Digital Heritage Serving Two Masters: the Great Public and the Academia

Jussi-Pekka Hakkarainen
Project Manager

Seminar on Fenno-Ugric Computational Linguistics
23.9.2016, Helsinki

Overview of the Project

- The **National Library of Finland** has been implementing the **Minority Languages Project** since 2012.

- Within the project we have digitized materials in 18 Uralic languages as well as developed tools to support the **1) linguistic research** and **2) citizen science**.

- Through this project, **1) researchers** would gain access to new corpora which they have not been able to study before and to which **2) all users** would have open access regardless of their place of residence.

Materials and Collection

- Within the project, **National Library of Finland** has digitized and published around **1150 monograph** titles and more than **100 newspapers** titles in 18 Uralic languages.

- The online collection, **Fenno-Ugrica**, will consist of 110,000 monograph pages, 150,000 newspaper and journal pages as well as some manuscripts. All in all, ca. **30 000 items** were published.

- The majority of materials belong to the collections of the **National Library of Russia** in Saint Petersburg.

The Content

- Aim to digitize items which **would precisely fill the gaps** in linguistic research and in corpora, vocabularies etc.

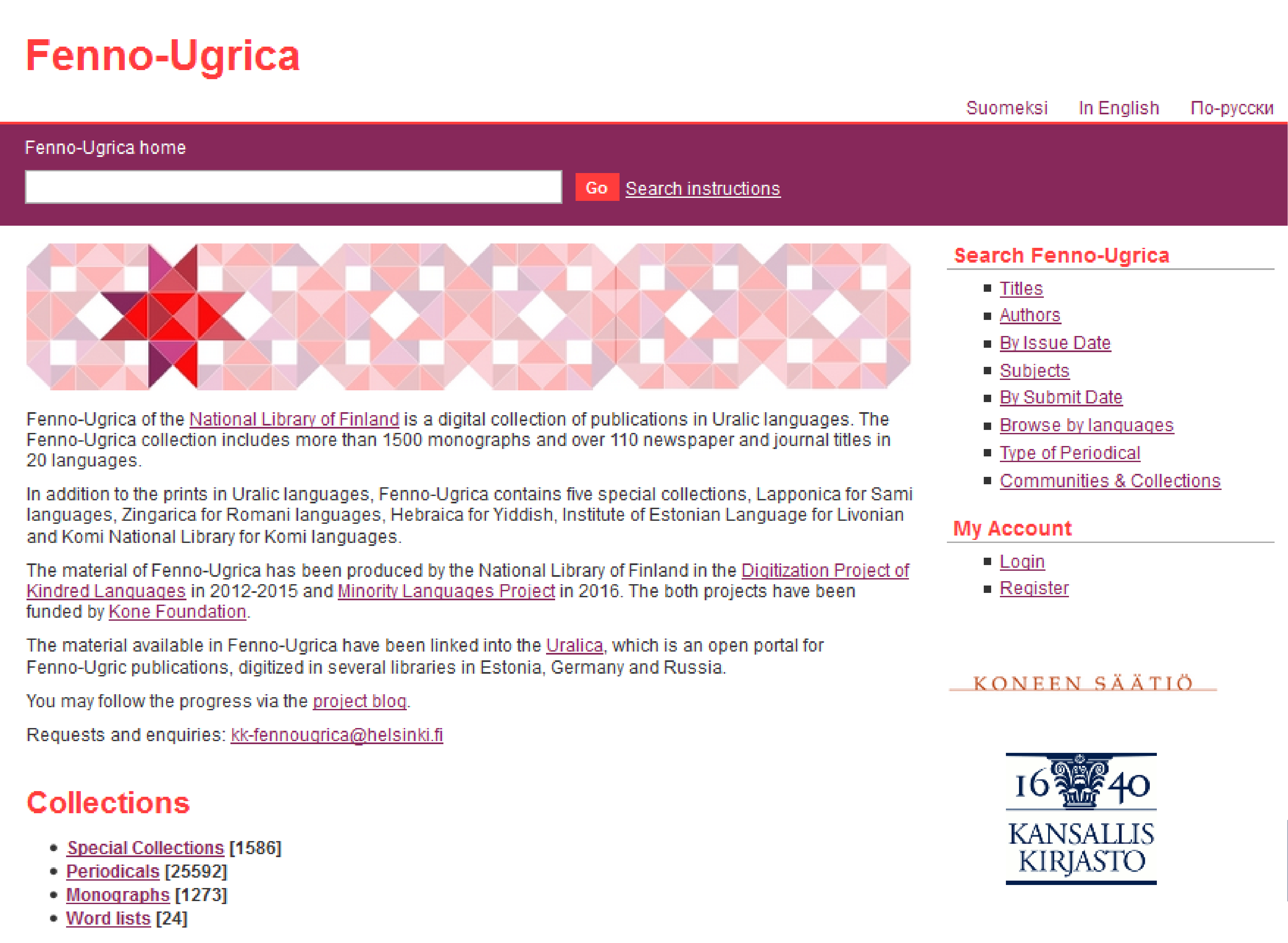
- After 1917, the languages were converted into a medium of **popular education**, **enlightenment** and **dissemination** of information pertinent to the developing political agenda of the Soviet state.

- The deluge of literature in 1920s-1930s suddenly challenged **the lexical orthographic norms** of the limited ecclesiastical publications from the 1880s. 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 are more tempting
  - cost efficiency – loads of parallel titles (translations)
  - **No-one else would digitize and publish this material!**



Fenno-Ugrica

Fenno-Ugrica of the National Library of Finland is a digital collection of publications in Uralic languages. The Fenno-Ugrica collection includes more than 1500 monographs and over 110 newspaper and journal titles in 20 languages.

In addition to the prints in Uralic languages, Fenno-Ugrica contains five special collections, Lapponica for Sami languages, Zingarica for Romani languages, Hebraica for Yiddish, Institute of Estonian Language for Livonian and Komi National Library for Komi languages.

The material of Fenno-Ugrica has been produced by the National Library of Finland in the Digitization Project of Kindred Languages in 2012-2015 and Minority Languages Project in 2016. The both projects have been funded by Kone Foundation.

The material available in Fenno-Ugrica have been linked into the Uralica, which is an open portal for Fenno-Ugric publications, digitized in several libraries in Estonia, Germany and Russia.

You may follow the progress via the project blog.

Requests and enquiries: kk-fennougrica@helsinki.fi

Collections

- Special Collections [1586]
- Periodicals [25592]
- Monographs [1273]
- Word lists [24]

Search Fenno-Ugrica

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

My Account

- Login
- Register

KONEEN SÄÄTIÖ

1640
KANSALLIS
KIRJASTO

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	156

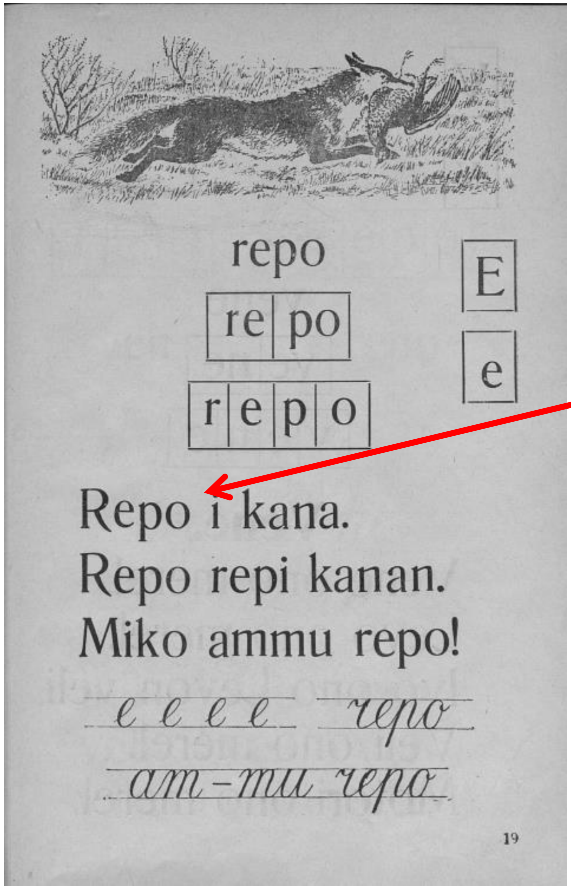
View/Open

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

Materials and Collection



repo

re | po

r | e | p | o

E

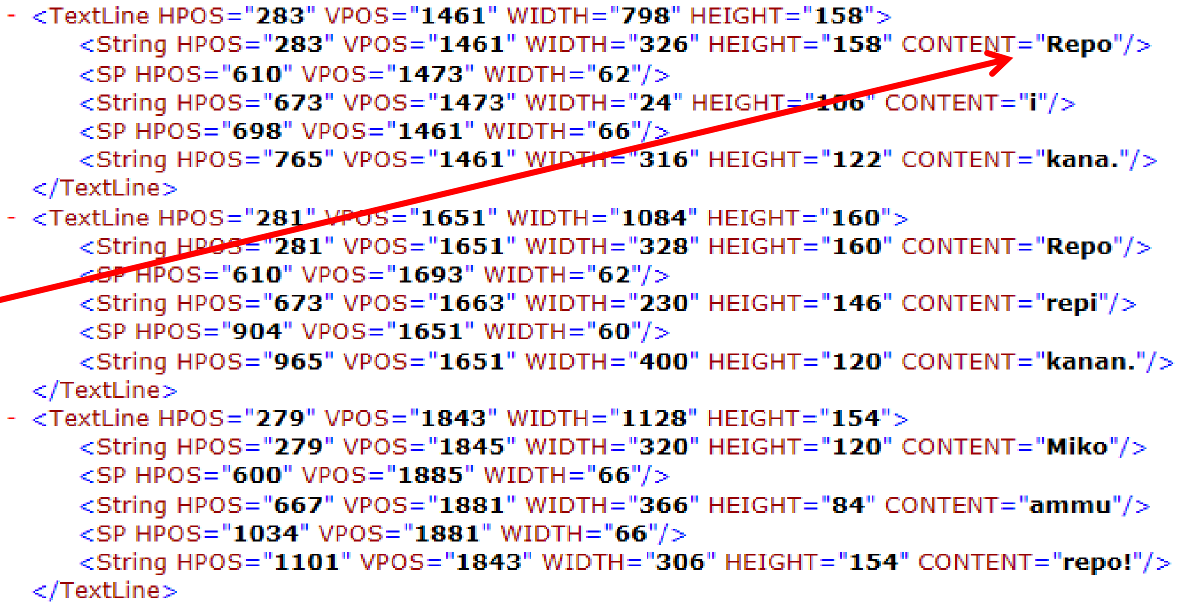
e

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>
```

Languages of Publications

Mari

- Meadow Mari
- Hill Mari

Permic

- Udmurt
- Komi-Zyrian
- Komi-Permyak

Ob-Ugric

- Khanty
- Mansi

Samoyedic

- Nenets
- Selkup

Mordvinic

- Erzyan
- Moksha
- (Shoksha)

Baltic Finns

- Ingrian
- Veps
- Karelian
- [Livonian]

Sami

- Kildin
- Skolt
- Northern

Engaging the Academia

- Key factor for success was the engaging the researchers to the project. The researchers were in the loop from the very beginning, even before the money was raised.

- Researchers did participate in
  - **Planning**
  - **Decision-making**
  - **Communications**
  - **Networking**
  - **Distribution of data**

Researchers and Decision-making

- Once the money was granted, we called researchers to the steering group, so they were part of decision-making.

- In addition to that, the meetings with the researchers in the field took place, both formally and informally:
  - Coffee sessions with the Helsinki-based researchers on the weekly basis
  - Annual meetings to discuss about the needs within the Academia
  - Participating the seminars, congresses, workshops etc.
  - **Forcing** them to communicate on our behalf.

Researchers and Networking

- The participation of researchers and the co-operation with the researchers took off once we got the reputation.
  - **2 projects in 2012**
  - **7 projects and 2 institutions in 2013**
  - **4 PhD students, 11 projects, 4 NGOs and 8 institutions in 2016.**

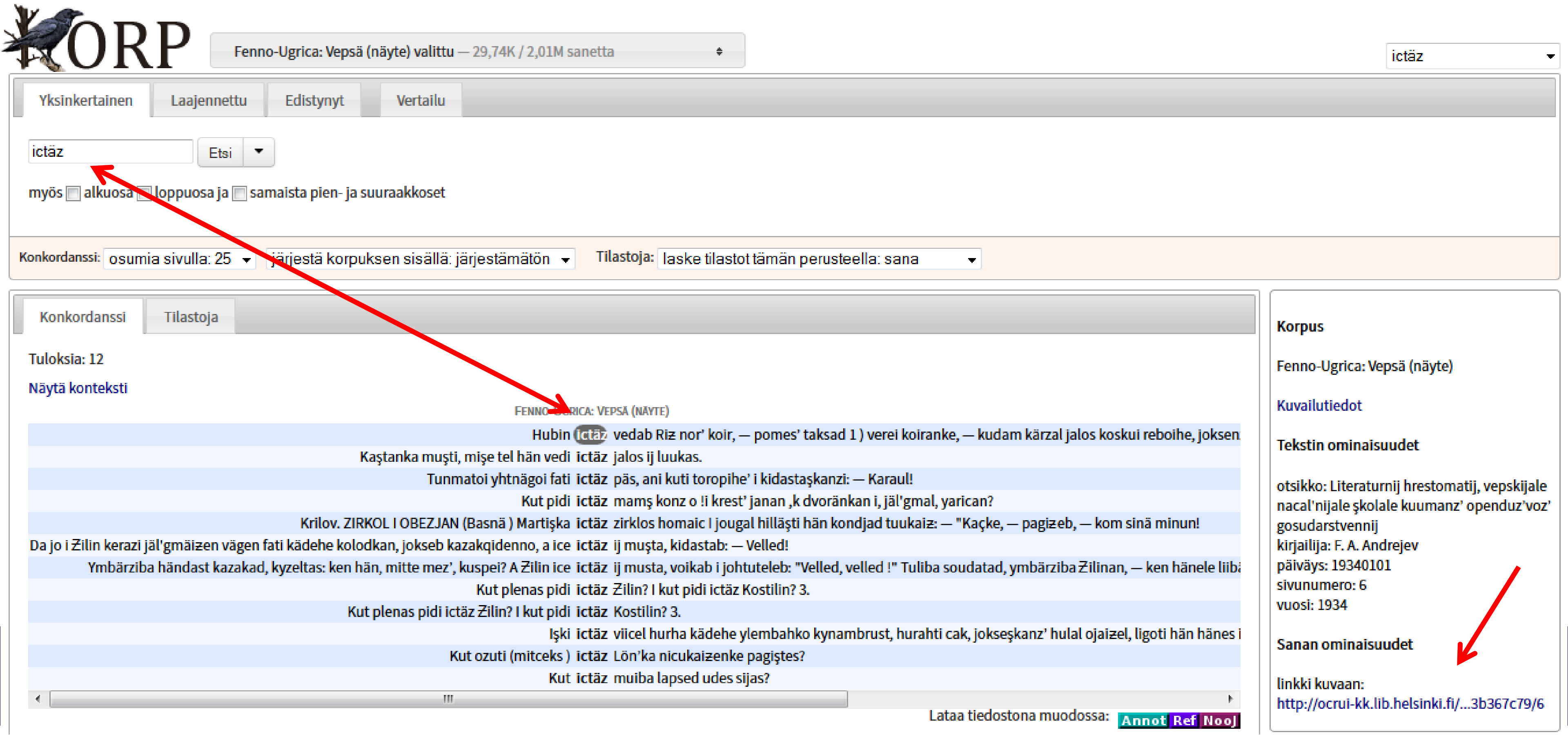
Datasets and Further Use

- We have created the data ourselves and released the data for other operators in **Fenno-Ugrica** as wordlists > raw data for linguistic solutions like online dictionaries, spell-checkers, corpora etc.

- Edited material is available also in **Korp,** which is the concordance search tool of the Finnish Language Bank.

- Some parts of available also at **Giellatekno** of Arctic University of Norway and **Institut für Deutsche Sprache** and **FU-Lab** in Syktyvkar, Komi Republic (Russia)

Datasets and Further Use



ictäz

myös □ alkuosa □ loppuosa ja □ samaista pien- ja suuraakkoset

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

Tuloksia: 12

Näytä konteksti

FENNO-UGRICA: VEPSÄ (NÄYTE)

Hubin	ictäz	vedab Riƶ nor' koir, — pomes' taksad 1) verei koiranke, — kudam kärzal jalos koskui reboihe, joksen
Kaştanka muşti, mişe tel hän vedi	ictäz	jalos ij luukas.
Tunmatoi yhtnägoi fati	ictäz	päs, ani kuti toropihe' i kidastaşkanzi: — Karaul!
Kut pidi	ictäz	mamş konz o !i krest' janan ,k dvoränkan i, jäl'gmal, yarican?
Krilov. ZIRKOL I OBEZJAN (Basnä) Martişka	ictäz	zirklos homaic I jougal hilläşti hän kondjad tuukaiƶ: — "Kaçke, — pagiƶeb, — kom sinä minun!
Da jo i Ƶilin kerazi jäl'gmäiƶen vägen fati kädehe kolodkan, jokseb kazakqidenno, a ice	ictäz	ij muşta, kidastab: — Velled!
Ymbärziba händast kazakad, kyzeltas: ken hän, mitte mez', kuspei? A Ƶilin ice	ictäz	ij musta, voikab i johtuteleb: "Velled, velled !" Tuliba soudatad, ymbärziba Ƶilinan, — ken hänele liib
Kut plenas pidi	ictäz	Ƶilin? I kut pidi ictäz Kostilin? 3.
Kut plenas pidi ictäz Ƶilin? I kut pidi	ictäz	Kostilin? 3.
Işki	ictäz	viicel hurha kädehe ylembahko kynambrust, hurahti cak, jokseşkanz' hulal ojaiƶel, ligoti hän hänes i
Kut ozuti (mitceks)	ictäz	Lön'ka nicukaiƶenke pagiştes?
Kut	ictäz	muiba lapsed udes sijas?

Lataa tiedostona muodossa: Annot Ref NooJ

Korpus

Fenno-Ugrica: Vepsä (näyte)

Kuvailutiedot

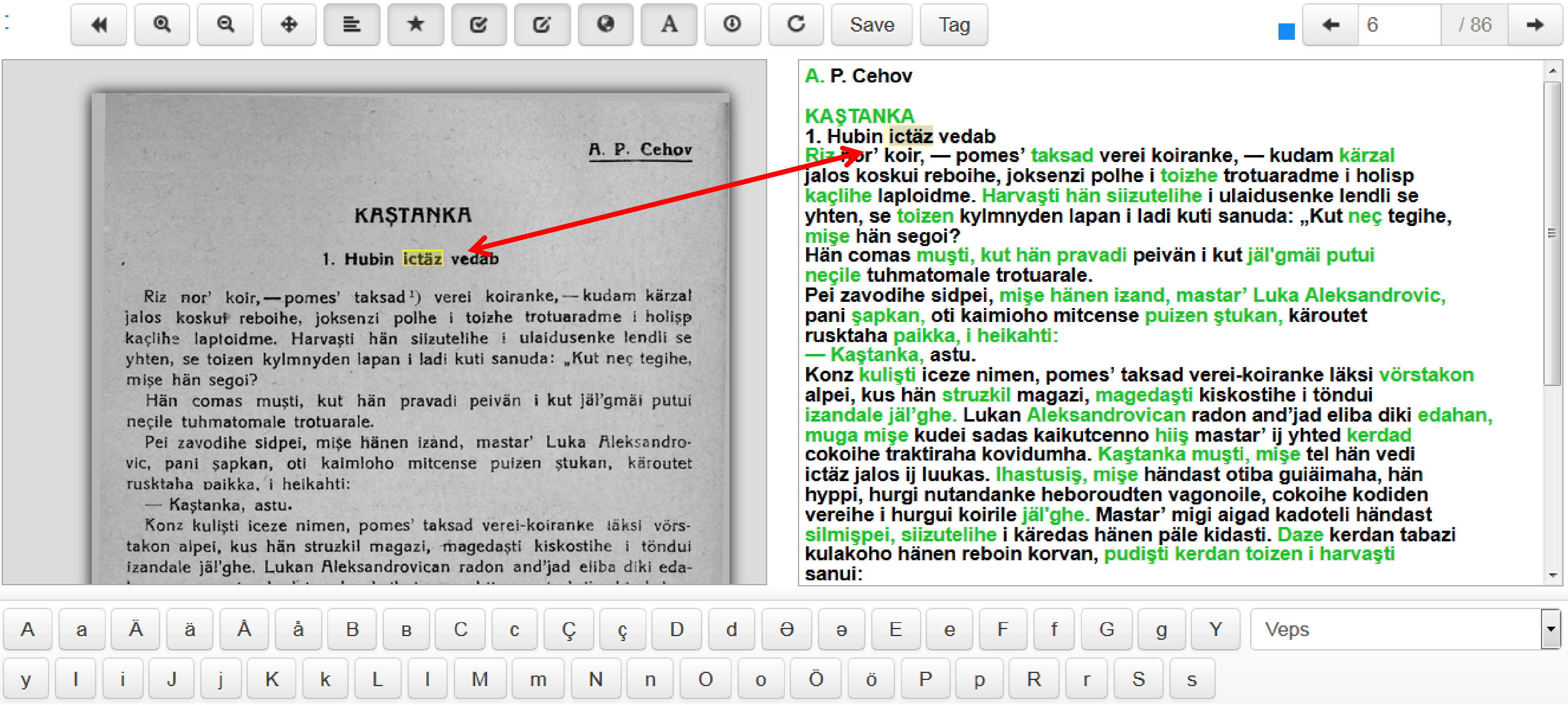
Tekstin ominaisuudet

otsikko: Literaturnij hrestomatij, vepskijale nacal'nijale şkolale kuumanz' openduz'voz' gosudarstvennij
kirjailija: F. A. Andrejev
päiväys: 19340101
sivunumero: 6
vuosi: 1934

Sanan ominaisuudet

linkki kuvaan:
http://ocrui-kk.lib.helsinki.fi/...3b367c79/6

Datasets and Further Use



A. P. Cehov

KAŞTANKA

1. Hubin ictäz vedab

Riz nor’ koir, — pomes’ taksad[1]) verei koiranke, — kudam kärzal jalos koskui reboihe, joksenzi polhe i toizhe trotuaradme i holişp kaçlihe laploidme. Harvaşti hän siizutelihe i ulaidusenke lendli se yhten, se toizen kylmnyden lapan i ladi kuti sanuda: „Kut neç tegihe, mişe hän segoi?

Hän comas muşti, kut hän pravadi peivän i kut jäl’gmäi putui neçile tuhmatomale trotuarale.

Pei zavodihe sidpei, mişe hänen izand, mastar’ Luka Aleksandrovic, pani şapkan, oti kaimloho mitcense puizen ştukan, käroutet rusktaha paikka, i heikahti:

— Kaştanka, astu.

Konz kulişti iceze nimen, pomes’ taksad verei-koiranke läksi vörstakon alpei, kus hän struzkil magazi, magedaşti kiskostihe i töndui izandale jäl’ghe. Lukan Aleksandrovican radon and’jad eliba diki eda-

A. P. Cehov

KAŞTANKA
1. Hubin ictäz vedab
Riz nor’ koir, — pomes’ taksad verei koiranke, — kudam kärzal
jalos koskui reboihe, joksenzi polhe i toizhe trotuaradme i holisp
kaçlihe laploidme. Harvaşti hän siizutelihe i ulaidusenke lendli se
yhten, se toizen kylmnyden lapan i ladi kuti sanuda: „Kut neç tegihe,
mişe hän segoi?
Hän comas muşti, kut hän pravadi peivän i kut jäl'gmäi putui
neçile tuhmatomale trotuarale.
Pei zavodihe sidpei, mişe hänen izand, mastar’ Luka Aleksandrovic,
pani şapkan, oti kaimioho mitcense puizen ştukan, käroutet
rusktaha paikka, i heikahti:
— Kaştanka, astu.
Konz kulişti iceze nimen, pomes’ taksad verei-koiranke läksi vörstakon
alpei, kus hän struzkil magazi, magedaşti kiskostihe i töndui
izandale jäl’ghe. Lukan Aleksandrovican radon and’jad eliba diki edahan,
muga mişe kudei sadas kaikutcenno hiiş mastar’ ij yhted kerdad
cokoihe traktiraha kovidumha. Kaştanka muşti, mişe tel hän vedi
ictäz jalos ij luukas. Ihastusiş, mişe händast otiba guiäimaha, hän
hyppi, hurgi nutandanke heboroudten vagonoile, cokoihe kodiden
vereihe i hurgui koirile jäl'ghe. Mastar’ migi aigad kadoteli händast
silmişpei, siizutelihe i käredas hänen päle kidasti. Daze kerdan tabazi
kulakoho hänen reboin korvan, pudişti kerdan toizen i harvaşti
sanui:

Finding the Language Communities

- How to locate people who could use the content for their benefit?

- Co-operation with universities and libraries didn’t really work out

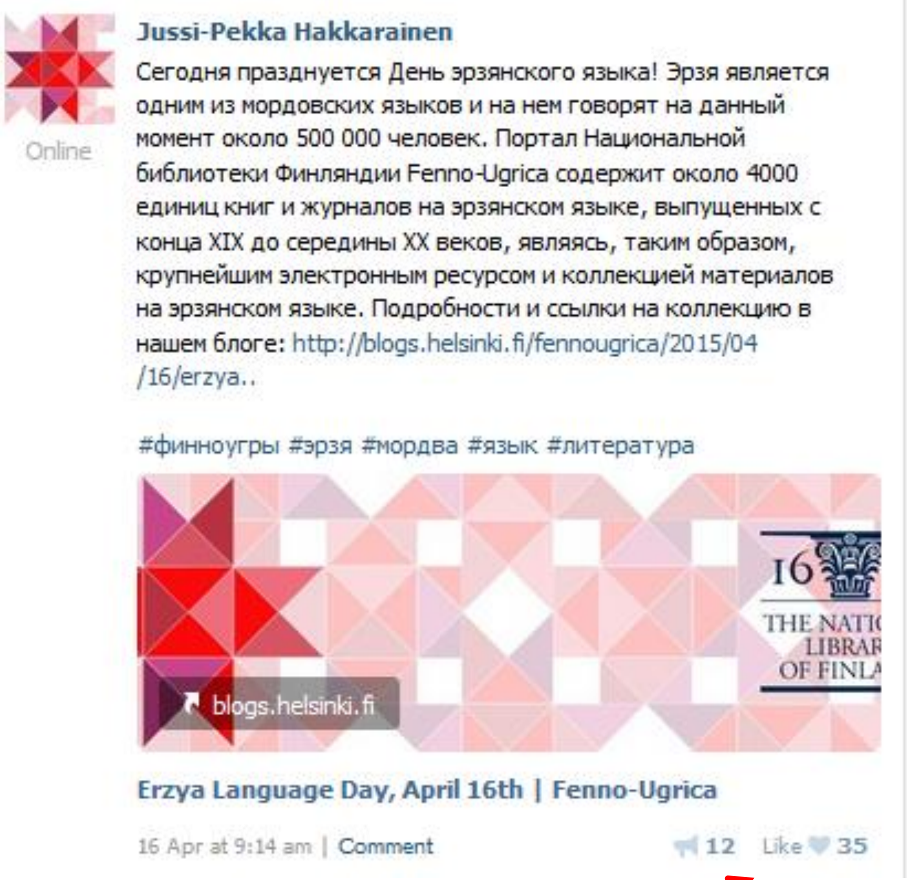
- Activity in English-oriented social media did not help us
  - No remarkable networking, contact or results via WWW, Twitter, Facebook or Project Blog
  - No interactivity with native-speakers

Language Communities, ie. The Great Public

- When thinking of the possible **wider audiences**, one must bear in mind that the most of the people, who speak these languages are located in Russia.

  - Communication and marketing
    - Schedule for blog posts and Vkontakte messages
  - Accessible user interface for Russian-speaking audience
    - Fenno-Ugrica, Uralica
  - Activitity in social media in Russian is necessary
    - Vkontakte
    - Chat forums (for linguists etc)
    - IRC channels

The Vkontakte Effect



Jussi-Pekka Hakkarainen

Online

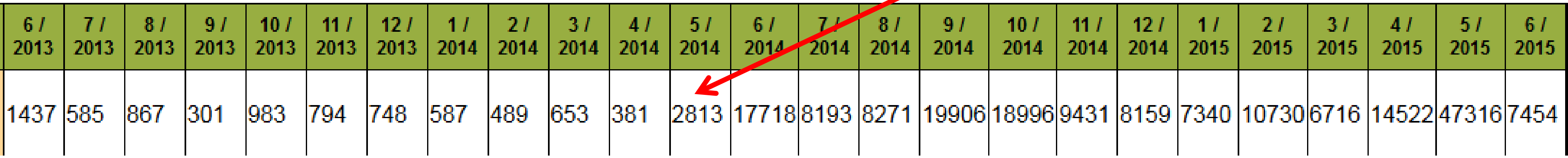
Сегодня празднуется День эрзянского языка! Эрзя является одним из мордовских языков и на нем говорят на данный момент около 500 000 человек. Портал Национальной библиотеки Финляндии Fenno-Ugrica содержит около 4000 единиц книг и журналов на эрзянском языке, выпущенных с конца XIX до середины XX веков, 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	5 / 2015	6 / 2015
1437	585	867	301	983	794	748	587	489	653	381	2813	17718	8193	8271	19906	18996	9431	8159	7340	10730	6716	14522	47316	7454

Some Conclusions

- Involvement of the researchers did build up our reputation, but only among the academia.

- The great audience was reached with the help of social media. One cannot underestimate its power. We benefitted on this in several ways.
  - More active and returning users, more audience and more opportunities to find suitable user-groups etc.
  - Increased co-operation with Russian institutions (when speaking Russian)
  - Makes our work open and understandable for the great audience.

Contact Details

jussi-pekka.hakkarainen@helsinki.fi
@hakkarainen

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