



KANSALLISARKISTO

Miten annotointi toteutettiin

Tekoälyllä lisäarvoa digiarkistojen asiakkaille

1.9.2023-31.8.2024

Vesa Laitinen

Sini Rajaniemi



**Euroopan unionin
osarahoittama**



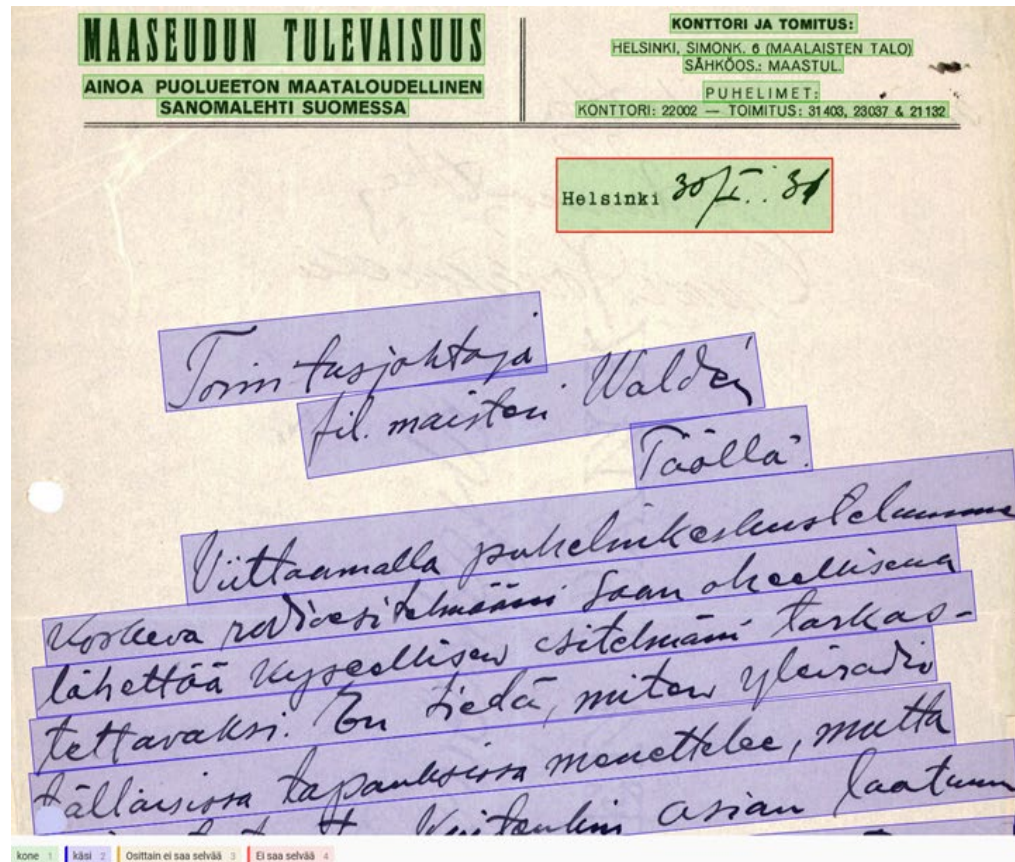
Kaakkois-Suomen
ammattikorkeakoulu



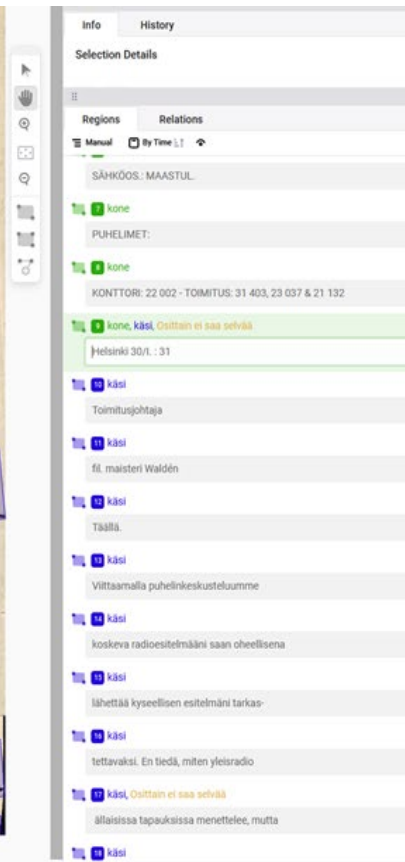
Miten annotointi toteutettiin

- 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



Esimerkkikuva ELKAN opetusaineistosta, YLE kirjeenvaihtoa 1927-1966



- 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

OCR-annotointi Label Studiolla

rationaalisuus kunta loiva
maaseutulla on vielä polja
sellaisia kunnallioite jotka
kaduainivat kunta vanha
poa tanssi musiikis aina
kun vanhemmat ihmiset
sillä eihän sitä sellaista
räminää vitse kunnella
kunin jalki on joka on
pakana kansojen tanssi oli
kohtuulit, että ainakin
tainen paroli tanssi koppa
leivä lähelleiitte vanhoja
välpejä polkia ja marurkka
ja sahtaan kolkkia sillä silloin

vanhempiin ihmisen niitä
kuuntelee muistuu silloin
miehen nuoruus aika ja
se että niitä tottunut
kuuntelemaan koko ikänsä
ja sopihan sitä vanhaa
tanssi musiikis grammo
foanilla lähetti silloin kun
grammofoanilla lähelöäie
jällelläköön sellaiset ope
reelli musiikit vaanyvä
lompoin ja yleensä lähetti
kii vähän enämpi kannan
tajuista ja amaista musiikit
objektmaa sillä nykyinen

Info History

Selection Details

Regions Relations

Manual By Time

ja saksan polkia sillä silloin

16 käsi
vanhempiin ihmisen niitä

17 käsi
kuuntelee muistuu silloin

18 käsi
mieleen nuoruus aika ja

19 käsi, Osittain ei saa selvää
|e että niitä tottunut

20 käsi
kuuntelemaan koko ikänsä

21 käsi
ja sopihan sitä vanhaa

22 käsi
tanssi musiikis grammo

kone 1 käsi 2 Osittain ei saa selvää 3 Ei saa selvää 4

Esimerkkikuva ELKAN opetusaineistosta, YLE kirjeenvaihtoa 1927-1966

OCR-annotointi Label Studiolla

4

jag har nisserligen och så erfärit att teknikererna
inte tro på dessa möjligheter, men min tanke i detta
är att det här möjligen finnes ännu antreda saker och
sådär mottagarna är det tåmmeligen likgiltigt vilka
säger vägar komma till mottagar apparaten, huvudsak-
ligen är att de komma fram möjligast stäringfria och
samtidigt med tillräcklig styrka. Så går ex.
Tallin och träligen ganska långt, och vad bättre är: Kasa gör detta även trots sin
svaghet, efter försök, men varför inte ska gör det underdiger
sig min försök att besluma saken, antuden hava här ingen
beträffande saken som försökade. men här och framgå. Skall minsta
Kasa detta stann i hur ringa min hjälp till att alla saken i vänt land Kasa i tillfälle att göra
till de saken som saken som utger är jag väj. Kåghäring fullt.
J. O. Kostas.

Info History

Selection Details

Regions Relations

Manual By Time

1 käsi
inte tro på dessa möjligheter, men min tanke i detta

4 käsi
är att det här möjligen finnes även antreda saker och

3 käsi
för mottagarna är det tåmmeligen likgiltigt vilka

4 käsi, Osittain ei saa selvää
vägar vägar komma till mottagar aparaten, huvudsak-

2 käsi
ken är att de komma fram möjligast stäringfria och

4 käsi
samtidigt med tillräcklig styrka. Så går ex.

4 käsi, Osittain ei saa selvää

kone 1 käsi 2 Osittain ei saa selvää 3 Ei saa selvää 4

Esimerkkikuva ELKAn opetusaineistosta, YLE kirjeenvaihtoa 1927-1966

OCR-annotointi Label Studiolla

VI. Palvelus puolustusvoimissa					
Huom. 1. Eri palvelusjakset erotetaan toisistaan sivun poikki vedetyllä vaakasuoralla viivalla ja merkitään: V = varusmiehenä, R = reservilisenä, P = palkattuna puolustusvoimain viran tai toimen haltijana (esim. vak. palv.) tai ystäväkylänä. Rauhajan kertausharjoitukset merkitään: K.					
Huom. 2. Uusi merkintä tehdään aina palvelustehtävän ollessaan vaihtuessa, samoin siirtojen ja tarkempien komennusten johdosta. Tälle sivulle tehdään myöskin sellaiset palvelusajankäytön vähentävät merkinnät kuin esim. sodan ajan vap., L.y. ja Tyl.-määräykset sekä maatalousomat ja palvelusjakson kestäessä suoritettavat asemi- ja vankeusarngitukset. Viimeksi mainituissa tapauksissa on myös peruste merkittävä.					
Aika	Palv. jaksot	Joukko-os. ja yksikkö	Palvelukseen astumiset, tehtävät, siirrot, koulut, kurssit, tärkeimmät komennukset, lomautukset, vapauttamiset, kotiuttamiset jne.	Palv. aika v. kk. pv.	Rintamakelpoisuusarvostelu
1929.	V.	RPR 2	Varusmies		
1929-30.	V.	RPR 2	Kem. esj.	452	
1930.	V.	RPR 2	Vap. varusmies palv. yht.	452	
3.7-38.	R.	RPR 2	Kent. harj. erik. koul.		
3.8-38.	R.	ha 2/10/50		70	
13.10-39.	R.	RPR 2	Thun. pääll.	1 1	
15.11-39.	R.		"	14	
30.11-39.	R.		"	19	
20.12-39.	R.		"	24	
15.1-40.	R.	RPR 27	"	3 8	
23.4-40.	R.	"	Kem. laistaminen yht.	6 6	
15.6-41.	R.	RPR 15	Thun. pääll.	7 29	
18.10-41.	R.	Pa/10/10 & 10/10/10		1 5 16	
5.4-43.	R.	ha 2/10/50	Pääll.	3 4	
10.7-43.	R.	ha 2/10/50		10 8	
19.5-44.	R.	Shp.		4 16	
6.10-44.	R.	Lan.		1 24	
30.11-44.	R.	"	Kotiutettiin yht.	3 5 7	
			Kuollut 27.9.1956 (kuelintod.)		
			Poistettu puol.vainain upseeri-		
			luettelosta		
			Tasapaino rly 20/27.1h/1956.		

VII. Ylennykset				VIII. Kunniamerkit		
Pvm.	Arvo	Pky no.		Pvm.	Merkki	Pky no.
30.9-30	1. Am.	3Pa 69/30		30.5-40.	VR 4	Yllp. 21/40
15.10-25	2. Am. huol.	" 60/25		30.8-42	VR 3	1227/42
31.5-40.	3. kapt.	Yllp. 102				

IX. Ampumaluokat ja -merkit			X. Urheiluharrastukset ja -saavutukset		

XI. Palveluskelpoisuusluokitukset 1)			XII. Haavoittunut, sairastunut, kaatunut, kuollut, kadonnut, karannut 2)		
Pvm.	Luokka ja L.T.O:n kohdat	Lääkäri	Pvm.	Missä	Selitys

XIII. Osanotto taisteluihin	
	Sota 1941-44. Tyrjän taist. Laatokan kuo- telijamies, Kannakom- taistelut Sota 1940. Pian-Pero ja Jal.

XIV. Hoidettu sairaaloissa 3)					
Saap-pvm.	Sairaus	Missä yksikössä	Vamma tai sairaus	Mihin yksikköön	Polst-pvm.

• 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

Sotapäiväkirja, Transkribus

26125 Tykistön tarkastajan toimisto, 16.6.1941 - 7.11.194... - #2

< 2 14 >

Ground Truth

Tykistön tarkastaja :n sotapäiväkirja

Päiväys	Kello	SISÄLTÖ
16.6.	22.20	Kenr. Nenosen ja ev. luutn. Ulfsson'in
18.6.	7.10	matkan Savonlinnaan. Tutustuminen III AK:n suunnitelmiin.
25.6.	23.35	Kenr. Nenonen
	13.00	Ev. Julenius, ev. luutn. Ulfsson sekä eversti Takkula ja rouva Julenius lähtivät Högistä autolla
	0.35	Toimisto (luutn. Appelgren vänr. Lindhölm ja ruut. Laine) lähti-

Region 2

- 1 Kello
- 2 22.20
- 3 7.10
- 4 23.35
- 5 13.00
- 6 0.35
- 7 0.30
- 8 12.
- 9 18.
- 10 8.30

Region 3

- 1 Tykistön tarkastaja :n sotapäiväkirja

Region 4

- 1 Kenr. Nenosen ja ev. luutn. Ulfsson'in
- 2 matka Savonlinnaan. Tutustuminen
- 3 III AK:n suunnitelmiin.
- 4 Kenr. Nenonen
- 5 Ev. Julenius, ev. luutn. Ulfsson
- 6 sekä eversti Takkula ja rouva
- 7 Julenius lähtivät H:gistä autolla
- 8 Toimisto (luutn. Appelgren vänr.

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

Esimerkkikuva Kansallisarkiston opetusaineistosta

KANSALLISARKISTO

Kiitos