KANSALLISARKISTO

Miten annotointi toteutettiin

Tekoälyllä lisäarvoa digiarkistojen asiakkaille

1.9.2023-31.8.2024

Vesa Laitinen Sini Rajaniemi

Euroopan unionin osarahoittama

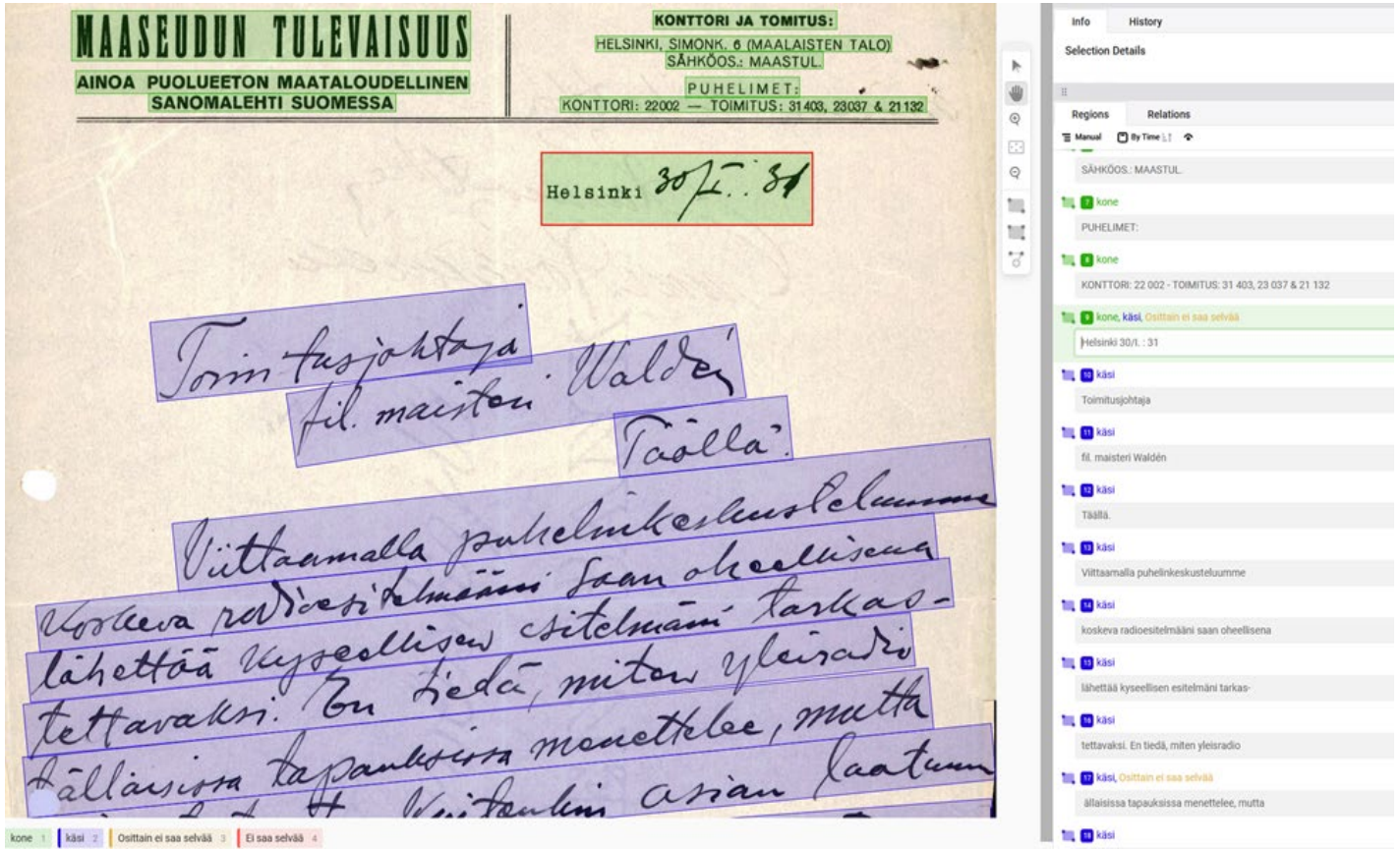
XAMK Kaakkois-Suomen ammattikorkeakoulu

elka SUOMEN ELINKEINOELÄMÄN KESKUSARKISTO

Miten annotointi toteutettiin

- ELKAn aineistojen lisäksi Kansallisarkiston aineistoja opetusaineistoina:
  - Kantakortit (sodanajan upseerien), OCR-annotoitiin *Label Studio* -sovelluksessa
  - Sotapäiväkirjat, OCR-tarkistuksia *Transkribus* -sovelluksessa
- Hankkeen aikana OCR-annotoinnin tuloksena 40000 tekstiriviä, joista 31621 konekirjoitettua ja 8310 käsinkirjoitettua, 1800 sivua
- Annotoinnin säännöt muotoutuivat aineistojen mukaan yhteisissä kokouksissa viikoittain
- Kansallisarkistolta 50% työajalla annotoijina Vesa Laitinen ja Sini Rajaniemi

OCR-annotointi Label Studiolla

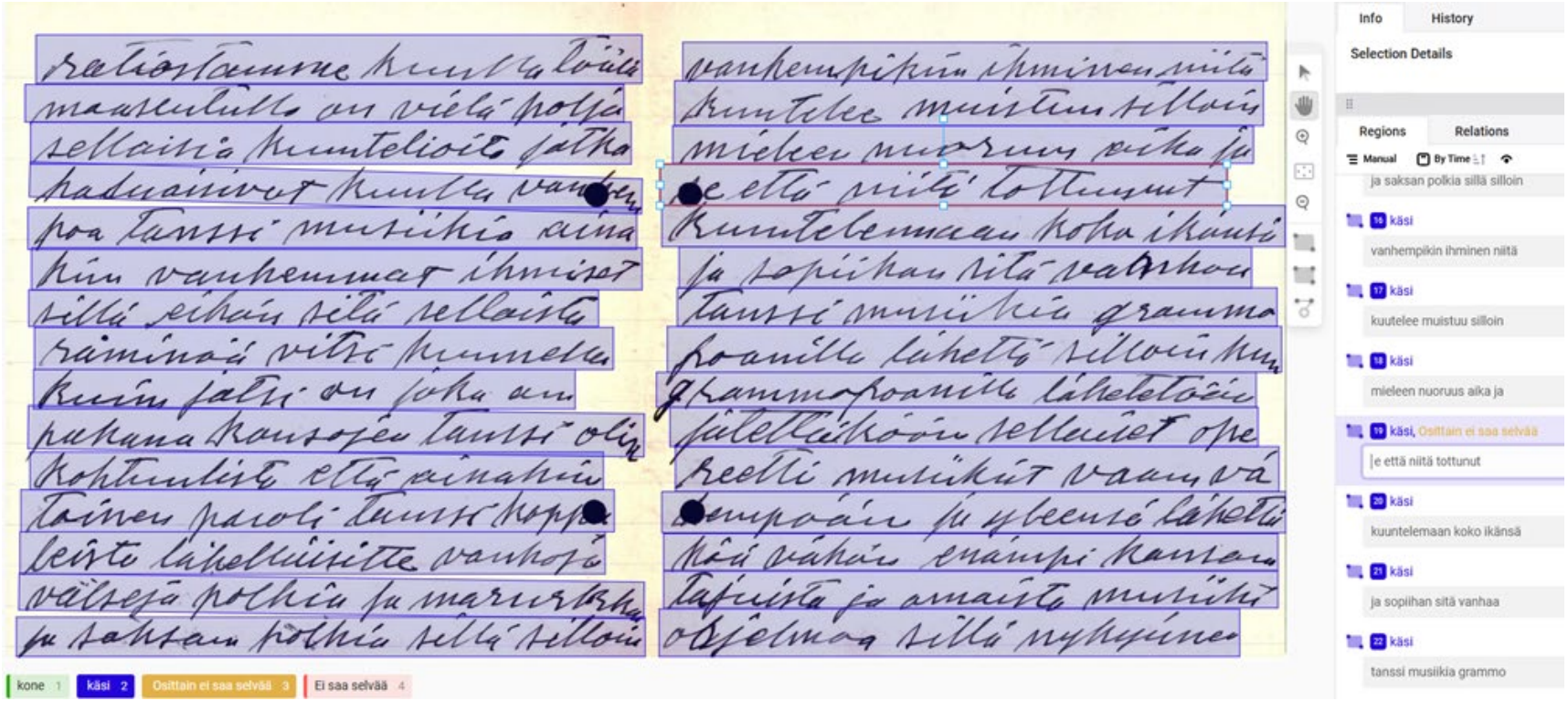


- Esiannotoidut kuvat
- Tekstialueiden korjaukset ja lisäykset
- OCR-tekstien korjaukset, suomen- ja ruotsinkielisiä pääasiassa
- Käytettiin kirjoitusmerkkejä, jotka sai suoraan näppäimistöltä
- Jos kirjoitusmerkkiä ei löytynyt näppäimistöltä, valittiin luokaksi ”ei saa selvää"
- Tekstialueiden luokittelu:
  - kone
  - käsi
  - osittain ei saa selvää
  - ei saa selvää

Esimerkkikuva ELKAn opetusaineistosta, YLE kirjeenvaihtoa 1927-1966

KANSALLISARKISTO

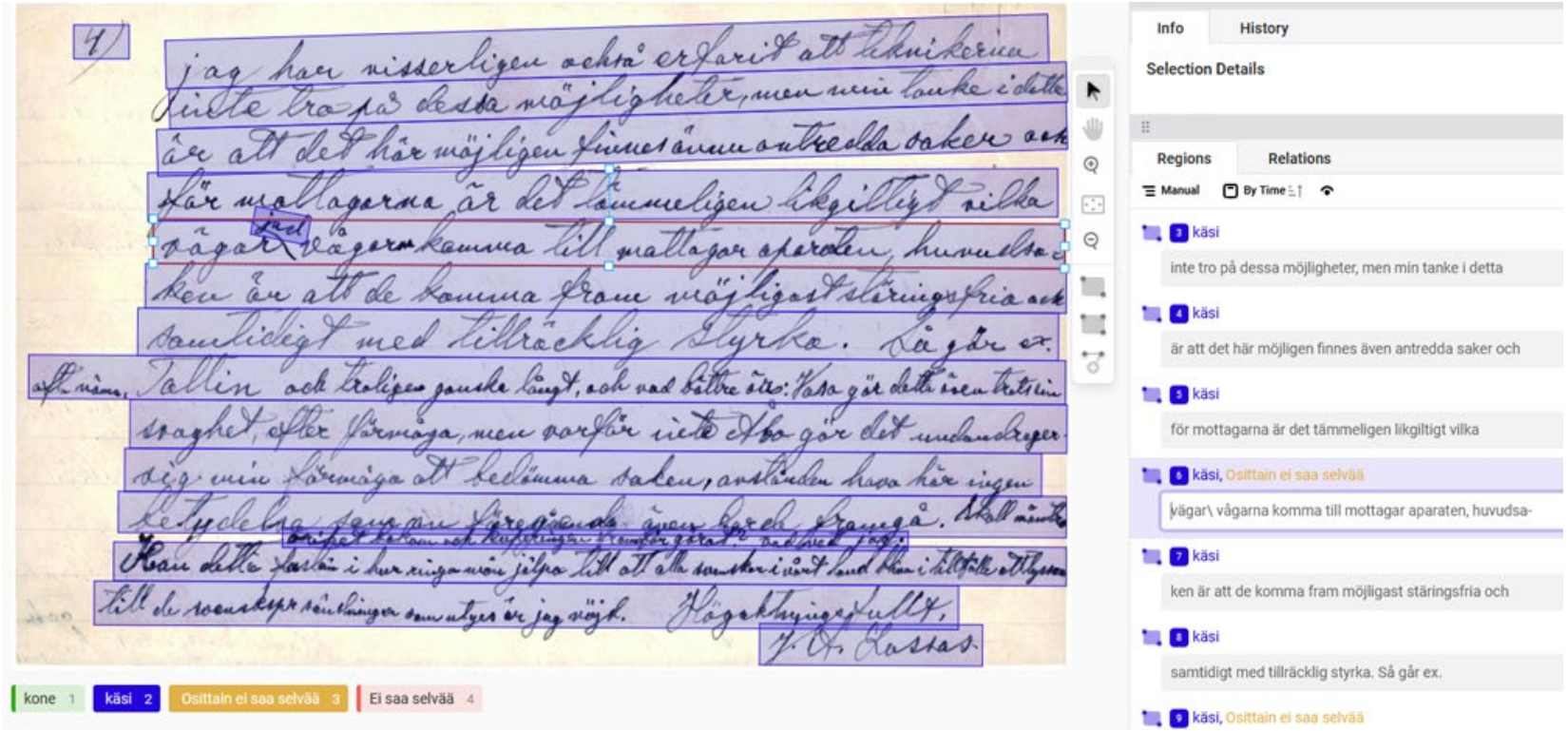
OCR-annotointi Label Studiolla



Esimerkkikuva ELKAn opetusaineistosta, YLE kirjeenvaihtoa 1927-1966

KANSALLISARKISTO

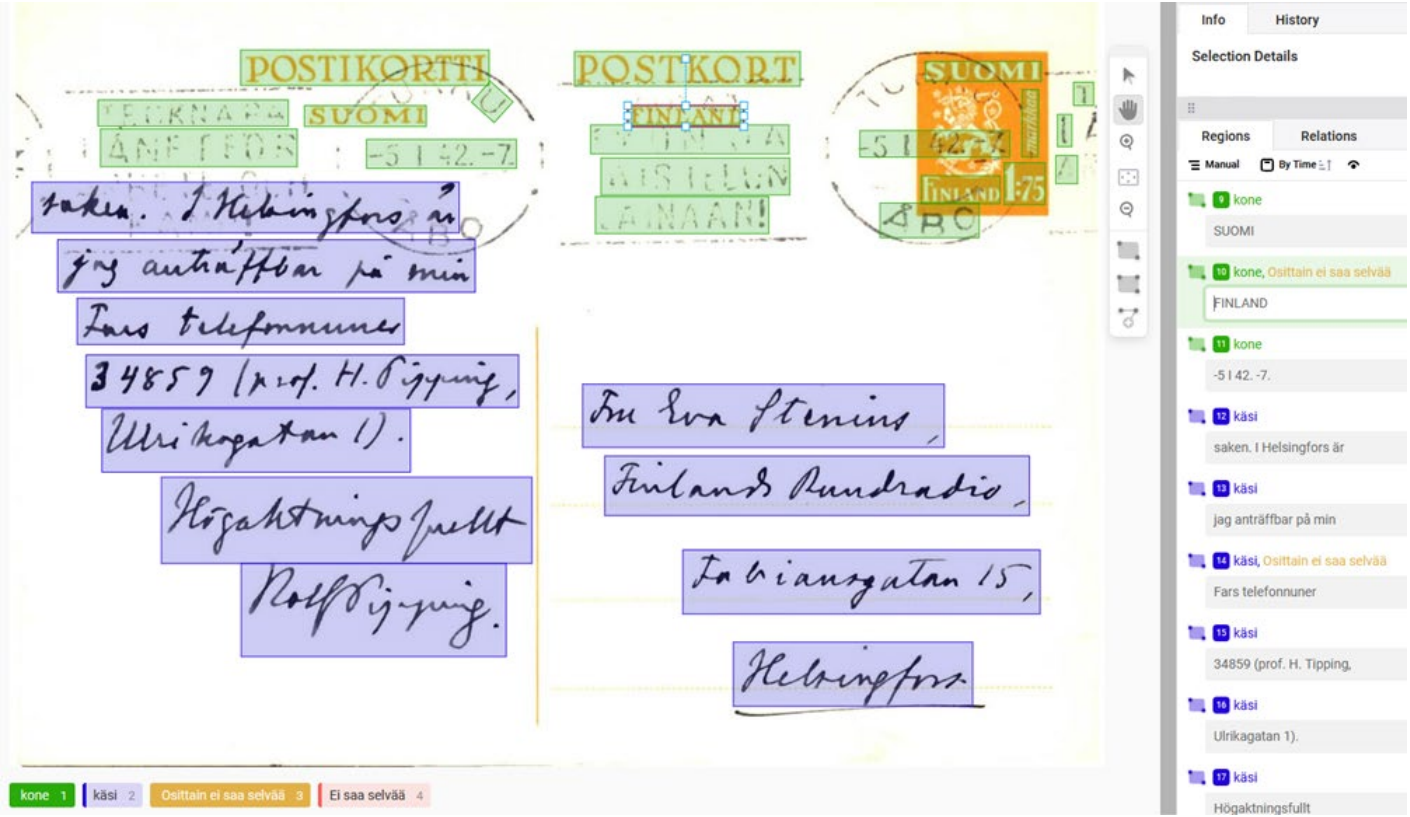
OCR-annotointi Label Studiolla



Esimerkkikuva ELKAn opetusaineistosta, YLE kirjeenvaihtoa 1927-1966

KANSALLISARKISTO

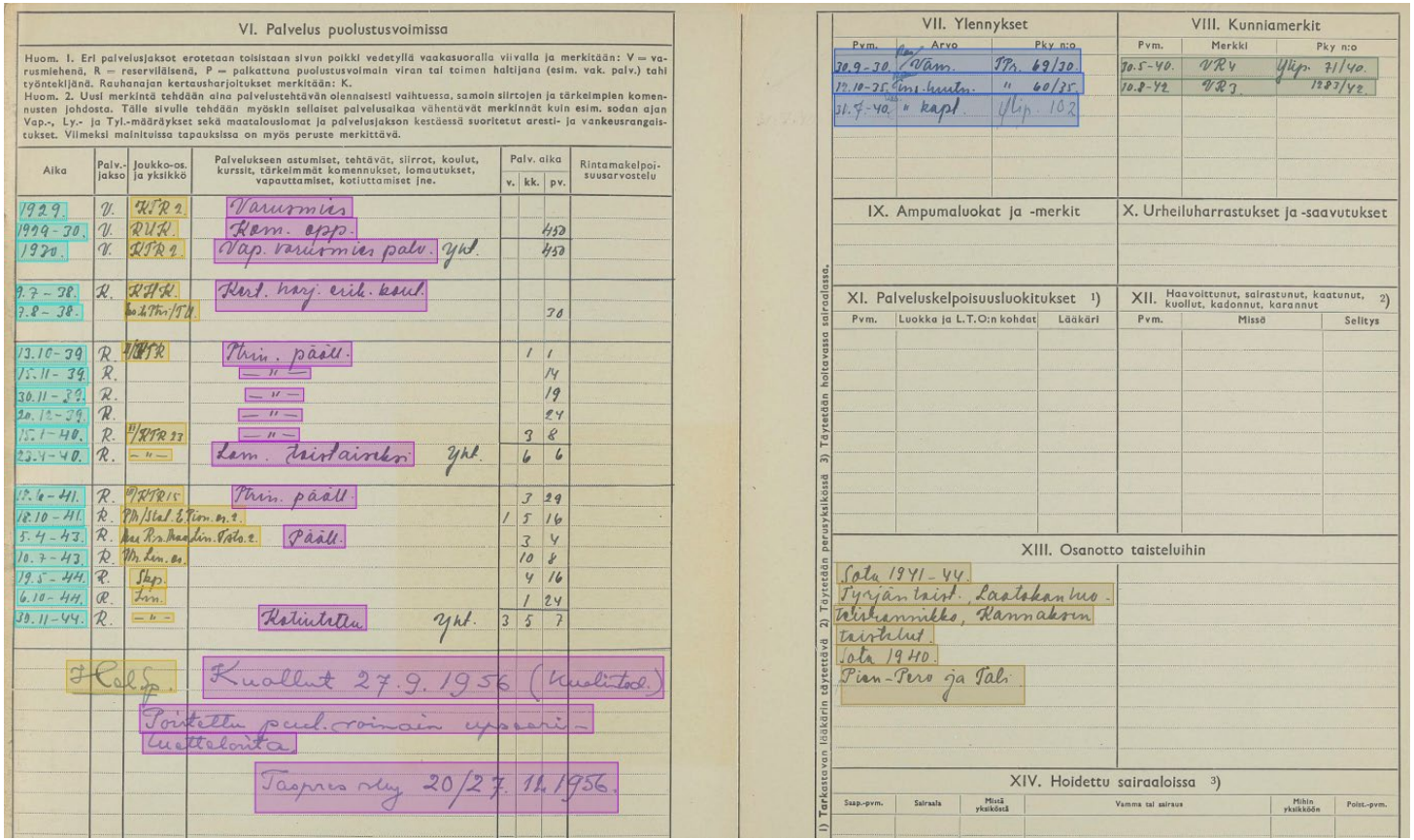
OCR-annotointi Label Studiolla



Esimerkkikuva ELKAn opetusaineistosta, YLE kirjeenvaihtoa 1927-1966

KANSALLISARKISTO

OCR-annotointi Label Studiolla

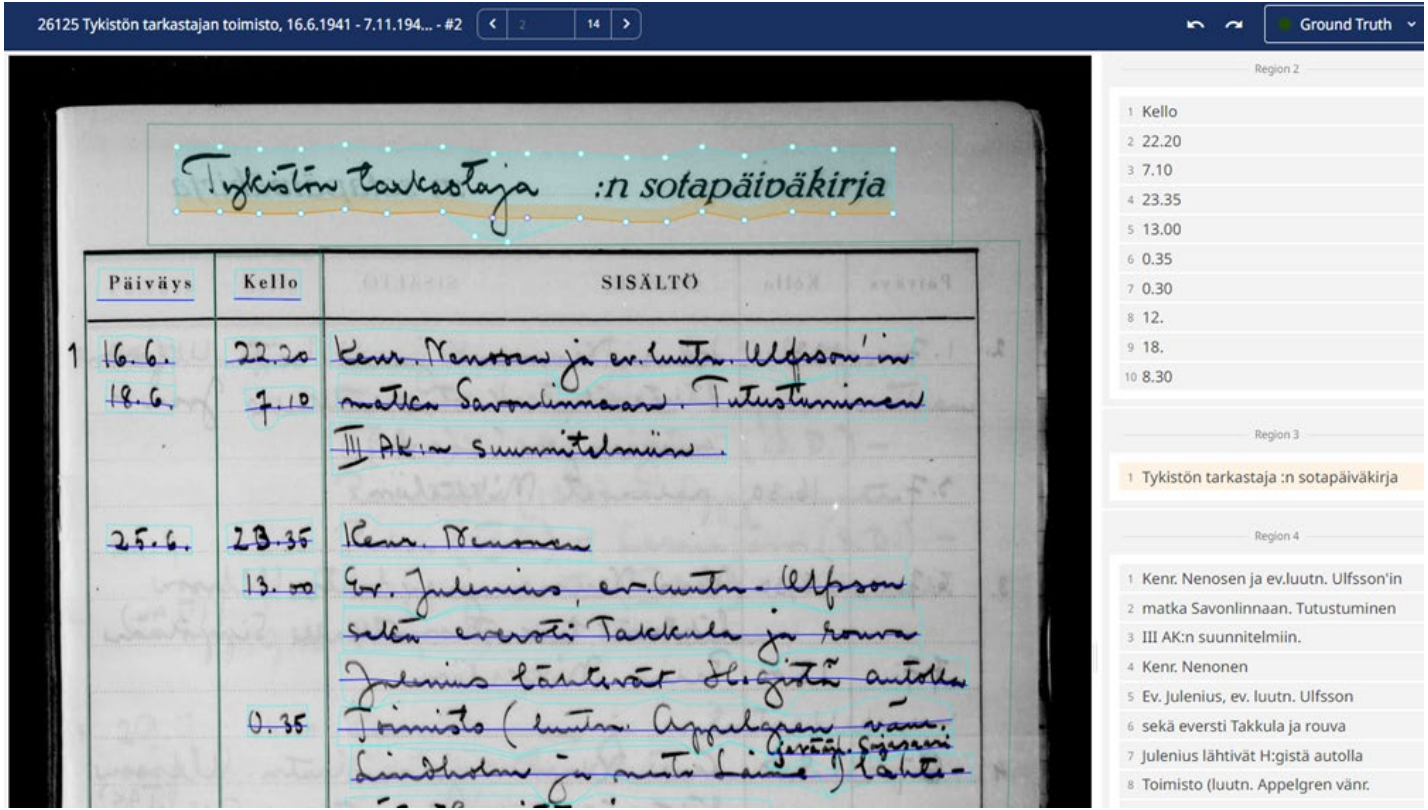


- Luokat kantakorteissa:
  - nimi
  - syntymäaika
  - syntymäpaikka
  - tuntolevyn numero
  - palvelukseen astumispäivämäärä
  - joukko-osasto
  - tehtävä
  - ylennykset
  - kunniamerkit
  - osanotto taisteluihin
  - osittain ei saa selvää
  - ei saa selvää

Esimerkkikuva Kansallisarkiston opetusaineistosta, kantakortti

KANSALLISARKISTO

Sotapäiväkirja, Transkribus



- Transcribus -sovelluksessa tehtiin OCR-tarkistusta, tekstialueiden rajausta ja lisäyksiä yli 100 kuvaan, suurin osa oli aukeamia

Esimerkkikuva Kansallisarkiston opetusaineistosta

KANSALLISARKISTO

Kiitos

KANSALLISARKISTO