Content development efforts in the NLF's historic newspaper and journal collection: re-OCR, Named Entity Recognition and layout analysis

Kimmo Kettunen

The National Library of Finland

Mikkeli unit

European Union
European Regional
Development Fund

Leverage from
the EU
2014—2020

Background

- The National Library of Finland’s Mikkeli unit has digitized historical newspapers, journals and ephemera published in Finland since 1998.
- The present collection consists of about 14 million pages mainly in Finnish and Swedish.
- Out of these about 7.45 million pages are freely available on the web site digi.kansalliskirjasto.fi (user interface in English & Swedish available)
- The time period of the open collection is from 1771 to 1929. Years 1920–1929 were opened in January 2018.

NLF’s Mikkeli Unit

- A digitization factory about 220 km north-east of Helsinki
- Ca. 40 employees
- Digitizes about 1-2 M pages of newspapers and journals annually
- Also other production

NLF’s digital humanities efforts with the newspaper and journal collection

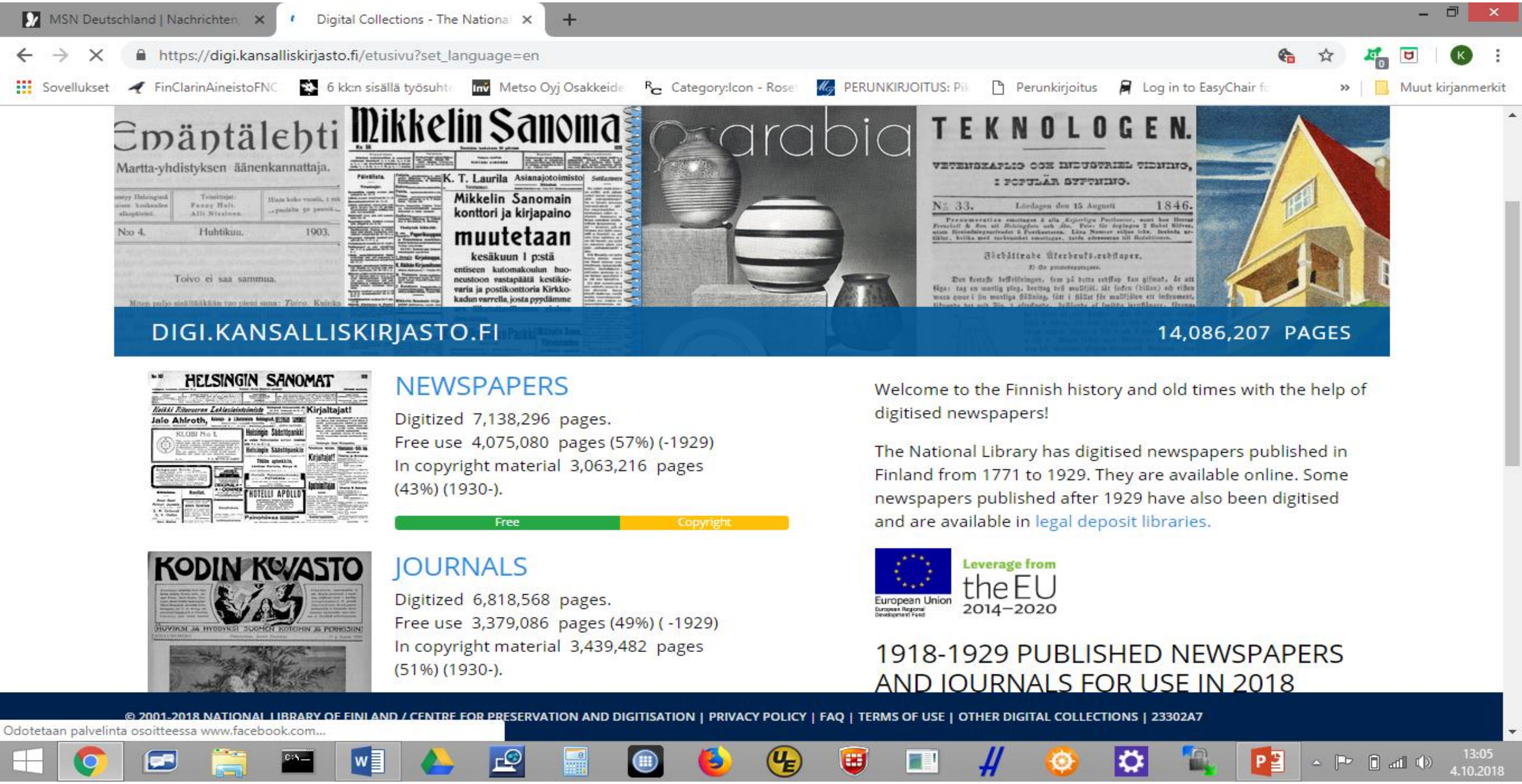
- Besides producing and publishing the digitized data all the time NLF has been involved in research and improvement of the digitized material during the last years. We ended in July 2017 a two year European Regional Development Fund project and started another two year ERDF project in August 2017 → **Digitalia (**together with South-Eastern Finland University of Applied Sciences**)**
- NLF is also involved in research consortium COMHIS that is funded by the Academy of Finland (2016–2019) and utilizes the newspaper and journal data in its research of historical changes of publicity in Finland.
- EU Horizon Project NewsEye started in May 2018 - NLF is one of the partners and provides data for the project

Data improvement and new ways to use the data

NLF has so far performed e.g. the following:

- Word level quality analysis for the Finnish part of data
- Open data delivery package of 1771-1910 newspapers and journals (available from digi.kansalliskirjasto.fi)
- Several improvements for the Web interface (time-line, notebook property etc.)
- Ground truth data of Finnish for new optical character recognition (open data)
- A new OCR process with Tesseract 3.04.01
- Named Entity Recognition evaluation collections (two phases: initial trial and present with GT OCR data)
- Layout analysis/article extraction work in progress

Historical Finnish Newspaper & Journal Web Collection: Digi


MSN Deutschland | Nachrichten
Digital Collections - The National
https://digi.kansalliskirjasto.fi/etusivu?set_language=en
Sovellukset
FinClarinAineistoFNC
6 kk:n sisällä työsuht
Metso Oyj Osakkeide
Category:Icon - Rose
PERUNKIRJOITUS: Pi
Perunkirjoitus
Log in to EasyChair f
Muut kirjanmerkit
Emäntälehti
Martta-yhdistyksen äänenkannattaja.
Mikkelin Sanomat
Mikkelin Sanomain konttori ja kirjapaino muutetaan kesäkuun 1 p:stä
arabia
TEKNOLOGEN.
Helsingin Sanomat
Kodin Kuvasto
Free
Copyright
European Union
Leverage from the EU 2014–2020
Odotetaan palvelinta osoitteessa www.facebook.com...
13:05
4.10.2018

DIGI.KANSALLISKIRJASTO.FI

14,086,207 PAGES

NEWSPAPERS

Digitized 7,138,296 pages.
Free use 4,075,080 pages (57%) (-1929)
In copyright material 3,063,216 pages (43%) (1930-).

JOURNALS

Digitized 6,818,568 pages.
Free use 3,379,086 pages (49%) (-1929)
In copyright material 3,439,482 pages (51%) (1930-).

Welcome to the Finnish history and old times with the help of digitised newspapers!

The National Library has digitised newspapers published in Finland from 1771 to 1929. They are available online. Some newspapers published after 1929 have also been digitised and are available in legal deposit libraries.

1918-1929 PUBLISHED NEWSPAPERS AND JOURNALS FOR USE IN 2018

© 2001-2018 NATIONAL LIBRARY OF FINLAND / CENTRE FOR PRESERVATION AND DIGITISATION | PRIVACY POLICY | FAQ | TERMS OF USE | OTHER DIGITAL COLLECTIONS | 23302A7

Three themes of the talk

- Quality of Optical Character Recognition/improvement of OCR
- Named Entity Recognition (NER) on historical Finnish data
- Layout analysis/article extraction work with the data

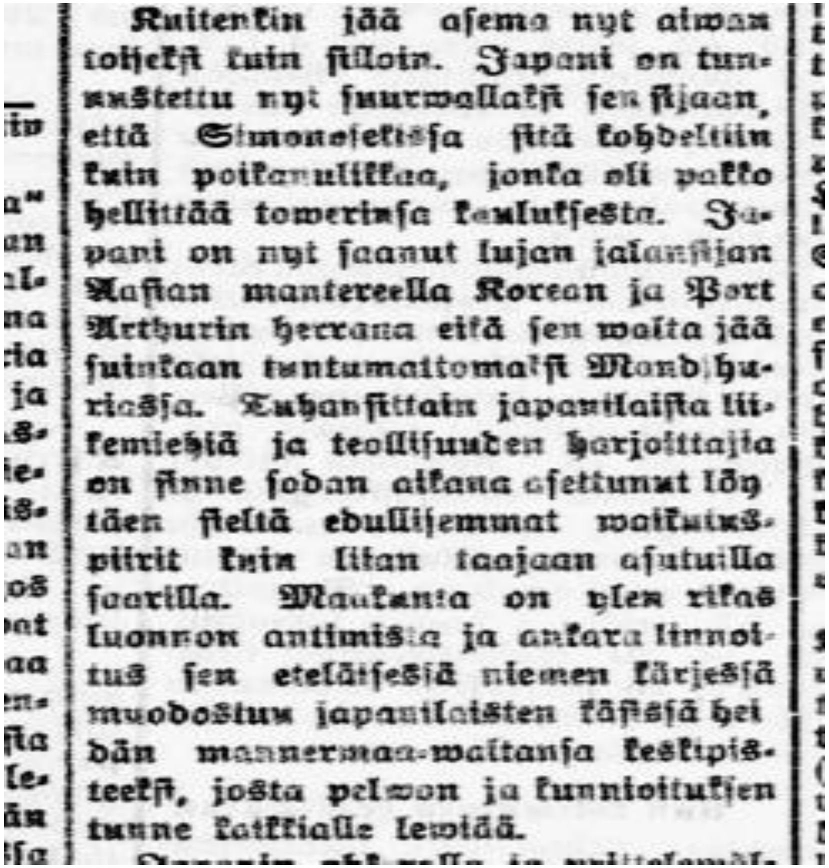
Quality: general problems of digitization in old newspaper collections

- Old newspaper collections are hard for digitization (paper and print quality, wear&tear, typeface etc.)
- In the output of the Optical Character Recognition (OCR) process errors are common especially when the texts are printed in the Fraktur (blackletter) typeface.
- E.g. Newspaper collection of British Library has a mean word correctness rate of **about 78 %** (19th Century Newspaper Project, http://www.dlib.org/dlib/july09/munoz/07munoz.html)
- Errors lower usability of corpora both from the point of view of human users as well as regarding potential text mining applications.

Quality issues of the OCRed Digi data

- Scanning of the contents of Digi was started in the early 2000s
- OCR software etc. was not on the same level then as now
- Quality of the originals has varied quite a lot
- Mostly Fraktur typeface used till the end of 19th century and early 20th century in Finland:
- → quality of the collection is varying due to OCR errors

- Aamulehti 17.9. 1905

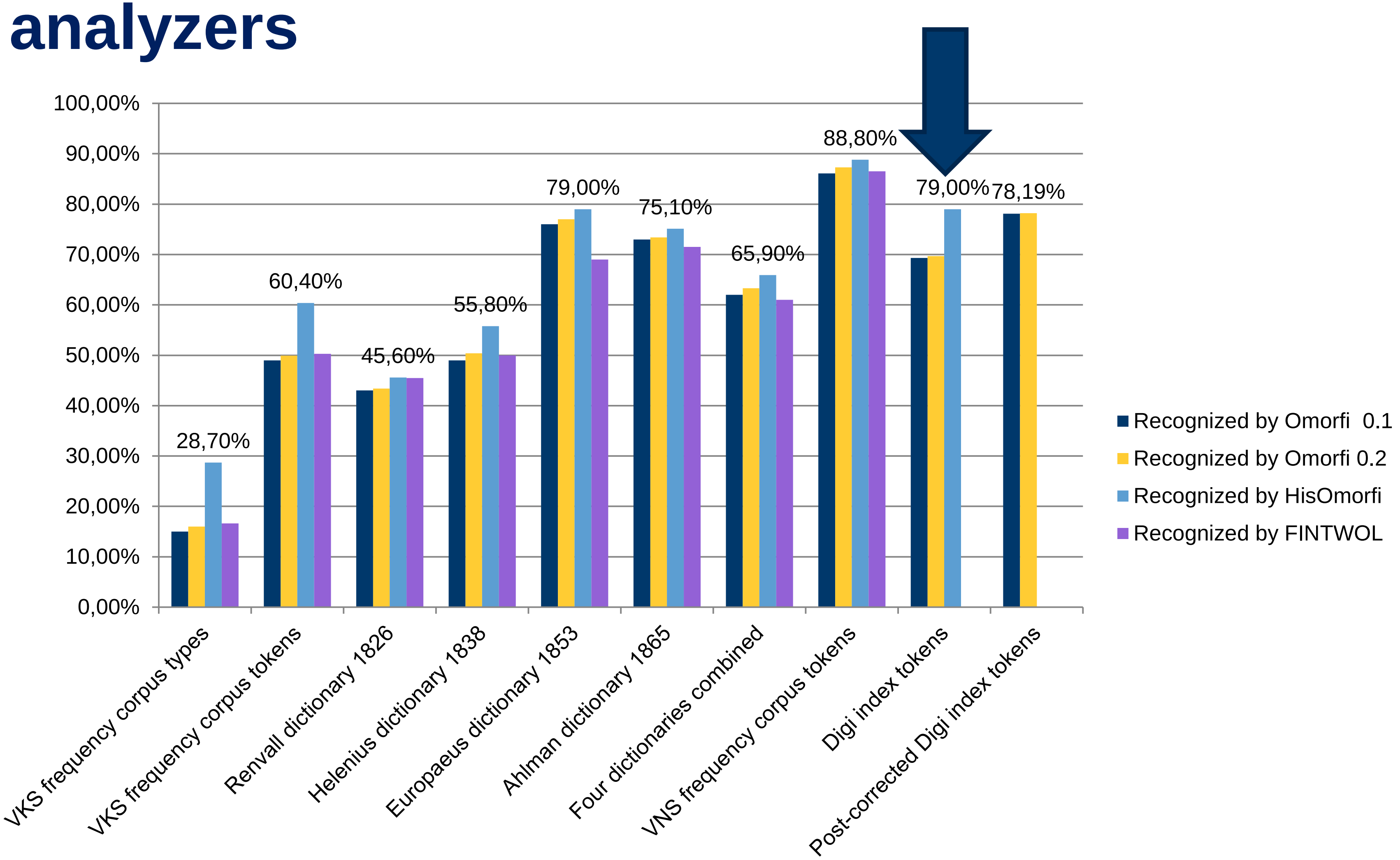


Kuitenkin jää aſema nyt aiwan toiſekſi kuin ſilloin. Japani on tunnustettu nyt ſuurwallakſi ſen ſijaan, että Simonoſekiſſa ſitä kohdeltiin kuin poikanulikkaa, jonka oli pakko hellittää tawerinſa kaulukſesta. Japani on nyt ſaanut lujan jalanſijan Aaſian mantereella Korean ja Port Arthurin herrana eikä ſen walta jää ſuinkaan tuntumattomakſi Mandſhuriasſa. Tuhanſittain japanilaiſia liikemiehiä ja teolliſuuden harjoittajia on ſinne ſodan aikana aſettunut löytäen ſieltä edulliſemmat waikutuspiirit kuin liian taajaan aſutuilla ſaarilla. Maakunta on yleen rikas luonnon antimista ja ankara linnoitus ſen eteläiſesſä niemen kärjesſä muodostuu japanilaisten käſisſä heidän mannermaa-waltanſa keskipisteekſi, josta pelwon ja kunnioituksen tunne kaikkialle lewiää.

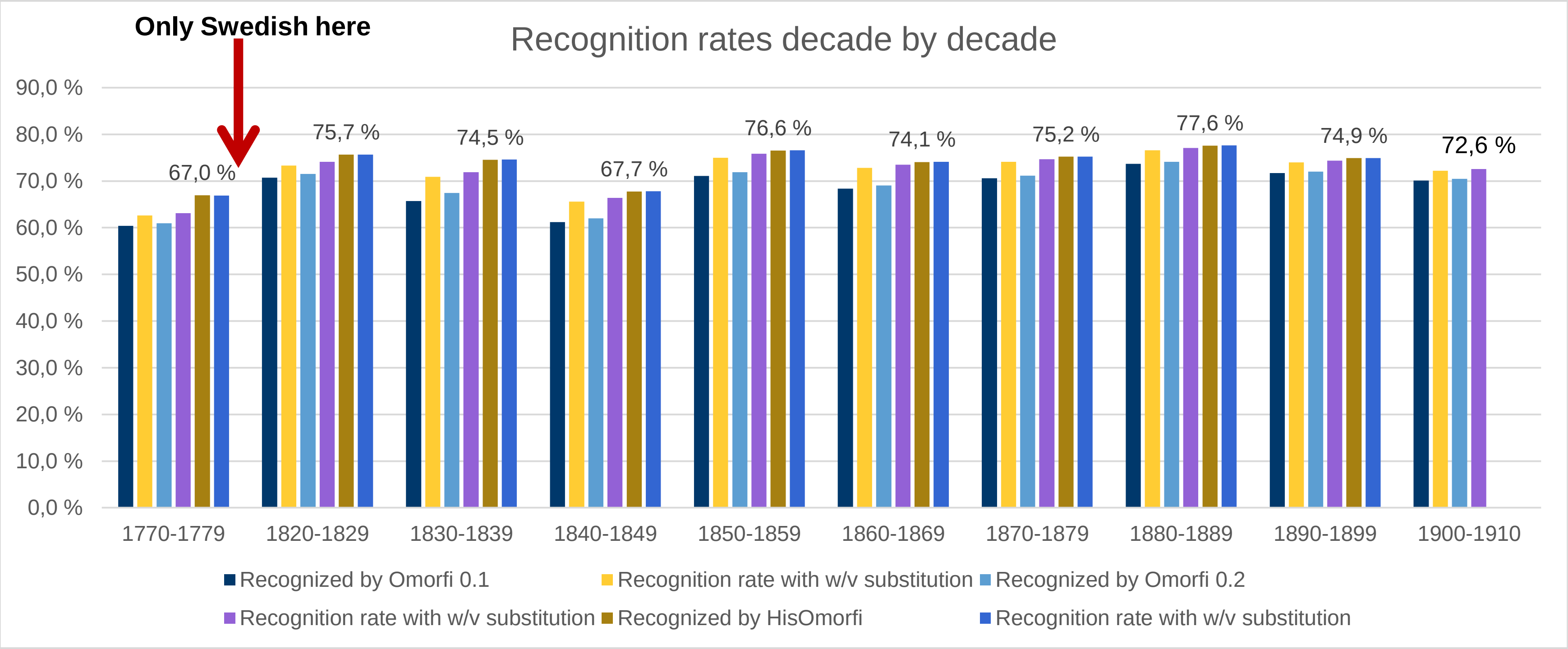
Quality assessment

- The Finnish part of the collection 1771-1910 has about 2.40 B words!
- Huge amount - what can you do? (evaluation based on samples or try to assess all?)
- We took four modern Finnish morphological analyzers (FINTWOL and 3 versions of Omorfi) and ran all the word data through them (word data taken out of the web collection's index)
- For comparison we used hand edited comparable word data available from the Institute for the Languages of Finland (VKS_korpus, VNS_korpus and four dictionaries from the 19th century: data is smallish, but the only available: largest c. 4.86 M, dictionaries 25K-90 K)

Word recognition rates: 4 morphologigal analyzers



Recognition of newspaper data decade by decade (word tokens)



Caveat – recognition ≠ correctness
misrecognition ≠ uncorrectness

- *mli* mli Num Roman Nom Sg → probably an OCR error
- *huu* huu Part

  tain tai N Gen Sg → wrong division to two parts based on hyphenation, should be *huutain* which is unrecognizable, although it is a correct form in 19th century Finnish
- *Hei* He Pron Nom Pl → should be *heidan,* unrecognizable *(heidän* would be correct*)*

  dan +?
- *Samoinkuin* +? → not recognized because written as a compound, OK otherwise
- *Ylöskannetaan* +? → not recognized because written as a compound, OK otherwise

Ways to improve the quality

1) Re-OCRing → ABBYY FineReader's Fraktur licensing policy/pricing impossible with new AbbyyFinereader, we are moving to open source Tesseract OCR.

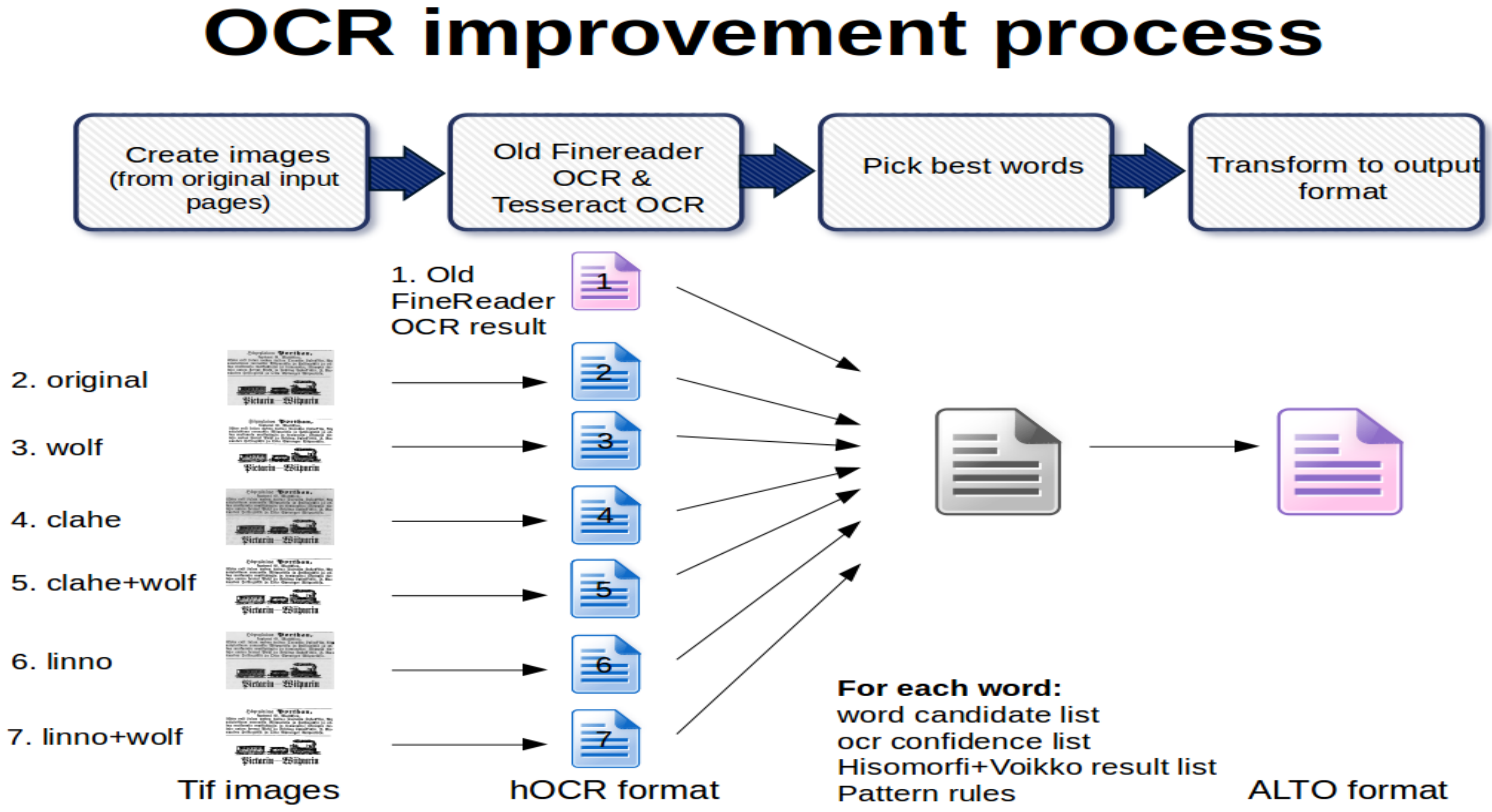
2) Post-correction with software: FIN-CLARIN: about 9-X % units recognition rate improvements so far (Omorfi 0.2, HisOmorfi)

3) Crowdsourcing (i.e. human correction): not feasible due to large amount of data

The new OCR process

- A new Fraktur font model was taught for Tesseract (using an existing German Fraktur as basis)
- A pipeline of image improving process was established
- OCR with Tesseract 3.04.01 (+ existing Abbyy FineReader v.7/8 results)
- Choosing the most probable word candidate from several suggestions
- Transformation of the output to ALTO XML

The new OCR process



OCR improvement process

Create images (from original input pages)

Old Finereader OCR & Tesseract OCR

Pick best words

Transform to output format

1. Old FineReader OCR result

2. original

3. wolf

4. clahe

5. clahe+wolf

6. linno

7. linno+wolf

Tif images

hOCR format

For each word:
word candidate list
ocr confidence list
Hisomorfi+Voikko result list
Pattern rules

ALTO format

Results - a 500 K ground truth word list – precision and recall results (latest)

Words without errors 374299
Words with errors 131008
Errorless not corrected 366043
Sum (lines 1 and 2) = 505307
True pos 99071
False negs 31937
False positives = 8256

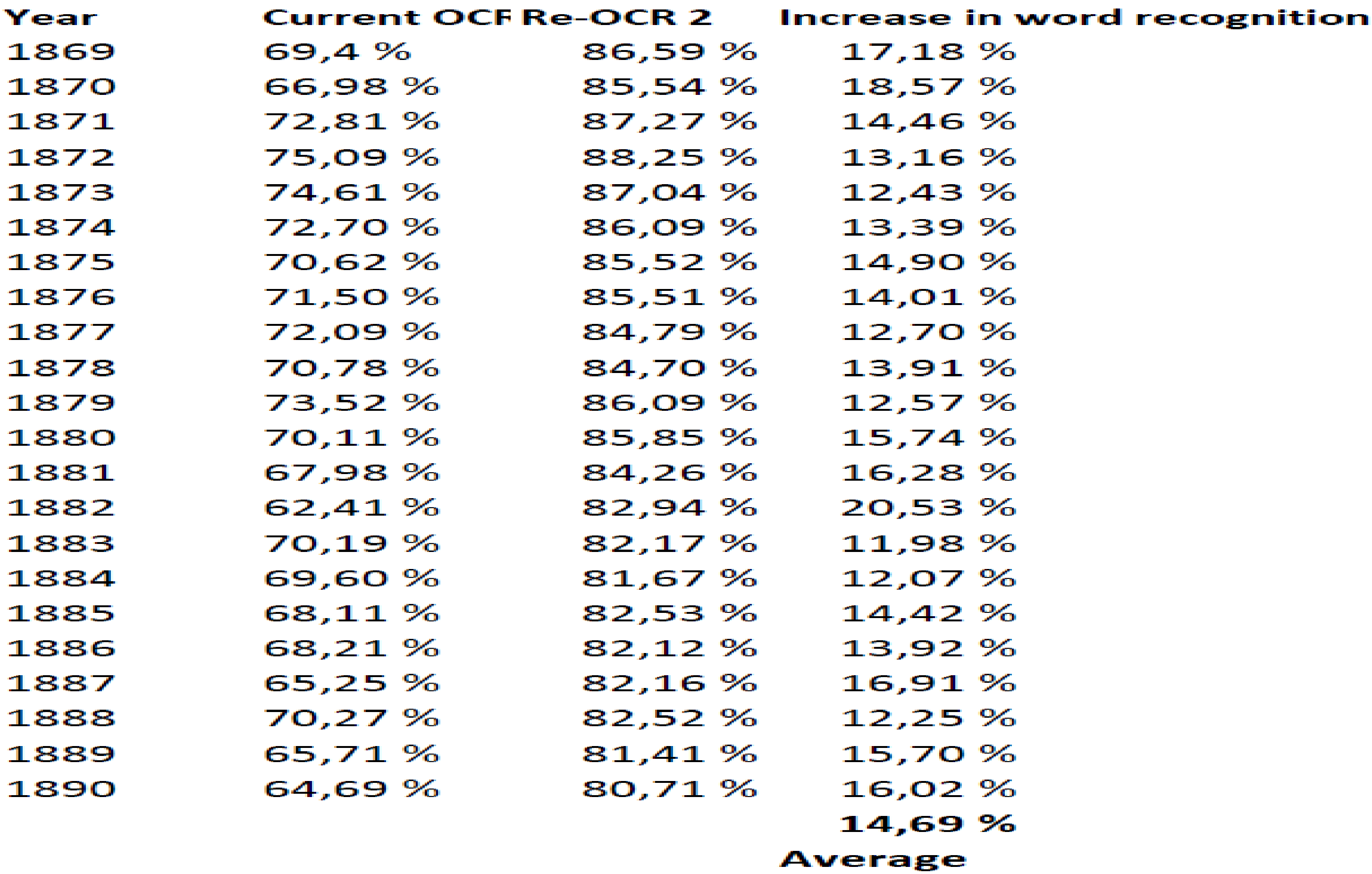
Recall   = 0.76
Precision = 0.92
F measure= 0.83

Correction rate = 0.69

Results of 500K GT: character and word error rates, character accuracy

	re-OCR of NLF	Current OCR
CER	2,05	6,47
WER	6,56	25,30
WER (order independent)	5,51	23,41
CAR	97,64	92,62

Uusi Suometar 1869-1890: average gain 14,7 % units in word recognizability



Year	Current OCF	Re-OCR 2	Increase in word recognition
1869	69,4 %	86,59 %	17,18 %
1870	66,98 %	85,54 %	18,57 %
1871	72,81 %	87,27 %	14,46 %
1872	75,09 %	88,25 %	13,16 %
1873	74,61 %	87,04 %	12,43 %
1874	72,70 %	86,09 %	13,39 %
1875	70,62 %	85,52 %	14,90 %
1876	71,50 %	85,51 %	14,01 %
1877	72,09 %	84,79 %	12,70 %
1878	70,78 %	84,70 %	13,91 %
1879	73,52 %	86,09 %	12,57 %
1880	70,11 %	85,85 %	15,74 %
1881	67,98 %	84,26 %	16,28 %
1882	62,41 %	82,94 %	20,53 %
1883	70,19 %	82,17 %	11,98 %
1884	69,60 %	81,67 %	12,07 %
1885	68,11 %	82,53 %	14,42 %
1886	68,21 %	82,12 %	13,92 %
1887	65,25 %	82,16 %	16,91 %
1888	70,27 %	82,52 %	12,25 %
1889	65,71 %	81,41 %	15,70 %
1890	64,69 %	80,71 %	16,02 %
			14,69 %
			Average

NER – Named Entity Recognition

What?

- Names of persons, locations, organisations are important factual data in texts
- They can be recognized automatically to a reasonable extent (70-90+ %)
- They can be used in information extraction out of the data
- Names of persons and locations are used heavily as keywords in text searches of on-line databases. Many times even 80 % of keywords are names.

An example of text with names

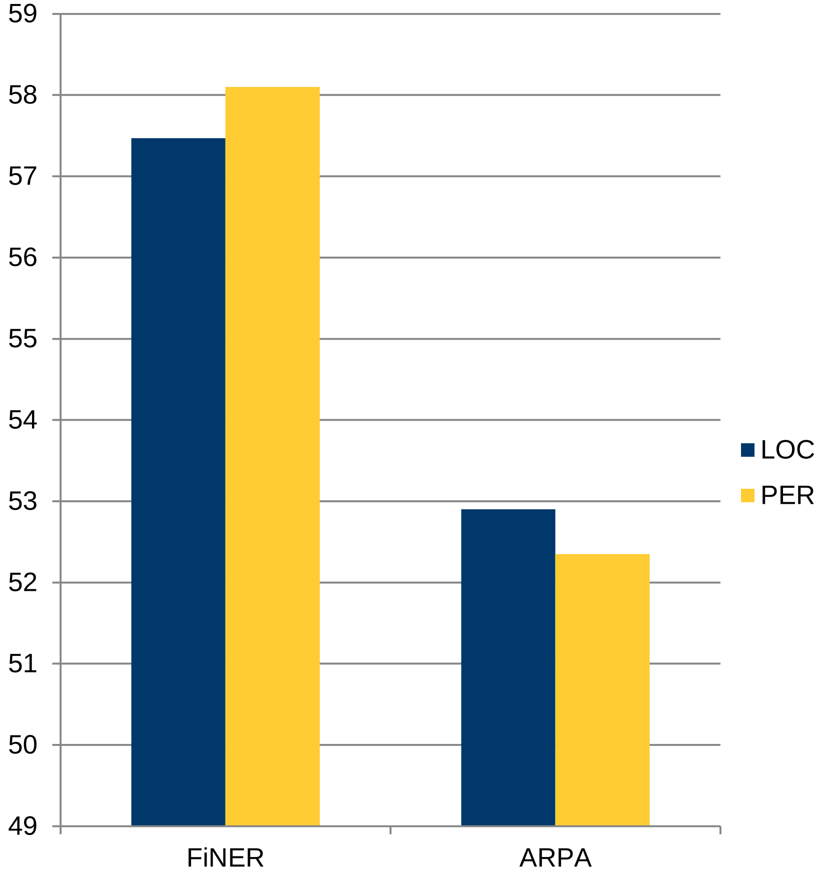
Edellä sanotun tarpeen maatimana ja kun ei täällä ole mallikelpoista maanmiljelystaloa, jonka puuttce» poistamiseksi toiwomme että usei»mai»ittu koulu perustettaisiin hra O. L, Vumeruksen omistmnalle Kupiala» tilalle Rantasalmen pitäjään, eikä lärmikylään Joroisissa, koska lärmikylä jo kaumcmman aikaa nykyisen omistajansa jalomielisestä ja paljon uhraaivaiscsta toimesta un ollut, ja luonnollisesti tulisi edelleen olemaan PohjoiS-samolaisille monien kokeittcnsa ja mallikelpoisen maanmiljcll,ksen

- 6 names, 2 of them spelled right

First NER trials with the OCRed newspaper data

- There are tools for doing NER for Finnish, but they are usually for modern language
- We evaluated first 5 different modern tools with an evaluation collection of about 76 000 words (manually tagged by T. Ruokolainen)
- Results were not very

good: F scores 0.52-0.58 (scale 0.0-1.0)



A new start with Stanford NER package

- We created a new training and evaluation collection of ca. 450 000 words
- Used trainable Stanford NER machine learning package to learn names from about 380 000 words that had been marked manually and semi-automatically
- Results quite good

Results of Stanford NER

- Ideal results with manually corrected data

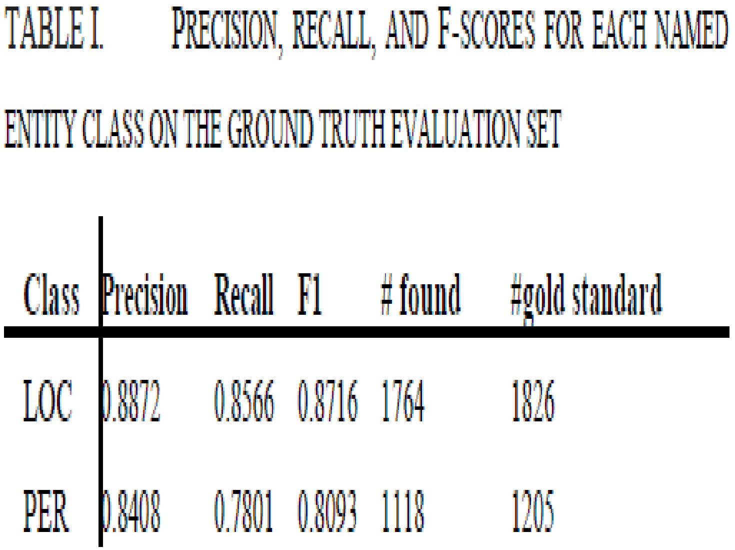


TABLE I. PRECISION, RECALL, AND F-SCORES FOR EACH NAMED ENTITY CLASS ON THE GROUND TRUTH EVALUATION SET

Class	Precision	Recall	F1	# found	#gold standard
LOC	0.8872	0.8566	0.8716	1764	1826
PER	0.8408	0.7801	0.8093	1118	1205

- Realistic results with achievable OCR: a 9-10 % unit drop

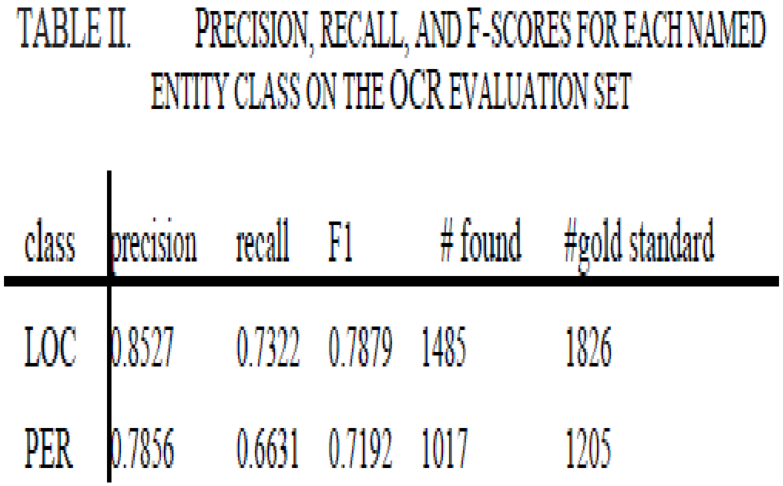
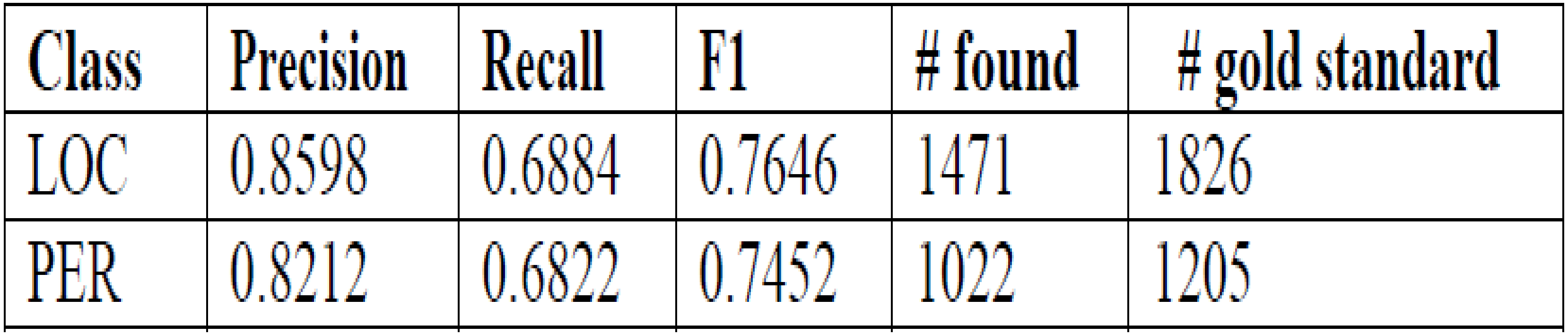


TABLE II. PRECISION, RECALL, AND F-SCORES FOR EACH NAMED ENTITY CLASS ON THE OCR EVALUATION SET

class	precision	recall	F1	# found	#gold standard
LOC	0.8527	0.7322	0.7879	1485	1826
PER	0.7856	0.6631	0.7192	1017	1205

Results of a LSTM-CRF (Lample et al. 2016)

- LSTM (long short-term memory), recurrent neural network model: state-of-the-art, receives very similar results with Stanford



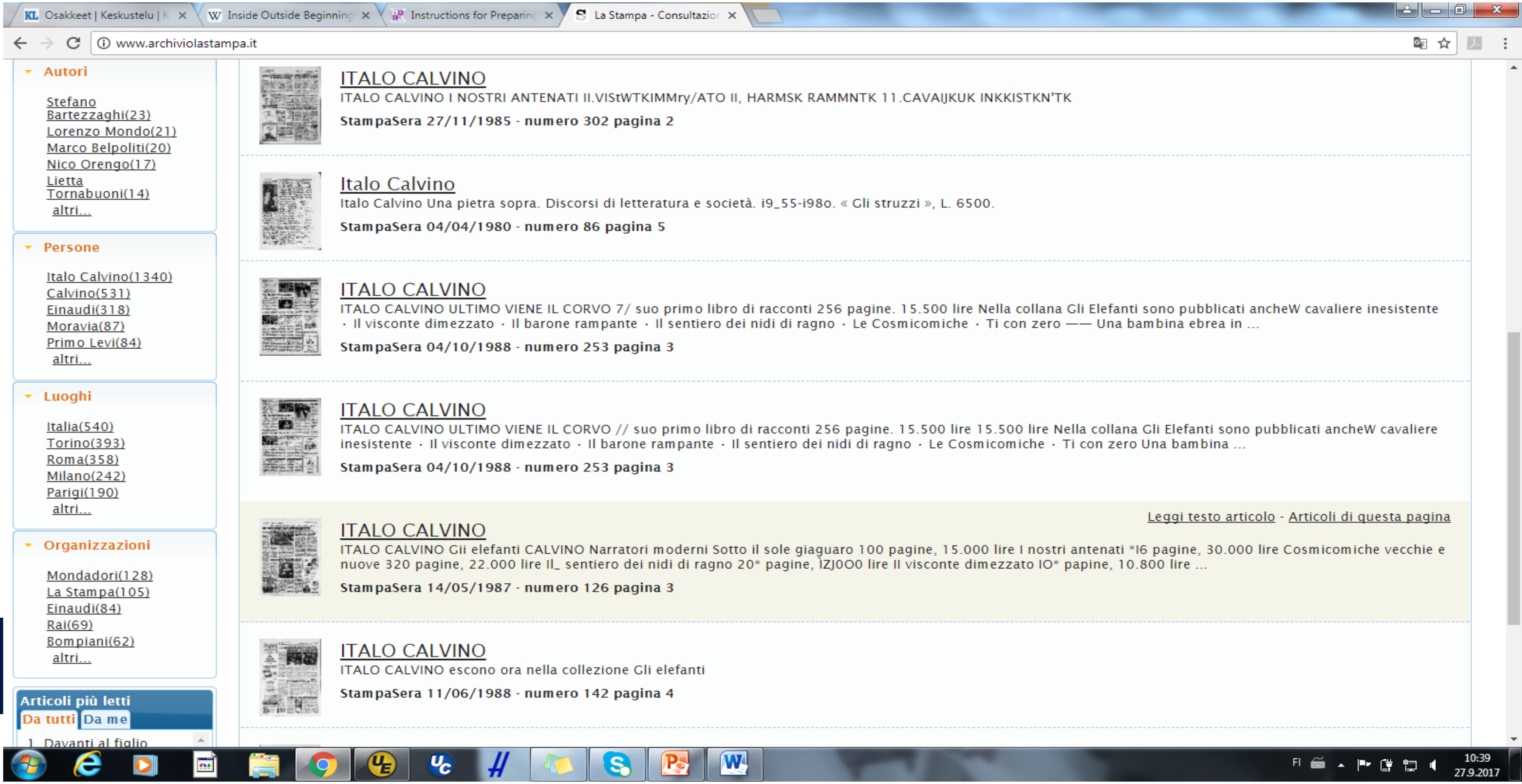
Class	Precision	Recall	F1	# found	# gold standard
LOC	0.8598	0.6884	0.7646	1471	1826
PER	0.8212	0.6822	0.7452	1022	1205

What to do with NER in a historical newspaper collection?

- NER is a tool that needs to be used for some purpose
- Two basic uses:
    - Enhancement of browsing in large collections
    - Linking names to knowledge sources (DBPedia, Wikipedia, bibliographical sources, geographical sources etc.)

An example of browsing enhancement

- La Stampa 1867-2005 http://www.archiviolastampa.it/ww.archiviolastampa.it



Autori

Stefano Bartezzaghi(23)
Lorenzo Mondo(21)
Marco Belpoliti(20)
Nico Orengo(17)
Lietta Tornabuoni(14)
altri...

Persone

Italo Calvino(1340)
Calvino(531)
Einaudi(318)
Moravia(87)
Primo Levi(84)
altri...

Luoghi

Italia(540)
Torino(393)
Roma(358)
Milano(242)
Parigi(190)
altri...

Organizzazioni

Mondadori(128)
La Stampa(105)
Einaudi(84)
Rai(69)
Bompiani(62)
altri...

Articoli più letti

Da tutti | Da me

1 Davanti al figlio

ITALO CALVINO

ITALO CALVINO I NOSTRI ANTENATI II.VIStWTKIMMry/ATO II, HARMSK RAMMNTK 11.CAVAIJKUK INKKISTKN'TK

StampaSera 27/11/1985 - numero 302 pagina 2

Italo Calvino

Italo Calvino Una pietra sopra. Discorsi di letteratura e società. i9_55-i98o. « Gli struzzi », L. 6500.

StampaSera 04/04/1980 - numero 86 pagina 5

ITALO CALVINO

ITALO CALVINO ULTIMO VIENE IL CORVO 7/ suo primo libro di racconti 256 pagine. 15.500 lire Nella collana Gli Elefanti sono pubblicati ancheW cavaliere inesistente · Il visconte dimezzato · Il barone rampante · Il sentiero dei nidi di ragno · Le Cosmicomiche · Ti con zero —— Una bambina ebrea in ...

StampaSera 04/10/1988 - numero 253 pagina 3

ITALO CALVINO

ITALO CALVINO ULTIMO VIENE IL CORVO // suo primo libro di racconti 256 pagine. 15.500 lire 15.500 lire Nella collana Gli Elefanti sono pubblicati ancheW cavaliere inesistente · Il visconte dimezzato · Il barone rampante · Il sentiero dei nidi di ragno · Le Cosmicomiche · Ti con zero Una bambina ...

StampaSera 04/10/1988 - numero 253 pagina 3

Leggi testo articolo - Articoli di questa pagina

ITALO CALVINO

ITALO CALVINO Gii elefanti CALVINO Narratori moderni Sotto il sole giaguaro 100 pagine, 15.000 lire I nostri antenati *I6 pagine, 30.000 lire Cosmicomiche vecchie e nuove 320 pagine, 22.000 lire Il_ sentiero dei nidi di ragno 20* pagine, ÌZJ0O0 lire Il visconte dimezzato IO* papine, 10.800 lire ...

StampaSera 14/05/1987 - numero 126 pagina 3

ITALO CALVINO

ITALO CALVINO escono ora nella collezione Gli elefanti

StampaSera 11/06/1988 - numero 142 pagina 4

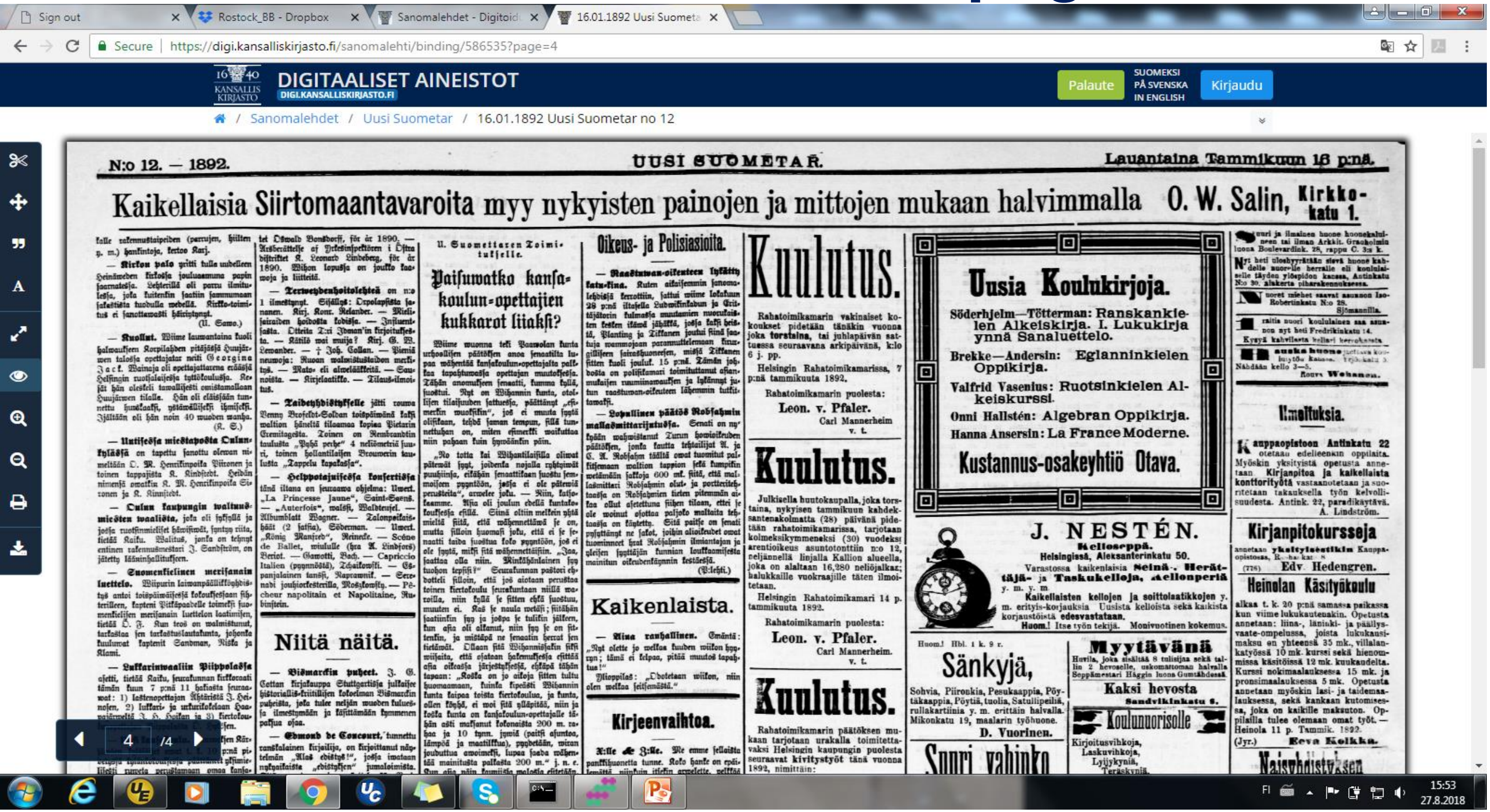
Layout analysis / article extraction

- Newspapers are usually digitised page by page (scanner takes a ”photo” of the page)
- Page images serve as the basic unit of whole processing
- → Trouble ahead!

- Page is not any kind of informational unit, only a typographical/printing unit

- Pages consist of different parts: text (news items, titles), pictures etc.

Uusi Suometar, 16.01.1892, page 4



N:o 12. — 1892. UUSI SUOMETAR. Lauantaina Tammikuun 16 p:nä.

Kaikellaisia Siirtomaantavaroita myy nykyisten painojen ja mittojen mukaan halvimmalla O. W. Salin, Kirkkokatu 1.

kalle rakennustaipeiden (parrujen, hiilten y. m.) hankintoja, kertoo Karj.

— **Kirkon palo** yritti tulla uhdelleen Heinäweden kirkoſſa jouluaamuna papin ſaarnateſſa. Lehterillä oli parru ilmituleſſa, joka kuitenkin ſaatiin ſammumaan ſakeistista tuodulla wedellä. Kirkko-toimitus ei ſanottawasti häiriytynyt. (U. Sawo.)

— **Kuollut.** Wiime lauwantaina kuoli halwaukſeen Korpilahden pitäjäſſä Huujärwen taloſſa opettajatar neiti Georgina Jack. Wainaja oli opettajattarena eräässä Helſingin ruotſalaiſeſſa tyttökouluſſa. Keſät hän oleskeli tawallisesti omistamallaan Huujärwen tilalle. Hän oli eläiſſään tunnettu hunskaaksi, ystäwälliseksi ihmiseksi. Ijältään oli hän noin 40 wuoden wanha. (K. S.)

— **Uutiseſſa miestaposta Oulunkylässä** on tapettu ſanottu olewan nimeltään O. M. Henrikinpoika Piironen ja toinen tappajista K. Kindſtedt. Heidän nimenſä owatkin K. M. Henrikinpoika Sironen ja K. Kinnſtedt.

— **Oulun kaupungin waltuusmiesten waalista,** joka oli ſyksyllä ja joſſa ruotſinmieliset häwisiwät, ſyntyy riita, tietää Kaiku. Walitus, jonka on tehnyt entinen rakennusmestari J. Sandſtröm, on jätetty lääninhallitukſeen.

— **Suomenkielinen meriſanain luettelo.** Wiipurin laiwanpäällikköyhdistys antoi toiſpäiwäiſeſſä kokoukſeſſaan ſihteerilleen, kapteeni Pitkäpaadelle toimeksi ſuomenkielisen meriſanain luettelon laatimiſen, tietää O. F. Kun teos on walmistunut, tarkastaa ſen tarkastuslautakunta, johonka kuuluwat kapteenit Sandman, Riska ja Klami.

— **Lukkarinwaaliin Piippolassa** asetti, tietää Kaiku, ſeurakunnan kirkkoraati tämän kuun 7 p:nä 11 hakiasta ſeuraawat: 1) lastenopettajan Ähtäristä J. Heinoſen, 2) lukkari- ja urkurikokelaan Haapajärweltä J. H. Hoikan ja 3) kiertokoulunopettajan Piippolasta O. [illegible]ſen.

— [illegible] p:nä pidetyssä [illegible] yksimielisesti ruweta perustamaan omaa kanſa-

tet Oswald Bonsdorff, för år 1890. — Årsberättelse af Yrkesinspektören i Östra distriktet K. Leonard Lindeberg, för år 1890. Wihon lopuſſa on joukko kaawoja ja liitteitä.

— **Terweydenhoitolehteä** on n:o 1 ilmestynyt. Sisällys: Orpolapſista ſananen. Kirj. Konr. Relander. — Mielisairaiden hoidosta kodiſſa. — Influentſasta. Otteita T:ri Idman'in kirjoitukſesta. — Kätilö wai muija? Kirj. G. W. Lewander. — † Joh. Collan. — Pieniä neuwoja: Ruoan walmistustaidon merkitys. — Mato- eli alweläkkeitä. — Saunoista. — Kirjelaatikko. — Tilaus-ilmoitus.

— **Taideyhdistykſelle** jätti rouwa Venny Brofeldt-Soldan toiſpäiwänä kaksi waltion häneltä tilaamaa kopiaa Pietarin Eremitagesta. Toinen on Rembrandtin taulusta „Pyhä perhe" 4 neliömetriä ſuuri, toinen hollantilaiſen Brouwerin taulusta „Tappelu kapakaſſa".

— **Helppotajuiſeſſa konſertiſſa** tänä iltana on ſeuraawa ohjelma: Uwert. „La Princesse Jaune", Saint-Saens. — „Auterfois", walssi, Waldteufel. — Albumblatt Wagner. — Talonpoikaishäät (2 ſatſia), Söderman. — Uwert. „König Manfred", Reinecke. — Scène de Ballet, wiululle (hra A. Lindfors) Beriot. — Gawotti, Bach. — Capriccio Italien (pyynnöstä), Tchaikowski. — Espanjalainen tanssi, Naprawnik. — Serenadi jousiorkesterilla, Moszkowsky. — Pêcheur napolitain et Napolitaine, Rubinſtein.

Niitä näitä.

— **Bismarckin puheet.** J. G. Cottan kirjakauppa Stuttgartiſſa julkaiſee historiallis-kriitillisen kokoelman Bismarckin puheista, joka tulee neljän wuoden kulueſſa ilmestymään ja käſittämään kymmenen pakſua oſaa.

— **Edmond de Concourt,** tunnettu ranskalainen kirjailija, on kirjoittanut näytelmän „Alas edistys!", joſſa iwataan nykyaikaiſte „edistyksen" jumaloimista.

U. Suomettaren Toimitukſelle.

Paiſuwatko kanſakoulun-opettajien kukkarot liiaksi?

Wiime wuonna teki Paawolan kunta urhoollisen päätökſen anoa ſenaatilta lupaa wähentää kanſakoulun-opettajalta palkkaa tapahtuwaſſa opettajan muutokſeſſa. Tähän anomukſeen ſenaatti, kumma kyllä, ſuostui. Nyt on Wihannin kunta, otollisen tilaiſuuden ſattueſſa, päättänyt „esimerkin wuoksikin", jos ei muuta ſyytä olisikaan, tehdä ſaman tempun, sillä tunnettuhan on, miten esimerkki waikuttaa niin pahaan kuin hywäänkin päin.

„No totta kai Wihantilaisilla oliwat pätewät ſyyt, joidenka nojalla ryhtyiwät puuhiinſa, eikähän ſenaattikaan ſuostu ſemmoiſeen pyyntöön, joſſa ei ole pätewiä perusteita", arwelee joku. — Niin, katſokaamme. Asia oli joulun edellä kuntakokoukſeſſa esillä. Siinä oltiin melkein yhtä mieltä siitä, että wähennettäwä ſe on, mutta silloin huomasi joku, että ei ſe ſenaatti taida ſuostua koko pyyntöön, jos ei ole ſyytä, miksi sitä wähennettäisiin. „Jaa, ſaattaa olla niin. Minkähänlainen ſyy tuohon keksisi?" Seurakunnan pastori ehdotteli silloin, että jos aiotaan perustaa toinen kiertokoulu ſeurakuntaan niillä waroilla, niin kyllä ſe ſitten ehkä ſuostuu, muuten ei. Kas ſe naula wetäsi; siitähän ſaatiinkin ſyy ja jospa ſe tulikin jälkeen, kun asia oli alkanut, niin ſyy ſe on sittenkin, ja mistäpä ne ſenaatin herrat ſen tietäwät. Ollaan sitä Wihanniſſakin sikſi wiiſaita, että osataan hakemukſeſſa esittää asia oikeaſſa järjestykſeſſä, ehkäpä tähän tapaan: „Koska on jo aikoja ſitten tultu huomaamaan, kuinka kipeästi Wihannin kunta kaipaa toista kiertokoulua, ja kunta, ollen köyhä, ei woi sitä ylläpitää, niin ja koska kunta on kanſakoulun-opettajalle tähän asti makſanut kokonaista 200 m. rahaa ja 10 tynn. jywiä (paitsi asuntoa, lämpöä ja maatilkkua), pyydetään, wiran jouduttua awoimeksi, lupaa ſaada wähentää mainitusta palkasta 200 m." j. n. e. Kun asia näin [illegible]

Oikeus- ja Polisiasioita.

— **Raastuwan-oikeuteen lykätty katu-kina.** Kuten aikaiſemmin ſanomalehdiſſä kerrottiin, ſattui wiime lokakuun 28 p:nä illaſella Ludwikinkadun ja Erottajatorin kulmaſſa muutamien nuorukaisten keskẽn ikäwä jähäkkä, joſſa kaksi heistä, Planting ja Tikkanen joutui siinä ſaatuja wammojaan parannuttelemaan kirurgilliseen ſairashuoneeſen, miſſä Tikkanen ſitten kuoli jouluk. 15 p:nä. Tämän johdosta on polisikamari toimituttanut asianmukaiſen ruumiinawaukſen ja lykännyt juttun raastuwan-oikeuteen lähemmin tutkittawaksi.

— **Lopullinen päätös Robſahmin mallasmittarijutuſſa.** Senati on nykyään wahwistanut Turun howioikeuden päätökſen, jonka kautta tehtailijat A. ja C. A. Robſahm täältä owat tuomitut palkitſemaan waltion tappion ſekä kumpikin wetämään ſakkoja 600 mk. siitä, että mallasmittari Robſahmin olut- ja portteritehtaaſſa on Robſahmien tieten pitemmän aikaa ollut asetettuna siihen tilaan, ettei ſe ole woinut osottaa paljoko maltaita tehtaaſſa on käytetty. Sitä paitsi on ſenati pysyttänyt ne ſakot, joihin alioikeudet owat tuominneet hrat Robſahmin ilmiantajan ja yleiſen ſyyttäjän kunnian loukkaamiſesta mainitun oikeudenkäynnin kestäeſſä. (P:lehti.)

Kaikenlaista.

— **Aina rauhallinen.** Emäntä: „Nyt olette jo welkaa kuuden wiikon hyyryn; tämä ei kelpaa, pitää muutos tapahtua!"

Ylioppilas: „Odotetaan wiikon, niin olen welkaa ſeitſemästä."

Kirjeenvaihtoa.

X:lle & Z:lle. Me emme ſellaista pankkihuonetta tunne. Koko hanke on epäilemättä niinkuin itsekin arwelette, pelkkää

Kuulutus.

Rahatoimikamarin vakinaiset kokoukset pidetään tänäkin vuonna joka **torstaina**, tai juhlapäivän sattuessa seuraavana arkipäivänä, k:lo 6 j. pp.

Helsingin Rahatoimikamarissa, 7 p:nä tammikuuta 1892.

Rahatoimikamarin puolesta:
Leon. v. Pfaler.
Carl Mannerheim
v. t.

Kuulutus.

Julkisella huutokaupalla, joka torstaina, nykyisen tammikuun kahdeksantenakolmatta (28) päivänä pidetään rahatoimikamarissa, tarjotaan kolmeksikymmeneksi (30) vuodeksi arentioikeus asuntotonttiin n:o 12, neljännellä linjalla Kallion alueella, joka on alaltaan 16,280 neliöjalkaa; halukkaille vuokraajille täten ilmoitetaan.

Helsingin Rahatoimikamari 14 p. tammikuuta 1892.

Rahatoimikamarin puolesta:
Leon. v. Pfaler.
Carl Mannerheim.
v. t.

Kuulutus.

Rahatoimikamarin päätöksen mukaan tarjotaan urakalla toimitettavaksi Helsingin kaupungin puolesta seuraavat **kivitystyöt** tänä vuonna 1892, nimittäin:

Uusia Koulukirjoja.

Söderhjelm—Tötterman: Ranskankielen Alkeiskirja. I. Lukukirja ynnä Sanaluettelo.

Brekke—Andersin: Eglanninkielen Oppikirja.

Valfrid Vasenius: Ruotsinkielen Alkeiskurssi.

Onni Hallstén: Algebran Oppikirja.

Hanna Ansersin: La France Moderne.

Kustannus-osakeyhtiö Otava.

J. NESTÉN.

Kelloseppä.
Helsingissä, Aleksanterinkatu 50.

Varastossa kaikenlaisia **Seinä-, Herättäjä- ja Taskukelloja, Kellonperiä** y. m. y. m.

Kaikellaisten kellojen ja soittolaatikkojen y. m. erityis-korjauksia. Uusista kelloista sekä kaikista korjaustöistä **edesvastataan.**

Huom.! Itse työn tekijä. Monivuotinen kokemus.

Huom.! Hbl. 1 k. 9 r.

Sänkyjä,

Sohvia, Piironkia, Pesukaappia, Pöytäkaappia, Pöytiä, tuolia, Satulipeiliä, rullakartiinia y. m. erittäin halvalla. Mikonkatu 19, maalarin työhuone.

D. Vuorinen.

Suuri vahinko

Myytävänä

Huvila, joka sisältää 8 tulisijaa sekä tallin 2 hevoselle, uskomattoman halvalla Seppämestari Häggin luona Gumtähdessä.

Kaksi hevosta
Sandvikinkatu 6.

Koulunuorisolle

Kirjoitusvihkoja,
Laskuvihkoja,
Lyijykyniä,
Teräskyniä,

Suuri ja ilmainen huone huonekaluineen tai ilman Arkkit. Graaholmin luona Boulevardink. 28, rappu C. 3:s k.

Nyt heti uloshyyrätään siewä huone kahdelle nuorelle herralle eli koululaiselle täyden ylöspidon kanssa, Antinkatu N:o 30. alakerta piharakennuksessa.

Nuoret miehet saavat asunnon Iso-Robertinkatu N:o 28.
Sjömannilla.

[illegible] raitis nuori koululainen saa asunnon nyt heti Fredrikinkatu 14.
Kysyä kahvilasta kellari kerroksesta.

Hauska huone [illegible] kouluytön kanssa. [illegible]katu 5.
Nähdään kello 3—5.
Rouva **Wehanen.**

Ilmoituksia.

Kauppaopistoon Antinkatu 22 otetaan edelleenkin oppilaita. Myöskin yksityistä opetusta annetaan. **Kirjanpitoa ja kaikellaista konttorityötä** vastaanotetaan ja suoritetaan takauksella työn kelvollisuudesta. Antink. 22, paradikäytävä.
A. Lindström.

Kirjanpitokursseja

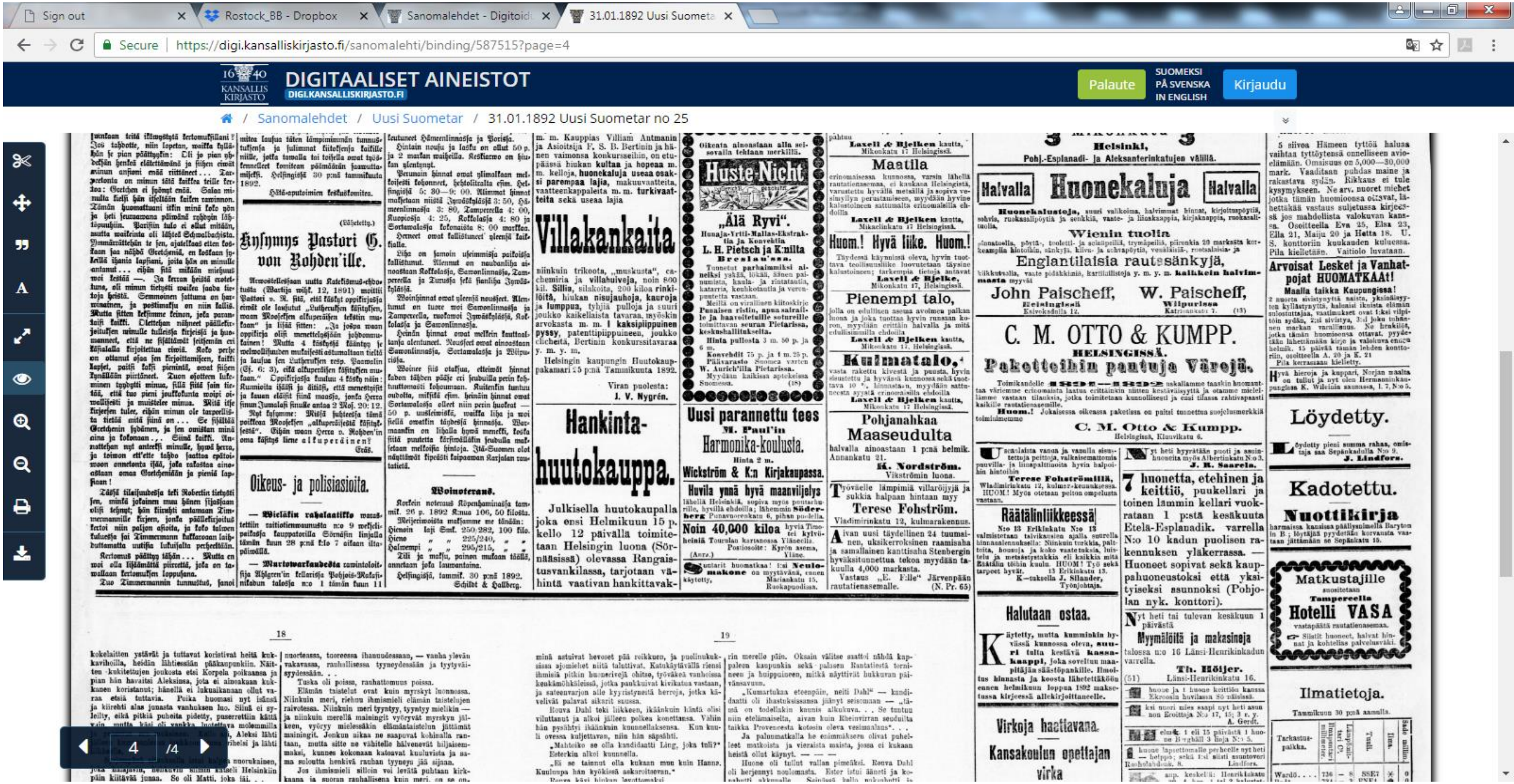
annetaan **yksityisestikin** Kauppaopistossa, R[illegible]katu 8
(776) **Edv. Hedengren.**

Heinolan Käsityökoulu

alkaa t. k. 20 p:nä samassa paikassa kun viime lukukautenakin. Opetusta annetaan: liina-, läninki- ja päällysvaate-ompelussa, joista lukukausimaksu on yhteensä 35 mk., villalankatyössä 10 mk. kurssi sekä hienommissa käsitöissä 12 mk. kuukaudelta. Kurssi nokimaalauksessa 15 mk. ja pronsimaalauksessa 5 mk. Opetusta annetaan myöskin lasi- ja taidemaalauksessa, sekä kankaan kutomisessa, joka on kaikille maksuton. Oppilailla tulee olemaan omat työt. — Heinola 11 p. Tammik. 1892.
(Jyr.) **Eeva Kolkka.**

Naisyhdistyksen

Uusi Suometar 31.01.1892, page 4



(Lähetetty.)

Kysymys Pastori G. von Rohden'ille.

Oikeus- ja polisiasioita.

Villakankaita

Hankinta-huutokauppa.

Huste-Nicht

„Älä Ryvi".

Uusi parannettu tees

Harmonika-koulusta.

Wickström & K:n Kirjakaupassa.

Huvila ynnä hyvä maanviljelys

Noin 40,000 kiloa

Maatila

Huom! Hyvä liike. Huom!

Pienempi talo,

Kulmatalo,

Pohjanakkaa Maaseudulta

Terese Fohström.

Räätälinliikkeessä

Halutaan ostaa.

Virkoja haettavana.

Kansakoulun opettajan virka

3 Helsinki, 3

Pohj.-Esplanadi- ja Aleksanterinkatujen välillä.

Halvalla Huonekaluja Halvalla

Wienin tuolia

Englantilaisia rautasänkyjä,

John Paischeff, W. Paischeff,

C. M. OTTO & KUMPP.

HELSINGISSÄ.

Paketteihin pantuja Värejä.

C. M. Otto & Kumpp.

7 huonetta, etehinen ja keittiö, puukellari ja toinen lammin kellari vuokrataan 1 p:stä kesäkuuta Etelä-Esplanadik. varrella N:o 10 kadun puolisen rakennuksen yläkerrassa. — Huoneet sopivat sekä kauppahuoneustoksi että yksityiseksi asunnoksi (Pohjolan nyk. konttori).

Nyymälöitä ja makasineja

Arvoisat Lesket ja Vanhatpojat HUOMATKAA!!

Löydetty.

Kadotettu.

Nuottikirja

Matkustajille

Hotelli VASA

Ilmatietoja.

18

19

Problems

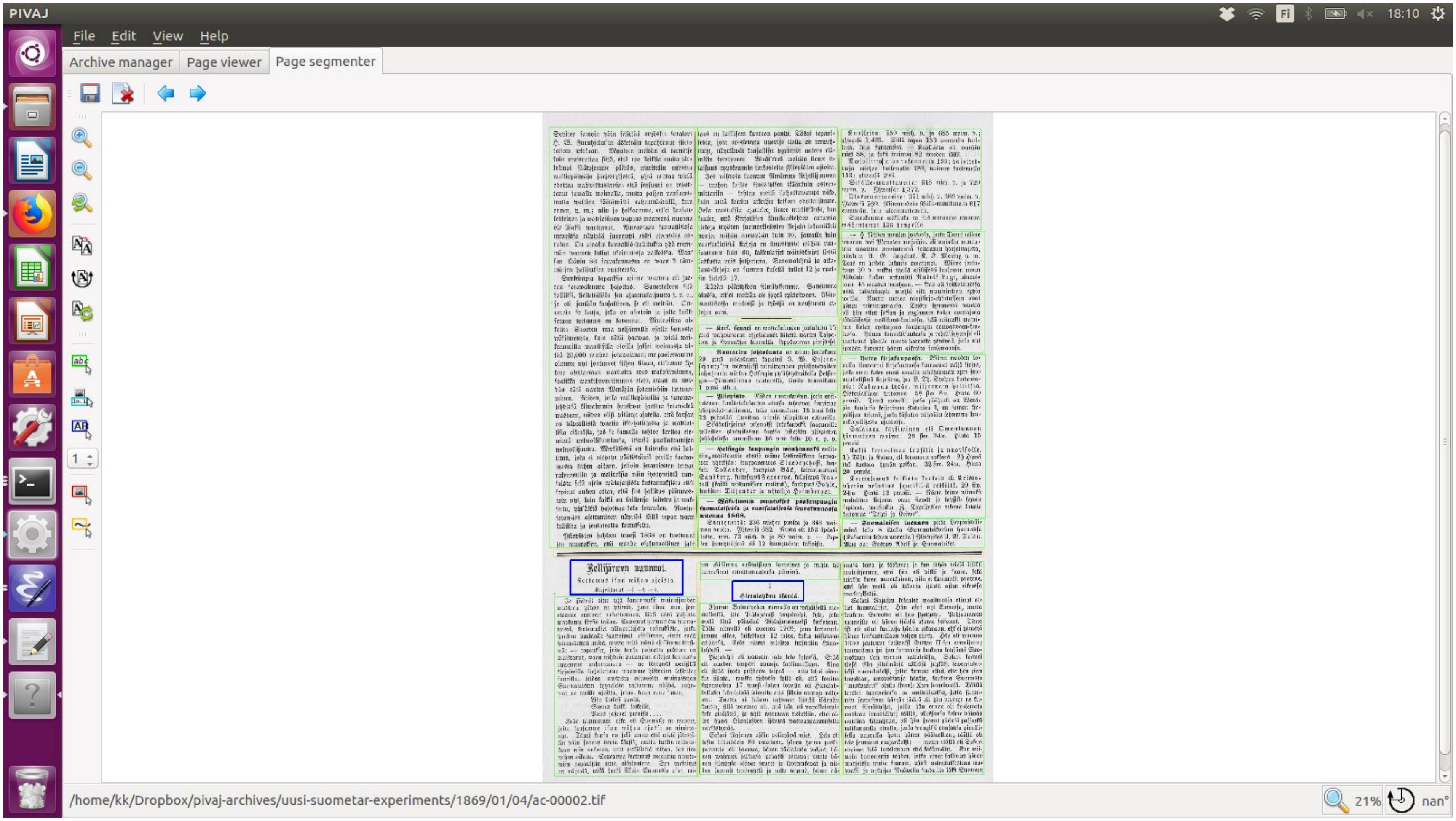
- Number of columns may keep changing even in a single number (and definitely every few years)
- Layout varies: advertisements and other pictures keep popping up in the middle of the text
- Different style of column structure in different parts of page
- ….

Currently best layout analysis software get about 85+% of the layout right in complex layouts/historical data (ICDAR competition from several years)

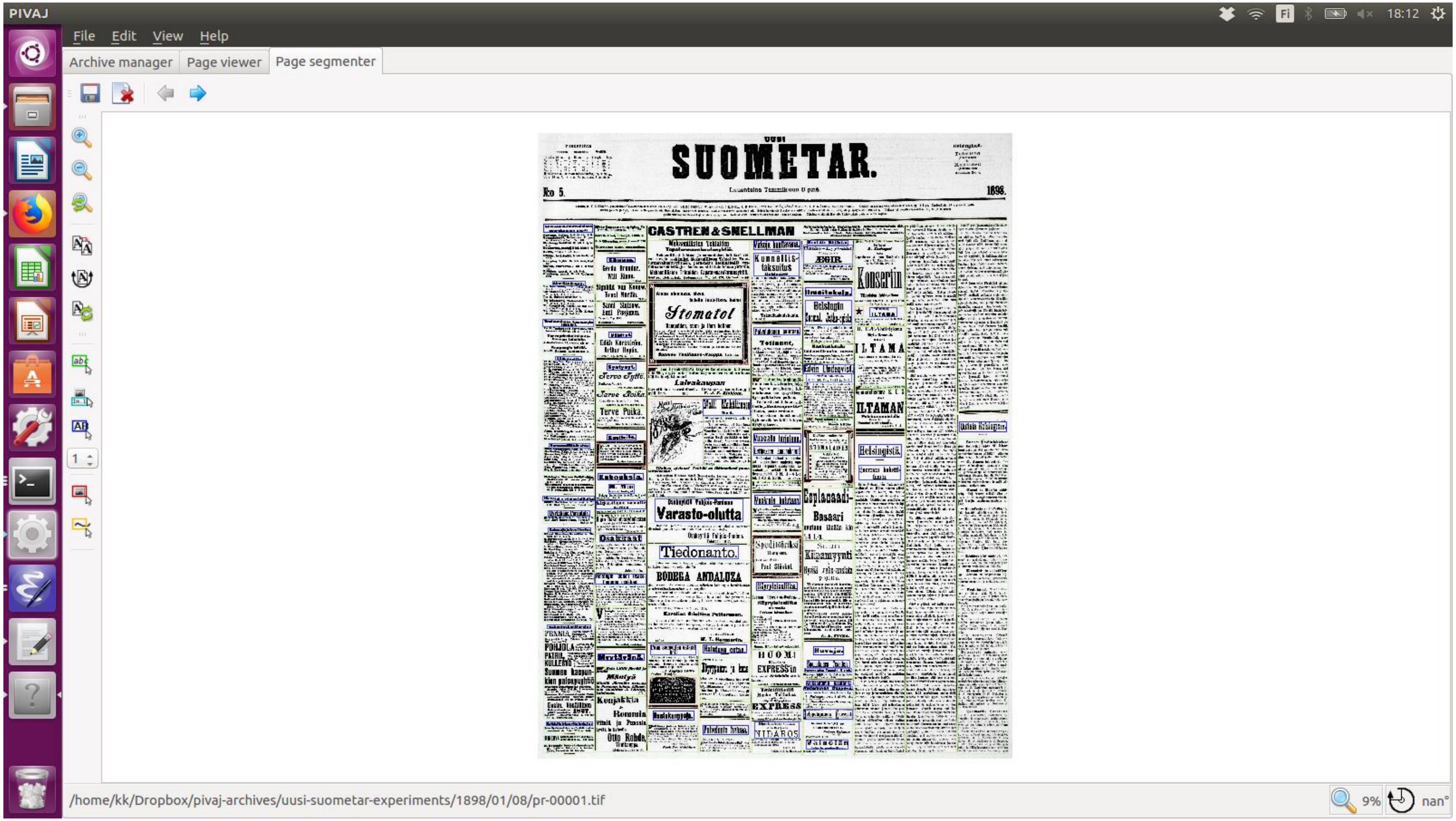
Article extraction trials at the NLF

- We are collaborating with University of Rouen, LITIS laboratory
- PIVAJ software: a machine learning software, that learns layouts and article separation from training data
- We have made a 224 page manually marked collection out of Uusi Suometar for PIVAJ
- We are trying to train PIVAJ to separate articles from pages of Uusi Suometar
- Results still forthcoming…

A simple layout page marked with PIVAJ’s annotation tool

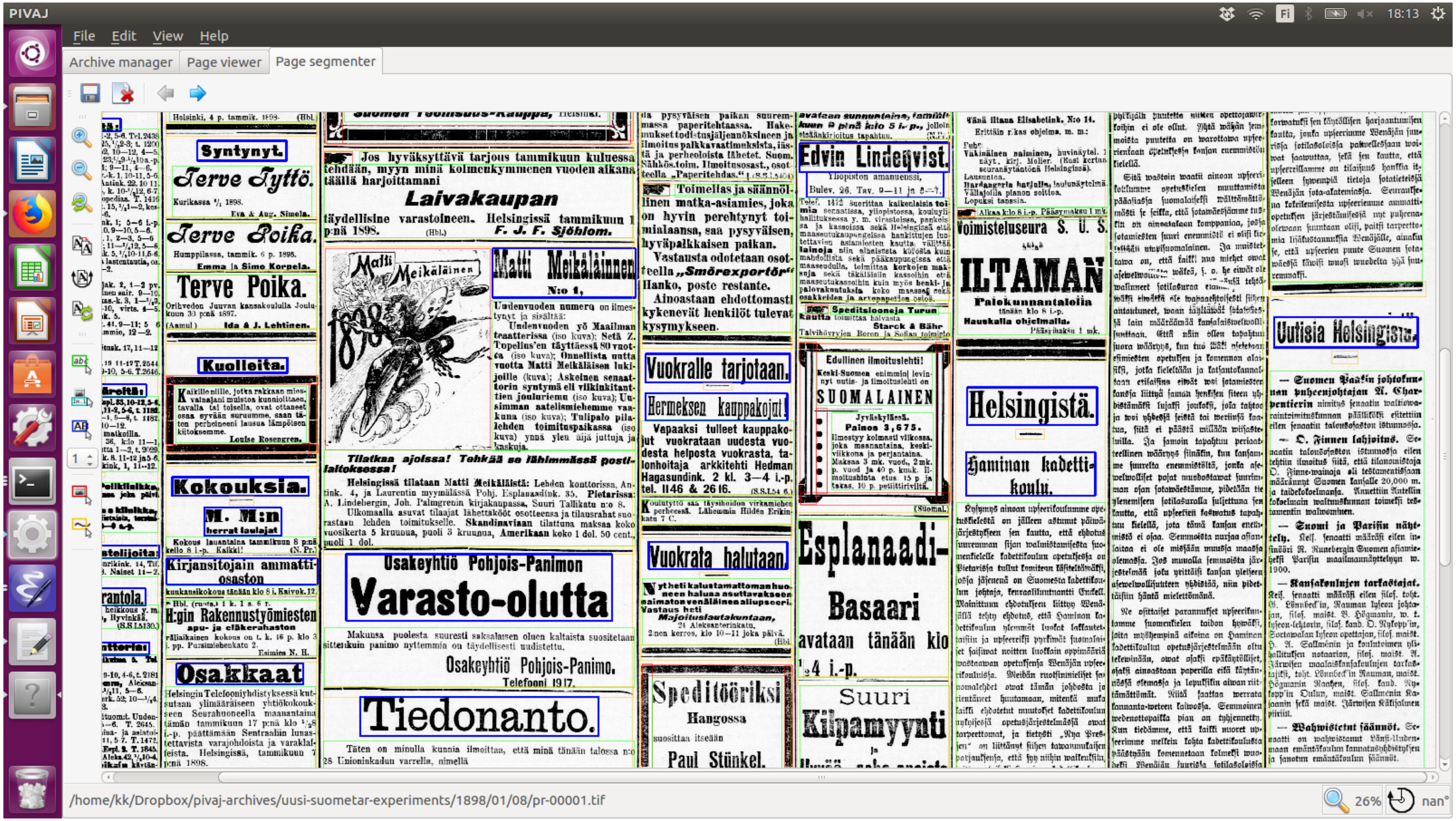


/home/kk/Dropbox/pivaj-archives/uusi-suometar-experiments/1869/01/04/ac-00002.tif



/home/kk/Dropbox/pivaj-archives/uusi-suometar-experiments/1898/01/08/pr-00001.tif

Uusi Suometar: different parts marked on a page with PIVAJ in the training data



/home/kk/Dropbox/pivaj-archives/uusi-suometar-experiments/1898/01/08/pr-00001.tif

In conclusion

- NLF has multilingual newspaper and journal data from 1771-1929 available on the web: digi.kansallikirjasto.fi
- We have been working with data and usability improvement since about 2014 in different projects
- NLF is part of the NewsEye consortium providing data and looking forward to have useful results out of the project

Thank you for your patience

Kimmo.kettunen@helsinki.fi

European Union
European Regional Development Fund

Leverage from the EU 2014–2020