Nichesourcing for the benefit of linguistic research and native speakers

Jussi-Pekka Hakkarainen
Project Manager

CIFU XIII
19.8.2015, Oulu

Overview of the Project

- The **National Library of Finland** is implementing the Digitization Project of Kindred (**Uralic**) Languages in 2012–15.

- Within the project we have digitized materials in 17 Uralic languages as well as developed tools to support the **1) linguistic research** and **2) citizen science**.

- Through this project, **1) researchers** will gain access to new corpora which they have not been able to study before and to which **2) all users** will 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.
- The online collection, **Fenno-Ugrica**, will consist of 110,000 monograph pages and 90,000 newspaper pages.
- The majority of materials belong to the collections of the **National Library of Russia** in Saint Petersburg.

Fenno-Ugrica

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]

Search Fenno-Ugrica

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

My Account

- Login
- Register

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)

Selection Criteria of Material

- 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.

- 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 are more tempting
  - cost efficiency – loads of parallel titles (translations)
  - **No-one else would digitize and publish this material!**

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.

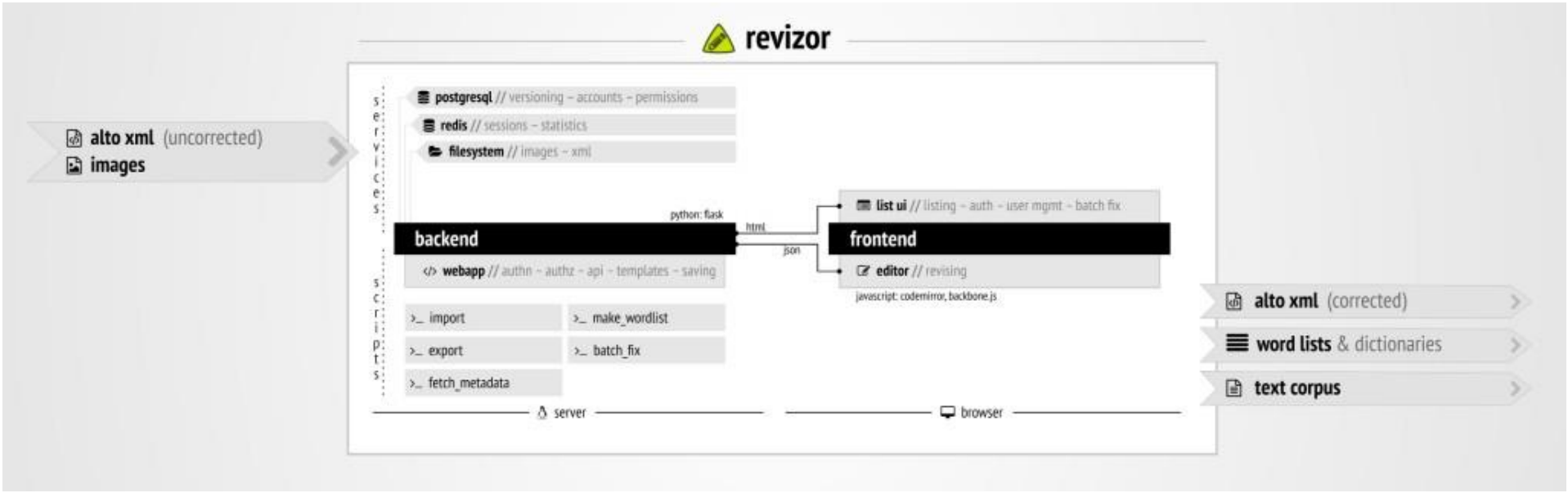
- The machined-encoded text (OCR) contain quite often too many mistakes to be used in research. **The mistakes in OCR'd texts must be corrected.** In order to meet the objective, we have developed an open source code editor that enables the editing of erroneous text.

Revizor (XML editor)

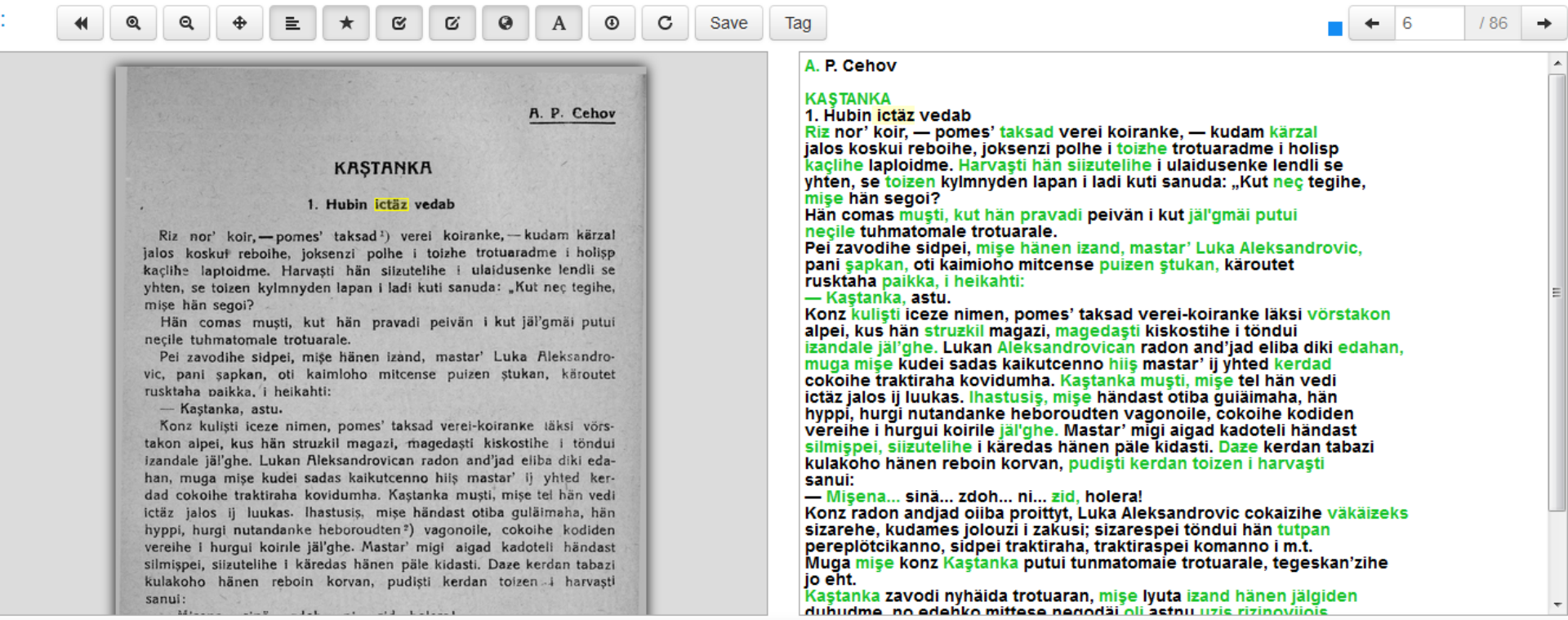
- The **Python back-end** of the software serves up the data to anybody authorised to make edits. It is also responsible for exporting the end result, which might be the corrected ALTO XML, plain text versions or word lists built from selected works. The exported product might be further processed by linguistic tools or imported into “corpus” sites specifically meant for facilitating searching and dissemination.

- **The front-end** of the editor is the part where manual changes are made by real human beings, as opposed to the automatic bulk processing in the back-end. By means of a two-pane window, the user can check the original work and make corrections to the text, as well as mark the language or relevance of words – this will aid the back-end in building accurate word lists per language.

Revizor (XML editor)



Revizor (XML editor)



http://ocrui-kk.lib.helsinki.fi/

Crowdsourcing the Finno-Ugrian material

- We have estimated that the Fenno-Ugrica collection could 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)

”Traditional” Crowdsourcing

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

Visualisation of the Problem

Trondheim,
Norway
April 19th, 2015

Visualisation of the Problem

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

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 the pages which we have digitized, but give the niches such assignments which **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

- E.g. 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 text editor.

Nichesourcing and Language Communities

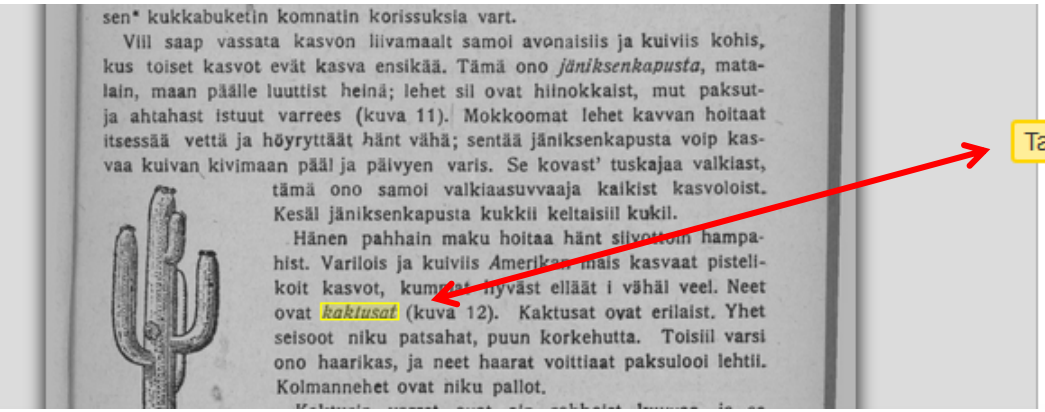
- When the language communities involve, it is essential that the **altruism** plays a central role.
- Upon the nichesourcing, our goal is to reach a certain level of **interplay**, where the language communities would benefit on the results.
- This objective of interplay can be understood as an aspiration 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 (Izhoran), an endangered language, spoken west of 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 and provide additional information.

Nichesourcing and Language Communities

- Skilled and educated people can do a lot!



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 paksut- ja 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 *Amerikaa* mais kasvaat pistelikoit kasvot, kumm[illegible] 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.
[illegible]an mais kasvaat pistelikoit
[illegible]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

Crowdsourcing vs Nichesourcing

- Traditional approach gives you only

bród

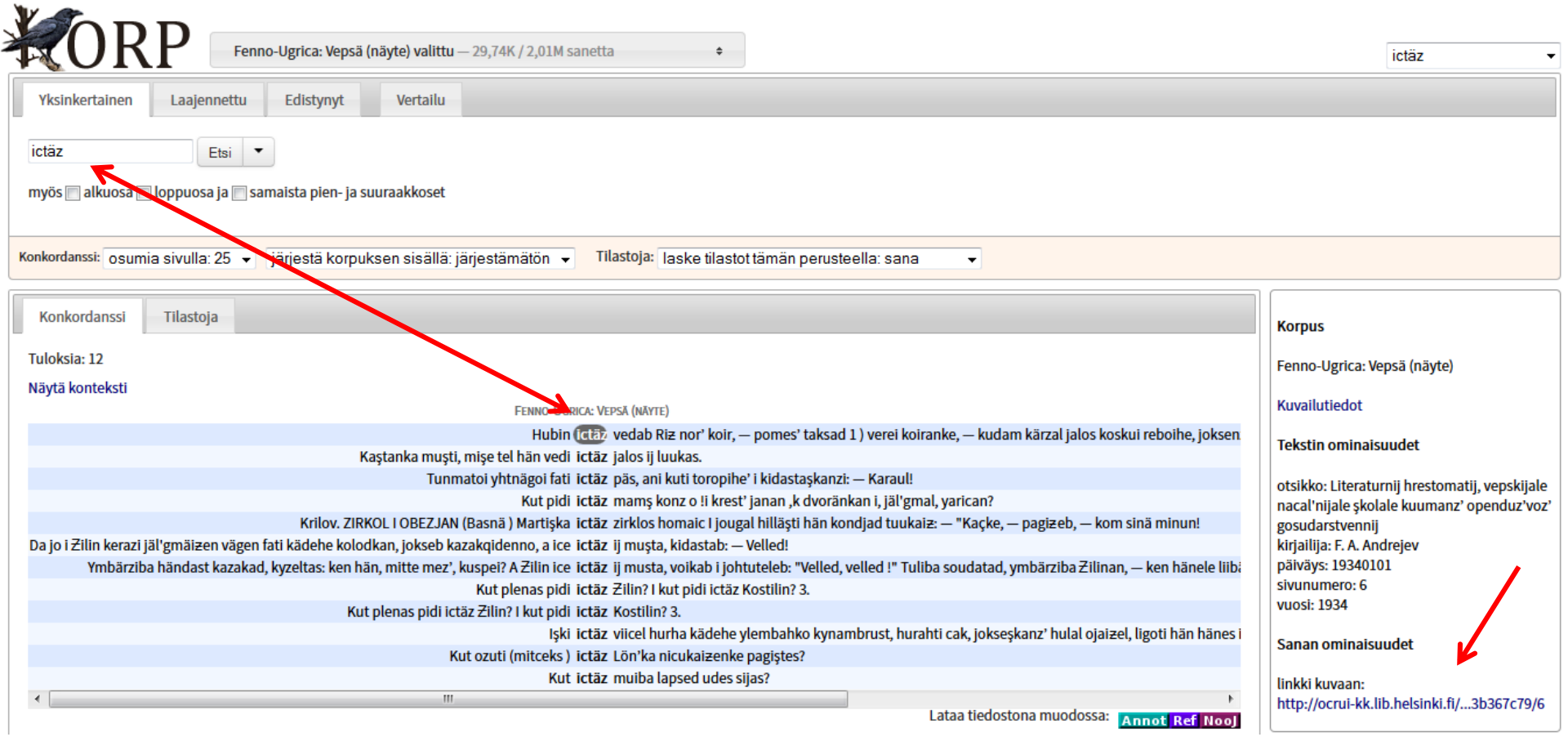
- Whereas nichesourcing gives you

brød

…or potentially more:

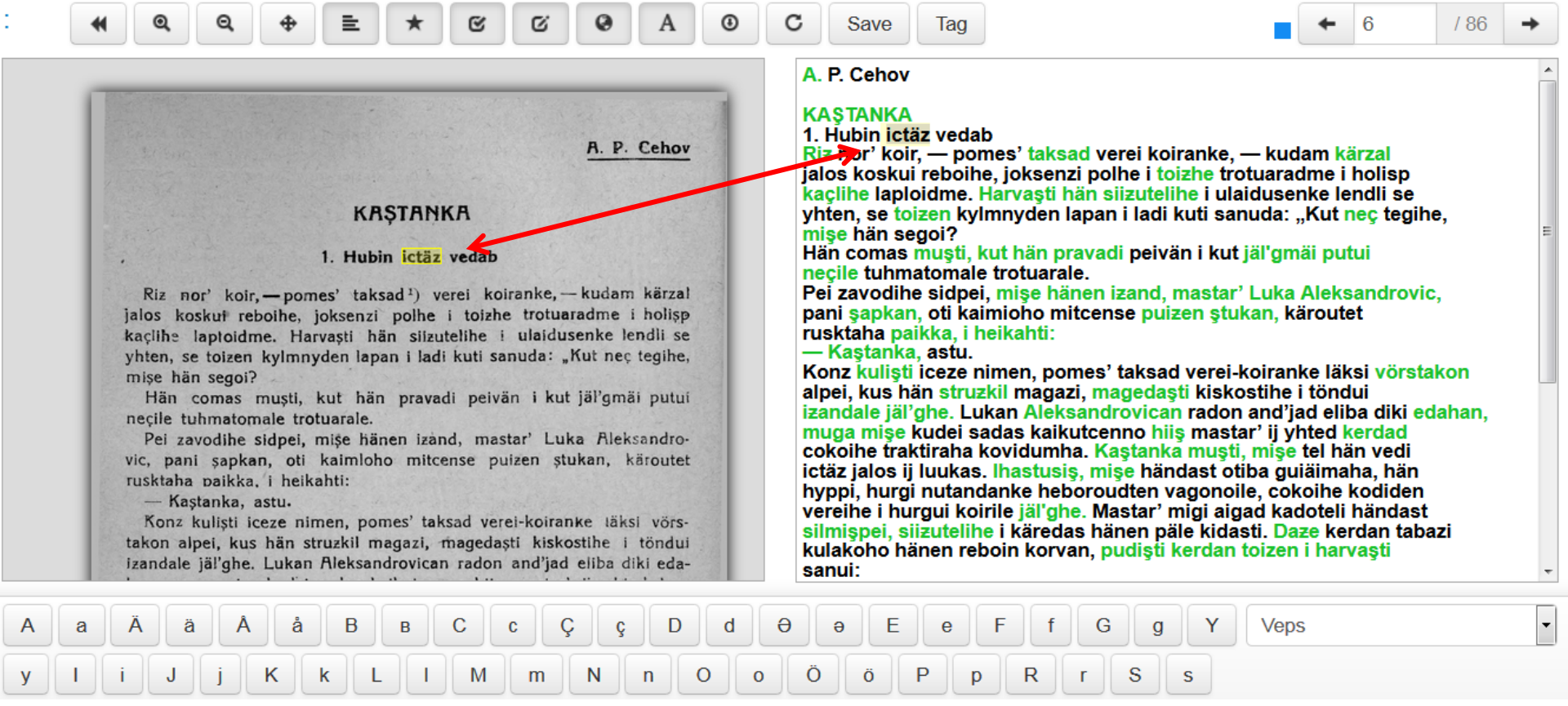
brød / bröd / bread / bröt / leipä / chleb etc.

Datasets and Further Use



[Link to the test material]

Datasets and Further Use



Nichesourcing and Language Communities

- How to locate suitable people to crowdsourcing / nichesourcing?
- Not easy to find the niches with purposeful capabilities for all languages
- 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

Nichesourcing and Language Communities

- When thinking of the possible **niches / crowds**, one must bear in mind that the most of the people 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

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 Revizor** 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**.

Contact Details

jussi-pekka.hakkarainen@helsinki.fi

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