OCR EDITOR FOR LINGUISTIC CORPORA

OCR resuHs are not alvvays 100% corred.

But they must be right in linguistic corpora, literary research, republishing, and other **digital humanities.**

Manual work is tedious. But some OCR editing can be **fun, educating, and rewarding.**

Our OCR editor **crowdsources** correction work among students, fellow researchers, citizen scientists, etc.

”Students used to complain that the course of literary old Finnish was difficult, boring, and useless.

OCR editing by students yielded high motivation and great learning results.”

1. Digitize

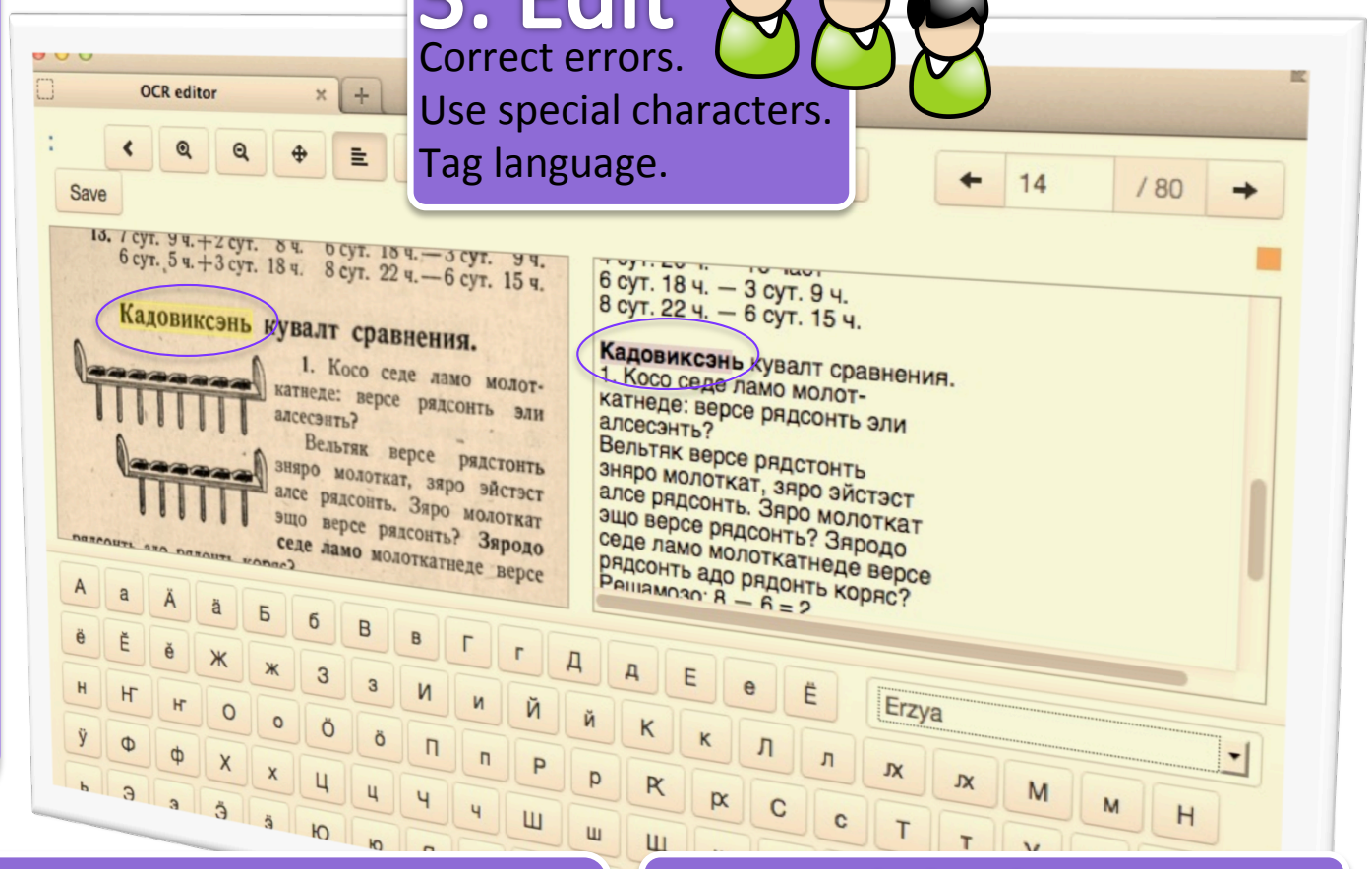
Digitize into TIFF. OCR into ALTO XML. Upload to the OCR Editor.

2. Engage

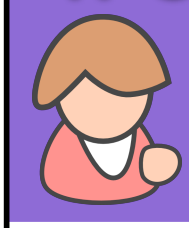
Get helpers.
Create accounts.
Assign editing rights.

3. Edit

Correct errors.
Use special characters.
Tag language.



4. Use



Lock ready documents.
Export ALTO XML.
Analyse.

5. Publish

E.g. fennougrica.kansalliskirjasto.fi

```
--- --- --- <TextLine HEIGHT="67"
WIDTH="1035" VPOS="1243" HPOS="343"
STYLEREFS="StyleId-9F2EEF00-
D8D9-4FCB-8CCB-20D078F0B86F-"><String
CONTENT="Кадовиксэнь" HEIGHT="60"
WIDTH="401" VPOS="1243" HPOS="343"/
><SP WIDTH="37" VPOS="1257" HPOS="745"/
><String CONTENT="кувалт" HEIGHT="53"
WIDTH="217" VPOS="1257" HPOS="783"/>
```

OCR Editor is an editor for ALTO XML files. ALTO is supported by most OCR software. It contains the characters, language and coordinate attributes for each word, as well as basic layout of sentences and paragraphs. It also may include customized editing marks.

The Python back-end of the OCR Editor manages documents and serves them to the editor. It also handles revisions, as well as user accounts and authorization management.

The OCR Editor web user interface (JavaScript) maps ALTO XML attributes to words on a web page. Users can correct errors side by side with the image, and edit meta-attributes.

To start a project, scanned images and ALTO files are imported into the system. JPG and thumbnails for the web interface are created.

An administrator supervises a collection and grants editing rights to collections or documents.

What we do with the OCR Editor:

- We produce **linguistic corpora.** We work with Soviet publications from 1920s and 1930s in Finno-Ugric languages, such as Veps, Erzya and Moksha. We also work with literary monuments of Finnish.
- We produce **texts suitable for critical study and republishing,** such as 19th century works of Finnish prose and poetry
- We publish the results in **Dspace.**

Authors: Esa-Pekka Keskitalo & Jussi-Pekka Hakkarainen.
Tool developers and contributors: Wouter van Hemel, Juho Vuori, and Anis Moubarik.

National Library of Finland, www.nationallibrary.fi, firstname.lastname@helsinki.fi

THE NATIONAL LIBRARY OF FINLAND

The project is funded by the Kone Foundation, www.koneensaatio.fi/

KONEEN SÄÄTIÖ

Presented at
OR2014
or2014.helsinki.fi