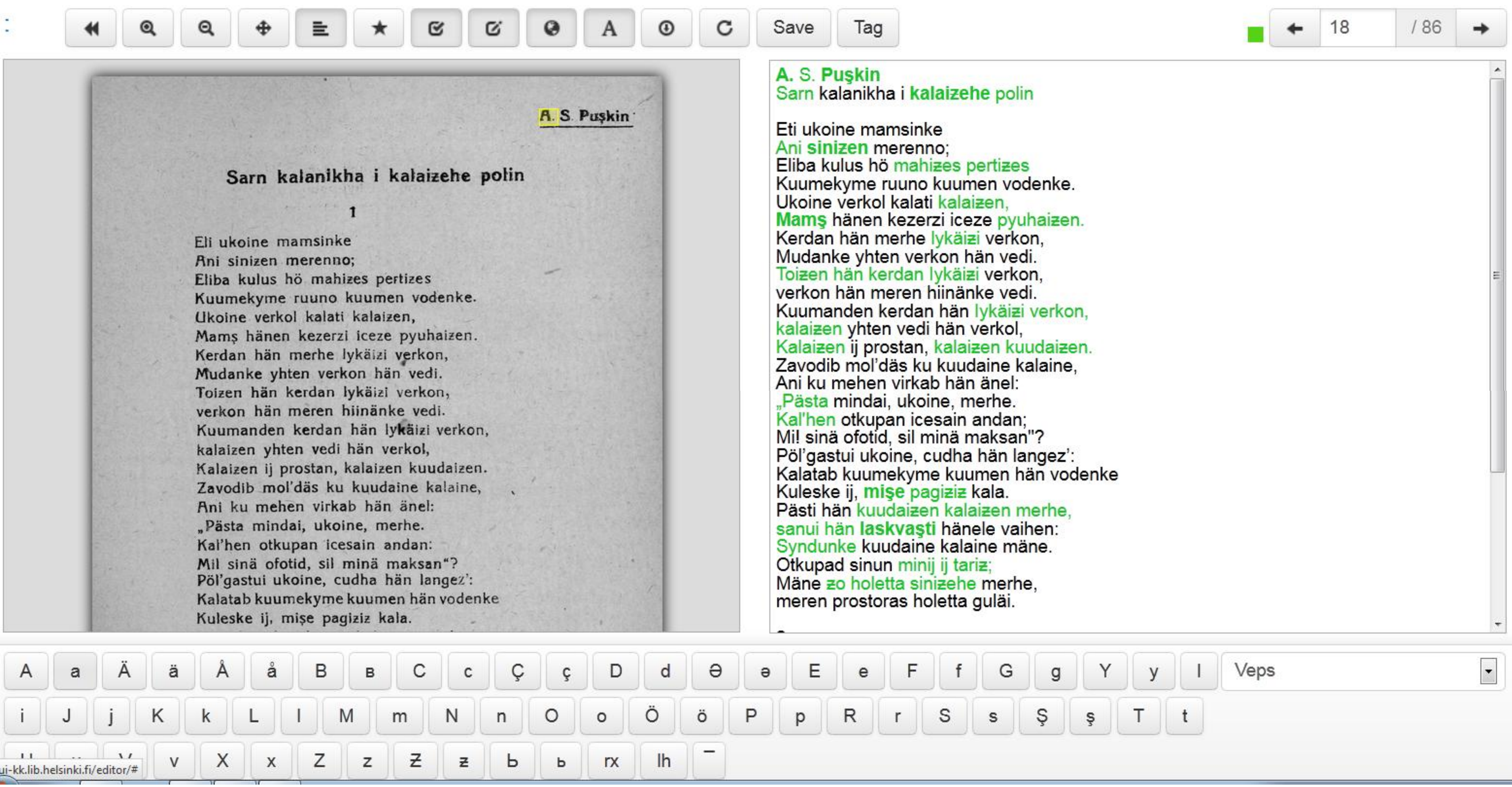
- The machined-encoded text (OCR) contain quite often too many mistakes to be used in research. **The mistakes in OCR'd texts must be corrected.** In order to meet the objective, we have developed an open source code **OCR editor** that enables the editing of erroneous text.

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
Veps
A a Ä ä Å å B ʙ C c Ç ç D d Ə ə E e F f G g Y y I
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 ‾

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.

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**.

Citizen Science and Crowdsourcing

- Traditional approach gives you only

bród

- Whereas nichesourcing gives you

brød

…or potentially more

brød / bröd / bread / bröt / leipä / chleb etc.

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.

- 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.

Nichesourcing and Language Communities

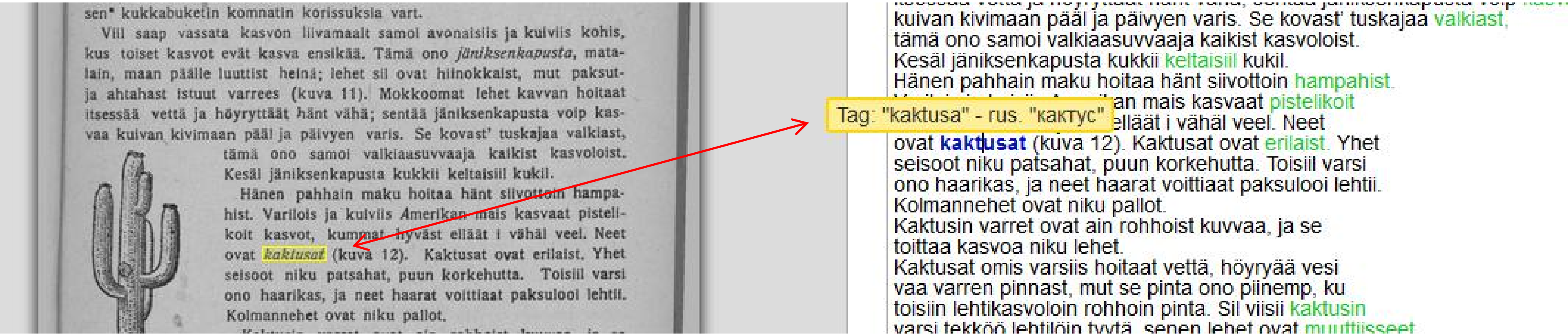
- Ingrian, an endangered language, spoken west from Saint Petersburg, around **300 native-speakers** left.

- No education available in native language, only voluntary lessons on Sundays every fortnight

- The focus group is no longer the old people, but **educated** and **assimiliated** Ingrians. They have enough **sparetime** and **opportunities** to execute the proof-reading.

Nichesourcing and Language Communities

- Skilled and educated people can do a lot
- The corrected words in Ingrian will be added onto the multilingual online dictionary, which is made freely available for the public.



sen* kukkabuketin komnatin korissuksia vart.

Viil saap vassata kasvon liivamaalt samoi avonaisiis ja kuiviis kohis, kus toiset kasvot evät kasva ensikää. Tämä ono *jäniksenkapusta*, matalain, maan päälle luuttist heinä; lehet sil ovat hiinokkaist, mut paksutja ahtahast istuut varrees (kuva 11). Mokkoomat lehet kavvan hoitaat itsessää vettä ja höyryttäät hänt vähä; sentää jäniksenkapusta voip kasvaa kuivan kivimaan pääl ja päivyen varis. Se kovast' tuskajaa valkiast, tämä ono samoi valkiaasuvvaaja kaikist kasvoloist. Kesäl jäniksenkapusta kukkii keltaisiil kukil.

Hänen pahhain maku hoitaa hänt siivottoin hampahist. Varilois ja kuiviis *Amerikan* mais kasvaat pistelikoit kasvot, kummat hyväst elläät i vähäl veel. Neet ovat *kaktusat* (kuva 12). Kaktusat ovat erilaist. Yhet seisoot niku patsahat, puun korkehutta. Toisiil varsi ono haarikas, ja neet haarat voittiaat paksulooi lehtii. Kolmannehet ovat niku pallot.

Tag: "kaktusa" - rus. "кактус"

kuivan kivimaan pääl ja päivyen varis. Se kovast' tuskajaa valkiast,
tämä ono samoi valkiaasuvvaaja kaikist kasvoloist.
Kesäl jäniksenkapusta kukkii keltaisiil kukil.
Hänen pahhain maku hoitaa hänt siivottoin hampahist.
an mais kasvaat pistelikoit
elläät i vähäl veel. Neet
ovat **kaktusat** (kuva 12). Kaktusat ovat erilaist. Yhet
seisoot niku patsahat, puun korkehutta. Toisiil varsi
ono haarikas, ja neet haarat voittiaat paksulooi lehtii.
Kolmannehet ovat niku pallot.
Kaktusin varret ovat ain rohhoist kuvvaa, ja se
toittaa kasvoa niku lehet.
Kaktusat omis varsiis hoitaat vettä, höyryää vesi
vaa varren pinnast, mut se pinta ono piinemp, ku
toisiin lehtikasvoloin rohhoin pinta. Sil viisii kaktusin
varsi tekköö lehtilöin tyytä, senen lehet ovat muuttiisseet

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' aspirations, 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.

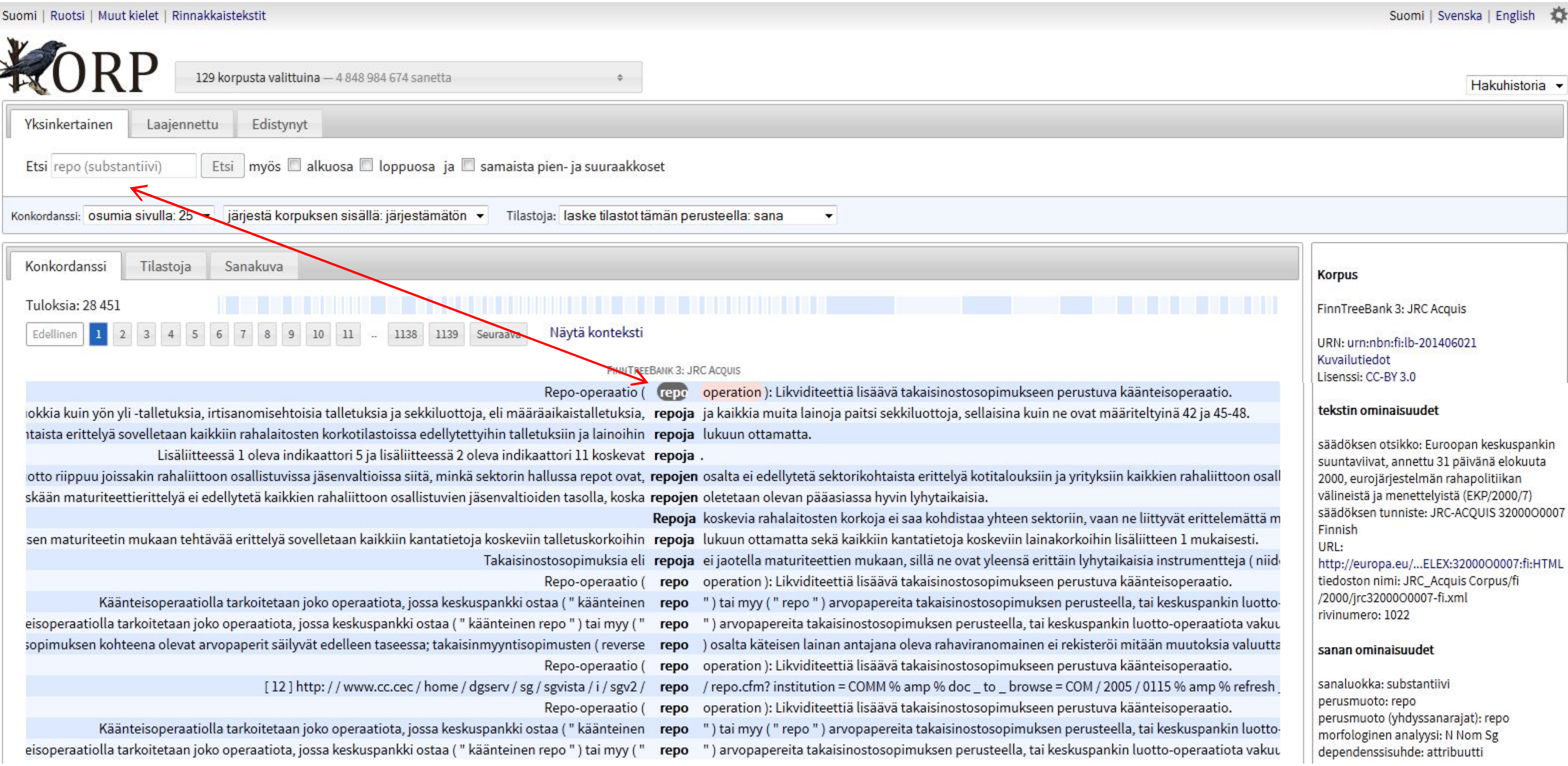
- The raw data hub, **data.kansalliskirjasto.fi** could be a solution.

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 Finnish (Swedish) Language Bank.



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

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

Tuloksia: 28 451

FINNTREEBANK 3: JRC ACQUIS

Repo-operaatio (	repo	operation): Likviditeettiä lisäävä takaisinostosopimukseen perustuva käänteisoperaatio.
ıokkia 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.
ıtaista 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**	.
ıotto 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 vakuı
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 vakuı

Korpus

FinnTreeBank 3: JRC Acquis

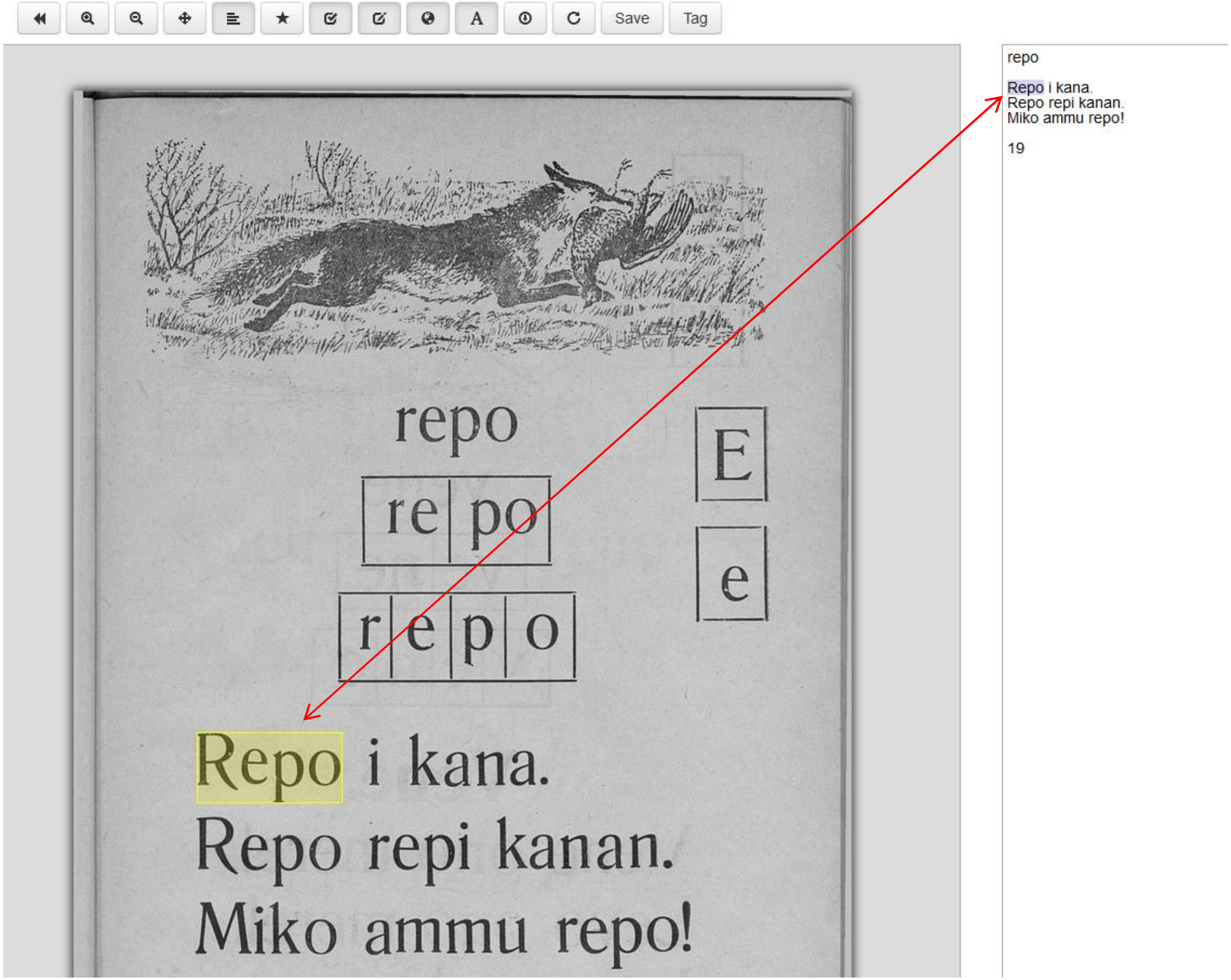
URN: urn:nbn:fi:lb-201406021
Kuvailutiedot
Lisenssi: CC-BY 3.0

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



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.

Thanks for Your Patience!

Contact Details

jussi-pekka.hakkarainen@helsinki.fi

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