OCR Webinar
Helsinki, March 5th, 2014

Subversioned OCR Editing,
an Opening for Community Involvement

Jack Rueter, “Morphological Parsers for Minority Finno-Ugrian Languages” project

Funded by The Kone Foundation, Helsinki, Finland

Subversioned OCR Editing, an Opening for Community Involvement

- Digitization Project of Kindred Languages
- Enhancement Tools
- Enhancement Tool = OCR Editor
- Opportunities
- Where do I use it?
- Ideas for proof-reading
- Where might the results go

Digitization Project of Kindred Languages

- Digitization
  - Preservation (limited space and facilities)
  - Open-Access (copyrights)
  - Searchability (text recognition)
- 1920s and 1930s Minority Media
  - Newspapers (1917-1940)
  - Monographs (1930s)
- Enhancement Tools

Enhancement Tools

- Motivations
  - High-quality commercial text recognition is excellent for main-stream texts in majority languages, but…
  - Lesser documented language forms do not endorse these tools
    - Stategies: character set recognition, word lists, simple regular expressions
    - Short-comings:
      - Not even all Unicode characters are acceptable
      - Where do you get the word lists before you even have the words
    - Lower quality output texts for lesser documented language forms
  - Smaller language communities have fewer work hour resources
    - Subversioned open editing possibilities
      - Location where research and language community members can contribute for mutual benefits and minimization of resource expenditures
  - Making open literature open, accessible, available and searchable
    - Shared treasures for mutual benefits made possible
    - An enhanced electronic library for preservation and world-wide utilization

Enhancement Tool = OCR Editor

- A sub-versioned open-source editor
- Opening for Community Involvement
  - Inter-community Involvement:
    - Research Community (invokers)
    - Language Community (source)
    - Education
    - Interested Public
  - Proof-read materials become available and searchable
    - Research Community and Education
    - Language Community and Interested Public

Opportunities

- Accessibility, availability and searchability of media and their enhancement:
  - Strengthening a sense of community through mutual planning and achievement
  - Documentation of ample materials previously ignored
    - Linguistic
    - Historic
  - Educational application
  - Open for use, modification and development
    - Subversioned proof-reading is parallel to morphological analyzer development

Where do I use it?

- Subversioned proof-reading allows
  - Correction of text recognition in source materials during test material collection
  - Extensive online access to test materials for development of morphological and syntactical analysis strategies
  - Improved analysis improves recognition of subsequent texts.

Ideas for proof-reading

- Incorporation of proof-reading in education
  - Experience
    - Old Written Finnish: University of Tampere (Postilla)
  - Wish list
    - Computer-assisted learning environments incorporating the proof-reading of OCR documents (citizen science)
    - Incorporation of proof-reading gamification (crowdsourcing)

Where might the results go

- Proof-read texts will be forwarded to the research community
  - The Finnish Language Bank (kielipankki)
- Open-access materials should be available to all interested
  - This means conversion of texts to download documents for reading devices

THANK YOU