Making the impact on research and society
a case study: crowdsourcing solutions developed for the linguistic research and citizen science

Jussi-Pekka Hakkarainen
Project Manager
Digitization Project of Kindred Languages
National Library of Finland

LIBER 43rd Annual Conference
Research Libraries in the 2020 Information Landscape
Riga, Latvia, 2 July 2014

Digitization Project of Kindred Languages

The **National Library of Finland** is implementing the **Digitization Project of Kindred Languages** in 2012–16.

Within the project we will digitize materials in the 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 it is part of the 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.

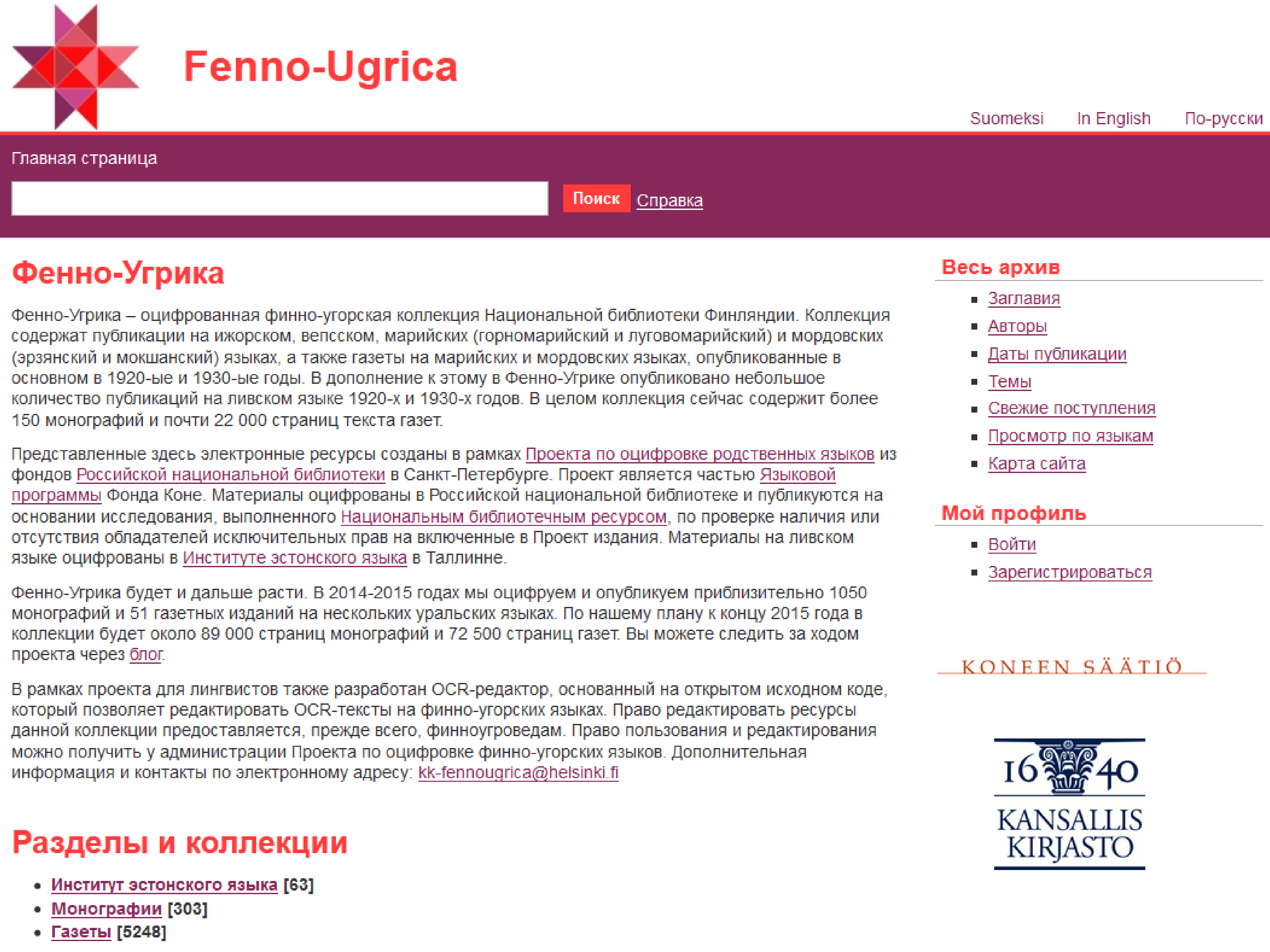
Our objective within the Language Programme is to make sure that both old and new corpora are made available for the open and interactive use of both the academic community and the language societies as a whole.

Materials and Collection

The project seeks to digitize and publish approximately 1200 monograph titles and more than 100 newspapers titles in various **Uralic languages.**

The digitization will be completed by the end of 2014, and the **Fenno-Ugrica** collection will consist of 110,000 monograph pages and about 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.



Фенно-Угрика

Фенно-Угрика – оцифрованная финно-угорская коллекция Национальной библиотеки Финляндии. Коллекция содержат публикации на ижорском, вепсском, марийских (горномарийский и луговомарийский) и мордовских (эрзянский и мокшанский) языках, а также газеты на марийских и мордовских языках, опубликованные в основном в 1920-ые и 1930-ые годы. В дополнение к этому в Фенно-Угрике опубликовано небольшое количество публикаций на ливском языке 1920-х и 1930-х годов. В целом коллекция сейчас содержит более 150 монографий и почти 22 000 страниц текста газет.

Представленные здесь электронные ресурсы созданы в рамках Проекта по оцифровке родственных языков из фондов Российской национальной библиотеки в Санкт-Петербурге. Проект является частью Языковой программы Фонда Коне. Материалы оцифрованы в Российской национальной библиотеке и публикуются на основании исследования, выполненного Национальным библиотечным ресурсом, по проверке наличия или отсутствия обладателей исключительных прав на включенные в Проект издания. Материалы на ливском языке оцифрованы в Институте эстонского языка в Таллинне.

Фенно-Угрика будет и дальше расти. В 2014-2015 годах мы оцифруем и опубликуем приблизительно 1050 монографий и 51 газетных изданий на нескольких уральских языках. По нашему плану к концу 2015 года в коллекции будет около 89 000 страниц монографий и 72 500 страниц газет. Вы можете следить за ходом проекта через блог.

В рамках проекта для лингвистов также разработан OCR-редактор, основанный на открытом исходном коде, который позволяет редактировать OCR-тексты на финно-угорских языках. Право редактировать ресурсы данной коллекции предоставляется, прежде всего, финноугроведам. Право пользования и редактирования можно получить у администрации Проекта по оцифровке финно-угорских языков. Дополнительная информация и контакты по электронному адресу: kk-fennougrica@helsinki.fi

Разделы и коллекции

- **Институт эстонского языка [63]**
- **Монографии [303]**
- **Газеты [5248]**

Весь архив

- Заглавия
- Авторы
- Даты публикации
- Темы
- Свежие поступления
- Просмотр по языкам
- Карта сайта

Мой профиль

- Войти
- Зарегистрироваться

Selection criteria of material

The selection of the materials has been made **in co-operation with the researchers** and we used several criteria to meet the requirements:

- 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)

Languages of publications

Mari
- Meadow Mari
- Hill Mari

Samoyedic
- Nenets
- Selkup

Ob-Ugric
- Khanty
- Mansi

Mordvinic
- Erzyan
- Moksha
- (Shoksha)

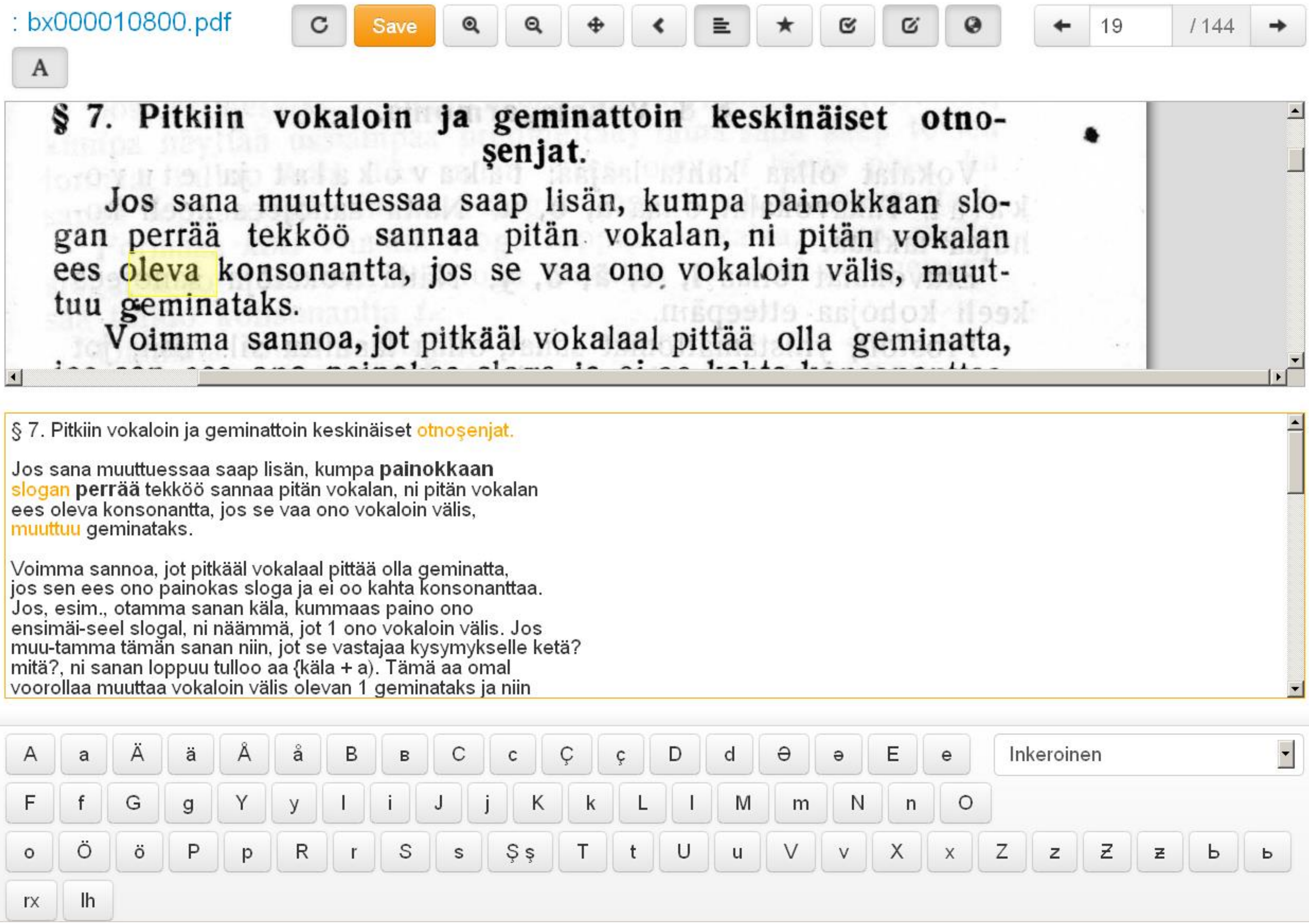
Permic
- Udmurt
- Komi-Zyrian
- Komi-Permyak

Baltic Finns
- Ingrian
- Veps
- [Livonian]

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 as such 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.


: bx000010800.pdf
Save
19
/ 144
A
Inkeroinen
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 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 U u V v X x Z z Ƶ ƶ Ь ь
rx lh

§ 7. Pitkiin vokaloin ja geminattoin keskinäiset otnoşenjat.

Jos sana muuttuessaa saap lisän, kumpa painokkaan slogan perrää tekköö sannaa pitän vokalan, ni pitän vokalan ees oleva konsonantta, jos se vaa ono vokaloin välis, muuttuu geminataks.

Voimma sannoa, jot pitkääl vokalaal pittää olla geminatta,

§ 7. Pitkiin vokaloin ja geminattoin keskinäiset otnoşenjat.

Jos sana muuttuessaa saap lisän, kumpa **painokkaan**
slogan **perrää** tekköö sannaa pitän vokalan, ni pitän vokalan
ees oleva konsonantta, jos se vaa ono vokaloin välis,
muuttuu geminataks.

Voimma sannoa, jot pitkääl vokalaal pittää olla geminatta,
jos sen ees ono painokas sloga ja ei oo kahta konsonanttaa.
Jos, esim., otamma sanan käla, kummaas paino ono
ensimäi-seel slogal, ni näämmä, jot 1 ono vokaloin välis. Jos
muu-tamma tämän sanan niin, jot se vastajaa kysymykselle ketä?
mitä?, ni sanan loppuu tulloo aa {käla + a). Tämä aa omal
voorollaa muuttaa vokaloin välis olevan 1 geminataks ja niin

Crowdsourcing the material of Fenno-Ugrica

We have estimated that the Fenno-Ugrica collection contains around 200 000 pages of editable text by the end of 2014. The researchers cannot spend so much time with the material that they could retrieve a satisfactory amount of edited words, so the participation of a crowd is needed.

Could crowdsourcing be used here to gain results?

Crowdsourcing and citizen science

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.

Crowdsourcing and citizen science

Often the targets in crowdsourcing have 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 be handled by **the faceless crowd** and the needs of linguistic research are not necessarily met. Also, the number of pages is **too high** to deal with.

The remarkable downside is the lack of shared goal or social affinity. There is **no reward** in traditional methods of crowdsourcing.

Nichesourcing with the help of 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. Communities have purpose, identity and their regular interactions engenders social trust and reputation.

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 with the help of 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 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 words related to **human organs** by assigning the citizen scientists to edit and collect words with OCR editor.

Interplay and altruism in crowdsourcing

From the crowdsourcing’s (nichesourcing’s) perspective, it is essential that the **altruism** plays a central role, when the language communities involve.

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 and the society can benefit too.

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.

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 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.

Additional Information and contact details

National Library of Finland
www.nationallibrary.fi/

Fenno-Ugrica Collection
fennougrica.kansalliskirjasto.fi/

Project Blog
blogs.helsinki.fi/fennougrica/

V Kontakte
vk.com/fennougrica