Submitted on: 7/7/2014

Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods

Kimmo Kettunen
Center for Preservation and Digitisation, National Library of Finland, Mikkeli, Finland

Timo Honkela
Center for Preservation and Digitisation, National Library of Finland, Mikkeli, Finland
Department of Modern Languages, University of Helsinki, Finland

Krister Lindén
Department of Modern Languages, University of Helsinki, Finland

Pekka Kauppinen
Department of Modern Languages, University of Helsinki, Finland

Tuula Pääkkönen
Center for Preservation and Digitisation, National Library of Finland, Mikkeli, Finland

Jukka Kervinen
Center for Preservation and Digitisation, National Library of Finland, Mikkeli, Finland

Abstract:

In this paper, we study how to analyze and improve the quality of a large historical newspaper collection. The National Library of Finland has digitized millions of newspaper pages. The quality of the outcome of the OCR process is limited especially with regard to the oldest parts of the collection. Approaches such as crowdsourcing has been used in this field to improve the quality of the texts, but in this case the volume of the materials makes it impossible to edit manually any substantial proportion of the texts. Therefore, we experiment with quality evaluation and improvement methods based on corpus statistics, language technology and machine learning in order to find ways to automate analysis and improvement process. The final objective is to reach a clear reduction in the human effort needed in the post-processing of the texts. We present quantitative evaluations of the current quality of the corpus, describe challenges related to texts written in a morphologically complex language, and describe two different approaches to achieve quality improvements.

Keywords: Digitization, optical character recognition, error correction, corpus linguistics, string matching

1 INTRODUCTION

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 (Bremer-Laamanen 2001, 2005). This collection contains approximately 1.95 million pages in Finnish and Swedish. In the output of the Optical Character Recognition (OCR) process, errors are common especially when the texts are printed in the Fraktur (blackletter) typeface. The errors lower the usability of the corpus both from the point of view of human users as well as regarding potential text mining applications. Automatic spell checking and correction is problematic, for instance, due to the historical spelling variants.

The spelling variation leads to the situation that modern morphological analysis tools are not fully applicable. Developing new morphological models for various historical periods is one option but would be costly for a morphologically highly complex language such as Finnish. One approach for improving the quality of the texts after the OCR process is to apply crowdsourcing that is also called human computing. When the number of documents is in millions and different word forms, among which many are incorrect, is in tens of millions, human efforts can provide only a partial solution. In this paper, we consider the use of language technology, corpus statistics and statistical machine learning methods as means to help the correction process. We describe the use of the collection through a search interface and as a corpus for researchers in linguistics, assess the current level of quality of the corpus, and present two approaches that can be used in the quality improvement.

2 HISTORICAL NEWSPAPER COLLECTION AND ITS USE

The duty of the National Library of Finland is to deposit and preserve everything published in Finland. The Digitisation Policy of the Library[1] outlines the strategic objectives, the regulations and recommendations, the content to be digitised, the life cycle management of digitised collections and the use and reuse of digital information resources. The current digitisation processes include various materials: newspapers, magazines, books, maps, ephemera and audio recordings. Regardless of the material type, they are processed in the same way to produce quality metadata and enabling access to all. An integrated workflow and tool-set enable cost-efficient digitisation.

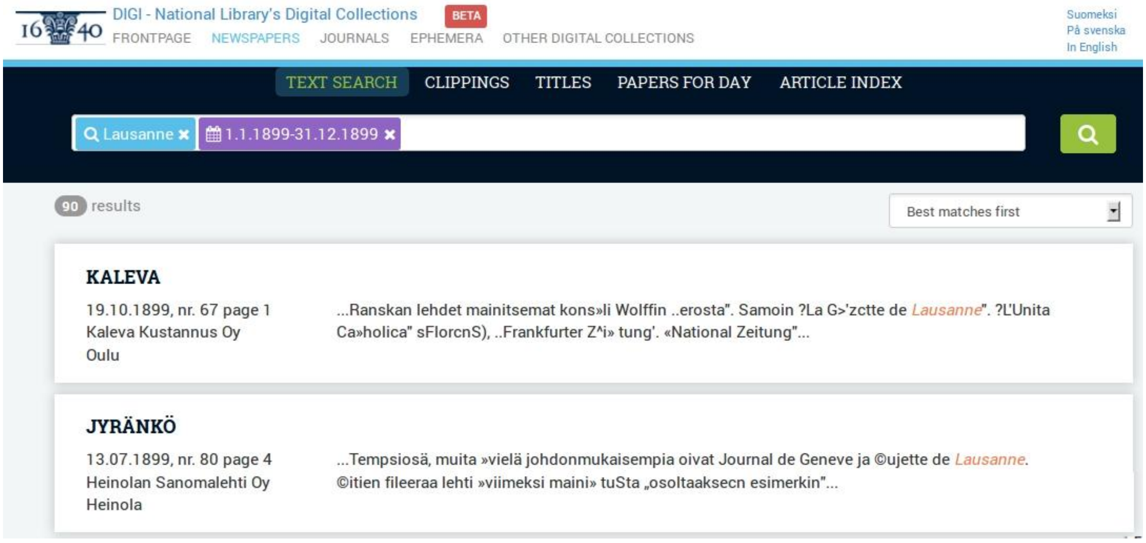
According to Legal Deposit law the National Library of Finland receives a copy of each newspaper and magazine published in Finland. The materials are processed according an internal concept called the digital chain. In the digital chain, the phases of material processing are 1) material deposit and return, 2) preparation and conservation if needed, 3) microfilming (of the newspapers), 4) scanning, 5) post-processing that includes a structural analysis, and 5) finally deployiment, use and preservation. In the first step, one crucial task is the selection of the material for the digitisation. There is naturally more incoming material than can be processed, so deciding what is selected for digitisation is important. Typically reasons vary from materials, which are in risk in the preservation sense, to materials which have high demand and would have several uses. If digital deposition is not used, the material is scanned with the resource available: automatic robotic scanner or manual scanner. Newspapers are scanned for microfilm.

[1] http://urn.fi/URN:NBN:fi-fe201401151119

Post-processing is at the core of the digitisation chain. In this phase, the material is processed so that it can be shared to the library sector and to the public use. In post-processing, the scanned images are improved and run through background software and processes, which create METS/ALTO metadata. The optical character recognition (OCR) is done at the same time for getting the text from the materials. Regardless of recent development with the OCR software, there are still challenges with it, as some material is very old, with varying paper quality, column amounts, various languages and varying from fraktur to antikva in font types. The images and metadata are combined to a package, where everything related to one issue is preserved. In the final phase these packages are deployed so that the materials are easily browsable in the on-line storage and retrieval system.

2.1 On-line storage and retrieval system

The National Library's Digital Collections are offered via the digi.kansalliskirjasto.fi web service. The web service contains different material types including newspapers, journals, and ephemera. Recently a new service was created to enable marking of clips and storing them to a personal scrapbook. The web service is used, for example, by genealogists, heritage societies, researchers, and history enthusiasts. There is also increasing desire to offer the material more widely for education purposes.



90 results

KALEVA

19.10.1899, nr. 67 page 1
Kaleva Kustannus Oy
Oulu

...Ranskan lehdet mainitsemat kons»li Wolffin ..erosta". Samoin ?La G>'zctte de *Lausanne*". ?L'Unita Ca»holica" sFlorcnS), ..Frankfurter Z^i» tung'. «National Zeitung"...

JYRÄNKÖ

13.07.1899, nr. 80 page 4
Heinolan Sanomalehti Oy
Heinola

...Tempsiosä, muita »vielä johdonmukaisempia oivat Journal de Geneve ja ©ujette de *Lausanne*. ©itien fileeraa lehti »viimeksi maini» tuSta „osoltaaksecn esimerkin"...

Fig. 1 A result of a search on the Finnish historical newspaper collection where articles with the keyword "Lausanne" and published in 1899 were looked for.

In the deployment phase of the material to the on-line retrieval system, the text produced is indexed in order to facilitate faster retrieval and accurate search results. Quite often the most common search terms are free text or approximate search terms, which allows partial matches to the vast amount of text in the digital collections. Approximate searches are most useful in the text-based use, as they allow some OCR errors to be omitted, and improve the recall of the search results, but naturally lowering the precision to some degree. The search log of the digi.kansalliskirjasto.fi web service is being followed in order to develop the system further taking into account the real needs of the users. In the following, we will consider the contents of the search log in some detail.

2.2 Log analysis

Every web service keeps a log of the searches that are performed on its pages. Search logs contain usually different amount of information about the user's transactions, and log analysis of searches can be useful in showing how collections are used and what the users are searching for. We had access to a log of about 220 000 searches in Finnish and Swedish covering a period of three years and analyzed briefly the search terms used by the users. The log consists of ca. 435 000 query term tokens and 96 000 query term types. The mean length of the queries is slightly less than two search terms, which is quite typical for web searches in general (Jansen et al. 2000). Longer queries are also used, but from three term queries the amount of queries drops heavily. About 85 per cent of the queries consist of 1-3 query terms. About 57 00 query terms occur only once in the query log. Thousand most frequent query terms occur about 150 000 times in the log, and 10 000 most frequent query terms occur about 295 000 times.

When one studies the most frequent query terms, different types of names are the most frequently used in searches. Among the 100 most frequent search terms about 70 % are names, most of them personal names, consisting mainly of first names. Surnames and place names (toponyms) are far less common in the top, toponyms being a bit more common than surnames. When the top 1000 most frequent search terms are studied (149 056 term tokens), about 80 % of them are names on type level: 30 % first names (42 % on token level), 30 % surnames (21 % on token level) and 20 % names of places (14 % on token level), mostly towns or other smaller localities with a small percentage of names of countries included. Only about 15 % of the search terms are common nouns on type level (12 % on token level). Heavy usage of proper names in searches is a typical phenomenon, as they give access to persons and places which are many times important in information-seeking behavior of humanists, genealogists etc. (Bremer-Laamanen 2006, Crane and Jones 2006).

2.3 Newspaper collection in FIN-CLARIN

The historical newspaper collection has also been made available to researchers through FIN-CLARIN. The FIN-CLARIN consortium is the Finnish part of the European_CLARIN collaboration that aims to build an infrastructure for language resources and technology for researchers in the Arts and Humanities. The corpora in FIN-CLARIN are provided by the University of Helsinki and hosted by CSC - IT Center for Science Ltd, in the Language Bank of Finland.[2] The collection can be accessed through the Korp[3] environment that has been developed by Språkbanken at the University of Gothenburg and extended by FIN-CLARIN team at the University of Helsinki to provide concordances of text resources. Other services in the Language Bank of Finland include FinnWordNet[4] with a full replica of WordNet in Finnish, and the Language Archive Technology[5] containing a collection of speech and video resources.

The foreseen use of the historical newspaper collection in FIN-CLARIN is twofold, i.e. the collection serves language researchers interested in various linguistic phenomena over time as well as researchers in Arts and Humanities interested in how phenomena in society have

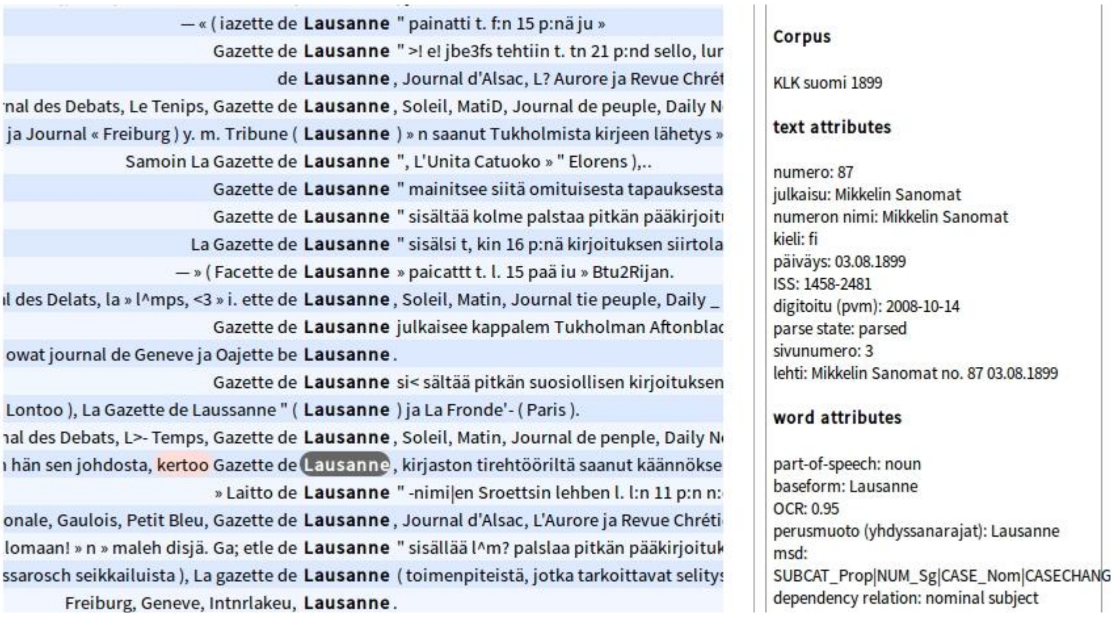
[2] http://www.kielipankki.fi

[3] https://korp.csc.fi/

[4] http://www.ling.helsinki.fi/en/lt/research/finnwordnet/

[5] https://lat.csc.fi

developed. The first group is interested in the collection on a sentential level, whereas the other group is interested in the texts they can read. To cater to both groups, the text collection has been mechanically processed both to provide more specific annotations of linguistic phenomena as well as to provide references to the full news items the search hits refer to. Both the linguistic and the bibliographic views of the hits can be downloaded by researchers for further use. To facilitate search for both groups of researchers, the collection will be further processed with named-entity recognition distinguishing person, place and organization names.



```
                          — « ( iazette de Lausanne " painatti t. f:n 15 p:nä ju »
                              Gazette de Lausanne " >! e! jbe3fs tehtiin t. tn 21 p:nd sello, lur
                                      de Lausanne , Journal d'Alsac, L? Aurore ja Revue Chrét
nal des Debats, Le Tenips, Gazette de Lausanne , Soleil, MatiD, Journal de peuple, Daily N
 ja Journal « Freiburg ) y. m. Tribune ( Lausanne ) » n saanut Tukholmista kirjeen lähetys »
                   Samoin La Gazette de Lausanne ", L'Unita Catuoko » " Elorens ),..
                              Gazette de Lausanne " mainitsee siitä omituisesta tapauksesta
                              Gazette de Lausanne " sisältää kolme palstaa pitkän pääkirjoit
                           La Gazette de Lausanne " sisälsi t, kin 16 p:nä kirjoituksen siirtola
                          — » ( Facette de Lausanne » paicattt t. l. 15 paä iu » Btu2Rijan.
l des Delats, la » l^mps, <3 » i. ette de Lausanne , Soleil, Matin, Journal tie peuple, Daily _
                              Gazette de Lausanne julkaisee kappalem Tukholman Aftonblac
owat journal de Geneve ja Oajette be Lausanne .
                              Gazette de Lausanne si< sältää pitkän suosiollisen kirjoituksen
Lontoo ), La Gazette de Laussanne " ( Lausanne ) ja La Fronde'- ( Paris ).
al des Debats, L>- Temps, Gazette de Lausanne , Soleil, Matin, Journal de penple, Daily N
 hän sen johdosta, kertoo Gazette de Lausanne , kirjaston tirehtööriltä saanut käännökse
                               » Laitto de Lausanne " -nimi|en Sroettsin lehben l. l:n 11 p:n n:
onale, Gaulois, Petit Bleu, Gazette de Lausanne , Journal d'Alsac, L'Aurore ja Revue Chréti
lomaan! » n » maleh disjä. Ga; etle de Lausanne " sisälläa l^m? palslaa pitkän pääkirjoituk
ssarosch seikkailuista ), La gazette de Lausanne ( toimenpiteistä, jotka tarkoittavat selitys
           Freiburg, Geneve, Intnrlakeu, Lausanne .
```

Corpus

KLK suomi 1899

text attributes

numero: 87
julkaisu: Mikkelin Sanomat
numeron nimi: Mikkelin Sanomat
kieli: fi
päiväys: 03.08.1899
ISS: 1458-2481
digitoitu (pvm): 2008-10-14
parse state: parsed
sivunumero: 3
lehti: Mikkelin Sanomat no. 87 03.08.1899

word attributes

part-of-speech: noun
baseform: Lausanne
OCR: 0.95
perusmuoto (yhdyssanarajat): Lausanne
msd:
SUBCAT_Prop|NUM_Sg|CASE_Nom|CASECHANG
dependency relation: nominal subject

Fig. 2 A Keyword-in-Content (KWIC) result provided by the Korp system.

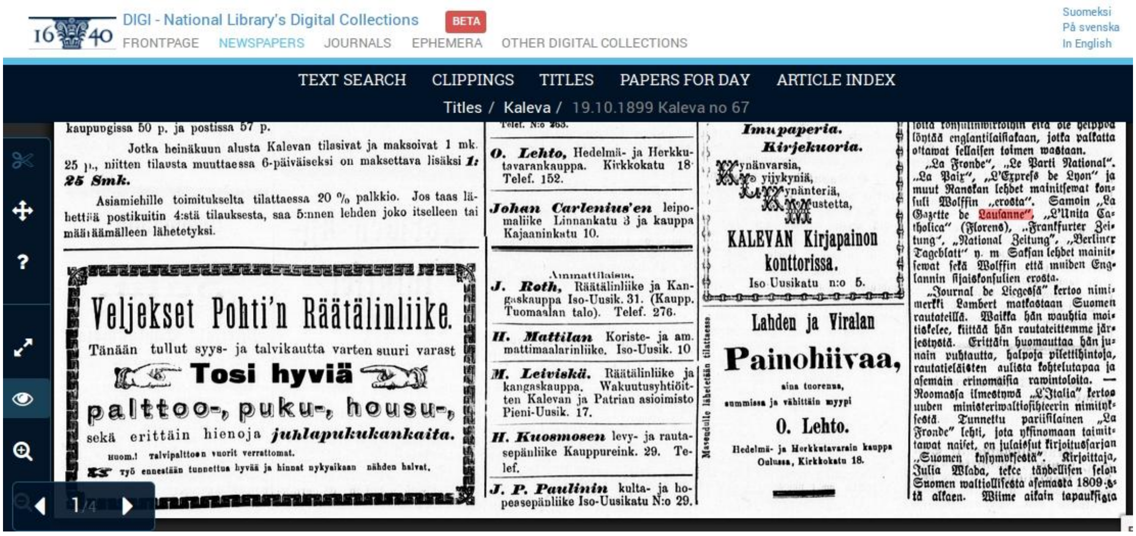
The collection is clearly valuable to diverse areas of research so it would be worthwhile to improve the OCR quality of the news items, many of which have originally been published in Gothic script which is less readable to current researchers. The correction needs to be conducted either by improved OCR of the scanned text images or by post-processing of the OCR output.

3 ANALYSIS OF CORPUS QUALITY

In the following, we consider in detail what is the current quality of the historical newspaper collection in its digital form.

3.1 Quality of the Digi corpus

Quality of the OCR of the Finnish digital newspaper collection (Digi) has not been analyzed systematically so far, only some small tests have been performed. Bremer-Laamanen (2001) reports an accuracy of 95-97% for the Finnish part of the documents, supposedly on character level. In Digitalkoot, a crowd-sourced gamified OCR correction project, two articles of the collection were examined closely. They had accuracy of about 77% and 84.5% on word level, and the accuracy was corrected to over 99 % with crowdsourcing (Chrons and Sundell 2011). Raitanen (2012) reports word error rates varying from ca. 15 to 24 % based on an analysis of three documents, and expects an overall error rate of 20% to a subset of the Digi collection that consists of 180,468 documents. This document collection has been extracted from the Digi and made into an evaluated Cranfield style search collection at the University of Tampere.



kaupungissa 50 p. ja postissa 57 p.

Jotka heinäkuun alusta Kalevan tilasivat ja maksoivat 1 mk. 25 p., niitten tilausta muuttaessa 6-päiväiseksi on maksettava lisäksi ***1: 25 Smk.***

Asiamiehille toimitukselta tilattaessa 20 % palkkio. Jos taas lähettää postikuitin 4:stä tilauksesta, saa 5:nnen lehden joko itselleen tai määräämälleen lähetetyksi.

Veljekset Pohti'n Räätälinliike.

Tänään tullut syys- ja talvikautta varten suuri varast

Tosi hyviä

palttoo-, puku-, housu-,

sekä erittäin hienoja ***juhlapukukankaita.***

Huom.! Talvipalttoen vuorit verrattomat.

Työ ennestään tunnettua hyvää ja hinnat nykyaikaan nähden halvat.

Telef. N:o 263.

O. Lehto, Hedelmä- ja Herkkutavarankauppa. Kirkkokatu 18 Telef. 152.

Johan Carlenius'en leipomaliike Linnankatu 3 ja kauppa Kajaaninkatu 10.

Ammattilaisia.

J. Roth, Räätälinliike ja Kangaskauppa Iso-Uusik. 31. (Kaupp. Tuomaalan talo). Telef. 276.

H. Mattilan Koriste- ja ammattimaalarinliike. Iso-Uusik. 10

M. Leiviskä. Räätälinliike ja kangaskauppa. Wakuutusyhtiöitten Kalevan ja Patrian asioimisto Pieni-Uusik. 17.

H. Kuosmosen levy- ja rautasepänliike Kauppureink. 29. Telef.

J. P. Paulinin kulta- ja hopeasepänliike Iso-Uusikatu N:o 29.

Imupaperia.

Kirjekuoria.

Kynänvarsia,
Kivijykyniä,
Kynänteriä,
Mustetta,

KALEVAN Kirjapainon konttorissa.

Iso Uusikatu n:o 5.

Maseudulle lähetetään tilattaessa

Lahden ja Viralan

Painohiivaa,

aina tuoreena,
summissa ja vähittäin myypi

O. Lehto.

Hedelmä- ja Herkkutavarain kauppa
Oulussa, Kirkkokatu 18.

jolla konſulinwirkoihin eikä ole helppoa löytää englantilaiſiakaan, jotka palkatta ottawat ſellaiſen toimen wastaan.

„La Fronde“, „Le Parti National“. „La Paix“, „L'Expreſs de Lyon“ ja muut Ranskan lehdet mainitſewat konſuli Wolffin „erosta“. Samoin „La Gazette de Lauſanne“, „L'Unita Catholica“ (Florens), „Frankfurter Zeitung“, „National Zeitung“, „Berliner Tageblatt“ y. m Sakſan lehdet mainitſewat ſekä Wolffin että muiden Englannin ſijaiskonſulien erosta.

„Journal de Liegessä“ kertoo nimimerkki Lambert matkastaan Suomen rautateillä. Waikka hän wauhtia moitiskelee, kiittää hän rautateittemme järjestystä. Erittäin huomauttaa hän junain puhtautta, halpoja pilettihintoja, rautatieläisten aulista kohtelutapaa ja asemain erinomaiſia rawintoloita. — Roomassa ilmestywä „L'Italia“ kertoo uuden ministeriwaltiosihteerin nimityksestä. Tunnettu pariisilainen „La Fronde“ lehti, jota yksinomaan toimittawat naiset, on julaissut kirjoitusſarjan „Suomen kysymyksestä“. Kirjoittaja, Julia Wlaba, tekee täydellisen ſelon Suomen waltiollisesta asemasta 1809:stä alkaen. Wiime aikain tapauksista

Fig. 3 An example of a newspaper retrieved by the Digi system. The search term "Lausanne" has been highlighted.

To get further insight of the quality of the OCRed publicly available older content of Digi, we have created small test collections of the digitized material that have hand edited equal content available. Some of the newspaper and journal materials of 19th century have been hand edited in the Institute for the Languages of Finland[6], and their collections include e.g. some newspapers and periodicals from the time. We compiled seven smallish parallel corpora where the ground truth for the comparison consists of the hand edited versions of the same material National Library of Finland has digitised for the Digi web service. It is obvious, that the hand edited versions of the material are not totally error free, but their quality is the best available comparison. As the amount of evaluation material (about 212,000 words) is very small compared to the amount of the digitised pages, about 1.95M, the results should be taken as tentative.

For compilation of the evaluation corpus, XML encoding of the digitised texts was first removed and also possible differences in order of the content were checked before alignment of the texts was performed. Alignments of the corpora were performed with *LF-Aligner* 4.05[7] software semi-automatically. Length of the aligned segments in each test corpus varies from a few words to a text paragraph, depending mostly on the alignment software's capability of finding similar passages. Table 1 shows the amount of the aligned material used in our tests.

[6] http://www.kotus.fi/?l=en&s=1
[7] http://sourceforge.net/projects/aligner/

Collection	Number of aligned segments	Number of words
Suometar 1847	1 032	ca. 20,700
Keski-Suomi 1871	107	ca. 1,870
Kirjallinen Kuukauslehti 1870	415	ca. 6,200
Mehiläinen 1859	9,185	ca. 149,000
Sanan Saattaja Wiipurista 1841	751	ca. 16,400
Turun Wiikko-Sanomat 1831	744	ca. 6,900
Oulun Viikko-Sanomia 1841	661	ca. 10,700

Table 1. Sizes of the OCR quality evaluation collections

As the numbers in Table 1 show, some of the parallel collections are quite small, but a few more representative. Due to availability of edited clean material, the evaluation collection consists only of older material. Edited material from the late 19th and early 20th century is not available and should be hand-compiled for evaluation purposes. Also edited material with more easily recognizable Antiqua font is not available without hand-crafting. A few peeks into Antiqua samples of the Digi collection, however, seem to indicate that the OCR quality with this letter type is much higher than with Fraktur.

Quality of the OCR can be measured with different measures, but usually quality assessment of OCRed data is based on precision and recall on word and/or character level. Other measures, such as F-measure and F1, which are derived from precision and recall, are also used generally. Tanner et al. (2009) emphasize that character level accuracy does not guarantee word level accuracy: even with a quite high character level accuracy the word level accuracy may be low depending on the distribution of the misspelled character in the words.

As there is no single software that could give us 'the truth' of the quality of the material, we have performed our initial trials with the following four software:

- GTM 1.3 (General Text Matcher)[8]
- Meteor 1.5[9]
- OCR Frontiers Toolkit 1.0[10]
- OCR evaluation tool[11]

GTM and Meteor are both machine translation quality measurement software. Both of them are based on precision and recall measurement on word or sub-word level. GTM's single figure metric is based on F-measure (Turian et al. 2003). Meteor's measurement is based on n-gram comparison and it shows precision, recall, F1 and Fmean measures for compared

[8] http://nlp.cs.nyu.edu/GTM/

[9] http://www.cs.cmu.edu/~alavie/METEOR/

[10] https://code.google.com/p/isri-ocr-evaluation-tools/

[11] http://impact.dlsi.ua.es/ocrevaluation/

segments (Denkowski and Lavie 2014). OCR Frontiers Toolkit 1.0 and OCR evaluation tool are dedicated OCR quality measurement packages. OCR Frontiers Toolkit consists of several small programs that perform different tasks (Bagdanov et al. 1999). Word and character level accuracies can be counted with it and calculations from several runs can be combined. OCR evaluation tool gives WER, Word Error Rate, and CER, Character Error Rate, figures for the input.

Table 2 presents the evaluation results of the four software for our evaluation corpora at word level.

Collection	**GTM**
	F-measure
		Meteor
		P = precision
		R = recall
		F = F mean
			OCR evaluation tool
			WER
			(word error rate)
				OCR Frontiers Toolkit 1.0
				Accuracy

Suometar 1847	0.64	P 0.69 R 0.65 F 0.66	39.7	71.1 %
Keski-Suomi 1871	0.57	P 0.63 R 0.60 F 0.61	49.5	60.5%
Kirjallinen Kuukauslehti 1870	0.80	P 0.80 R 0.83 F 0.82	21.3	82.1%
Mehiläinen 1859	0.90	P 0.90 R 0.90 F 0.90	8.8	N/A
Sanan Saattaja Wiipurista 1841	0.65	P 0.69 R 0.65 F 0.66	34.8	73.8%
Turun Wiikko-Sanomat 1831	0.76	P 0.79 R 0.77 F 0.78	25.2	80.4%
Oulun Viikko-Sanomia 1841	0.80	P 0.82 R 0.81 F 0.81	19.7	83.0%

Table 2. OCR error evaluation results at word level

The results show that the quality of the OCRed data seems to be in general quite low in Suometar, Keski-Suomi and Sanansaattaja Wiipurista collections. Word level precision and recall in Suometar, Keskisuomalainen and Sanan Saattaja Wiipurista varies from about 0.57 to 0.69 when measured with GTM and Meteor. Kirjallinen Kuukauslehti, Mehiläinen, Turun Wiikko-Sanomat and Oulun Viikko-Sanomia are quite a lot better. In Kirjallinen Kuukauslehti and Oulun Viikko-Sanomia precision and recall are about 0.80 and in

Mehiläinen around 0.90. Thus it seems that about 30-50 % of the words in the test material of Suometar, Keskisuomalainen and Sanan Saattaja Wiipurista are misrecognized, and about 10-20 % of words in Kirjallinen Kuukauslehti, Mehiläinen and Oulun Viikko-Sanomia. OCR Frontiers Toolkit 1.0, the most detailed software in the test, is giving 65-74 % accuracy on word level for Suometar, Keski-Suomi and Sanansaattaja Wiipurista, and 82-83 % accuracy for Kirjallinen Kuukauslehti and Oulun Viikko-Sanomia. Data of Mehiläinen could not be run in OCR Frontiers Toolkit due to its size. The WER figures given by OCR evaluation tools show varyingly 35-49 % of errors in Suometar, Keski-Suomi and Sanan Saattaja Wiipurista which is in the same range as GTM and Meteor results. WER figures for Kirjallinen Kuukauslehti, Turun Wiikko-Sanomat and Oulun Viikko-Sanomia vary from 8.8 to 25. Although the measures differ in each software, the results converge and show the approximate error rate in the corpuses well enough.

In general, our results are lower than the accuracy results of the British Library newspaper collection reported by Tanner et al. (2009) with larger material and more accurate evaluation methods. Their average word level accuracy figures vary from 65 to 78 %. About half of the titles in their 19th Century Newspaper collection get word level accuracy of 80 to 89 %. Accuracy between 60-79 % is achieved in 47 % of the titles. Our tentative results seem to be low but not uncommonly low when compared to the BL results. Part of our material seems to have lots of errors; part of the material is not too bad, although there is a lot to achieve in error correction of the material.

3.2 Morphological recognition of the OCRed data

Higher level processing of documents involves usually morphological analysis. It is clear that OCR errors harm this partly; part of the problems is caused by OOV words that are not recognized by the morphological analyzers in historical texts. Part of the unanalyzed words are also words of different languages than Finnish, mostly Swedish. Words in Russian, German, and Latin may also occur in the texts. It is difficult to estimate, what is the effect of OCR errors and what the effect of OOV words.

Collection/ number of word types	**FinTWOL**	**Voikko**
Digi / 72.1 M	4.3 %	N/A
Tampere search collection / 7.03 M	13 %	11.8 %
VNS_Kotus / 0.53 M[12]	58.1%	N/A

Table 3. Recognized words for morphological analysers in the Digi corpus

We analyzed a substantial part of corpus of Digi with modern Finnish morphological analyzer, FinTWOL[13]. Rantanen (2012) has analyzed the index word list of the Tampere search collection with another analyzer, Voikko[14].For comparison we analyzed also the edited

[12] http://kaino.kotus.fi/sanat/taajuuslista/vns_frek.zip

[13] http://www2.lingsoft.fi/cgi-bin/fintwol

[14] http://voikko.puimula.org/

data of a 19th century word list compiled in the Institute for the Languages of Finland. Results for recognized words in the corpora are shown in Table 3. As the figures show number of recognized words for a modern language analyser are very low in the Digi corpus. In comparison over half of the words of the Kotus word list are recognized by FinTWOL, which suggests that majority of the the unrecognized words in the Digi collection are OCR errors.

Figure 4 shows, how the number of unknown words to FinTWOL increases in Digi when the data is analyzed starting from the most frequent word types ranging from 10K to the whole word list.

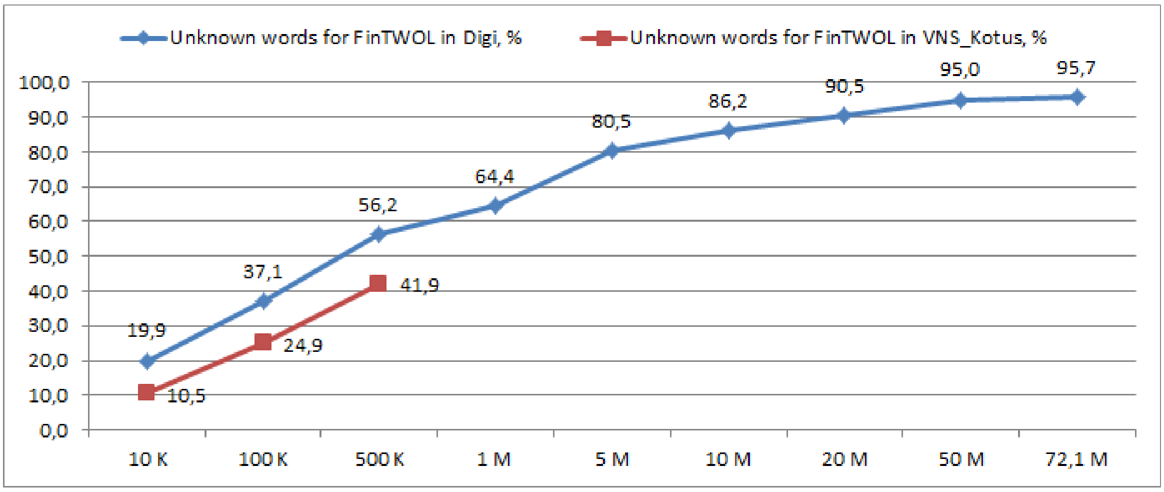


Fig. 4 Percentage of unknown words for FinTWOL in the top frequent word lists of Digi and VNS_Kotus ranging from 10 K to all of the data

The number of unknown words increases to over 50 % after the 500 000 most frequent word types and reaches 90 % in 20 million words. In the VNS_Kotus corpus the number of unknown words increases slower, but the size of the corpus is over fourteen times smaller.

Almost three quarters (73.4 %) of the unknown words in Digi occur only once. Figure 5 shows how the same numbers of word types occur as word form tokens in the Digi word list. The number of all tokens is 837.1 M.

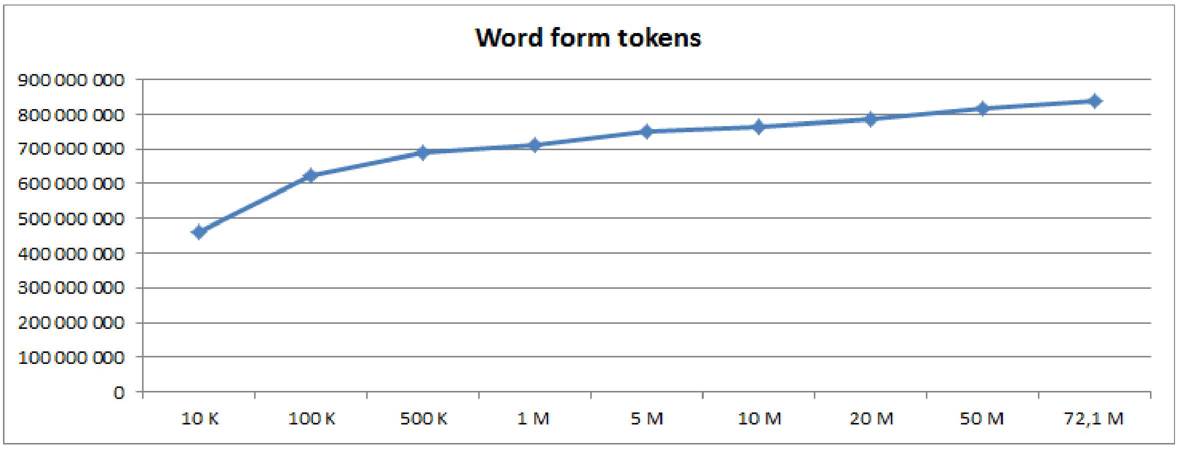


Fig. 5 Number of word form tokens in the Digi corpus related to frequency

The 10,000 most frequent word form types constitute 55.2 % of the word form tokens of the text (462.6 M). 1 million word form types constitute 85.1 % of the word form tokens of the text (712.1 M). From there on the token amount increases slowly.

4 IMPROVING CORPUS QUALITY

In the following, we discuss the challenges related to the task of improving the quality of a historical newspaper collection. We describe three approaches in which a central motivation is to deal with a large text collection written in a morphologically complex language. We do not present conclusive results but show promising routes for solutions that can later be used also in the production system.

4.1 Correction of word forms

Due to the large size of the Digi corpus, manual correction of the word instances in the texts is not possible. Furthermore, as there are tens of millions of separate word form types generated by the OCR process, even the types cannot be manually corrected even if this could be made in a context independent manner. Therefore, any means that can be used to automate the correction process are highly necessary.

In the previous section, we have shown the results of analyzing the Digi corpus with morphological analysis methods. These language technology tools can be used to find correctly recognized word forms. However, due to the fact that the historical newspaper corpus contains old forms and names, a number of false negatives would remain. A more important problem is that the knowledge of correct form does not directly indicate how incorrect forms should be transformed to correct forms. In the following, we study some approaches on how to correct word forms when availability of good quality training data is either assumed or not assumed. Before discussing methods for correcting the texts, we next present an approach for finding out the words that seem to require correction.

4.2 Detecting incorrect word forms

The first problem in the correction process is to determine whether a word form needs to be corrected or not. In principle, we could use a morphological analysis program in this task. If the word form is recognized by the program, it needs to be corrected, otherwise not. This procedure is widely used in spell checkers. With the linguistic models currently available, one problem is that the the overlap between modern and historical lexica is not sufficient. Another problem is that the word formation principles in Finnish have also evolved to some extent during the past 100-200 years. Therefore, we have applied a technique widely used in language detection algorithms based on collecting statistics of n-grams (see, e.g., Schmitt 1991, Häkkinen & Tian 2001, Vatanen et al. 2010).

In our approach, we collected a small corpus of corrected historical texts with 17,468 different word forms (types). The frequency of each trigram of calculated using this corpus. In order to estimate the “Finnishness” of a word, we did not apply a strictly probabilistic model, but used the product of the corpus frequencies of the trigrams in the word. Logarithm of zero is not defined, and therefore we used a small positive value (0.1) for out-of-vocabulary trigrams. The approach is well motivated and serves well our purposes.

The algorithm was heuristically tuned to give values above one for words that appear to be Finnish. For instance, for the incorrect form “ytsimicliscsti” (see Table 5 below), the algorithm gives the value 7.6 * 10^{-7}, whereas for the corresponding correct form “yksimielisesti” the Finnishness value is 21.4. The algorithm may, of course, provide high values for words that look like Finnish but are not within Finnish vocabulary. This can be a suitable complementary feature in comparison with morphological analysis tools with fixed lexica.

A problematic OCR error for this n-gram-based algorithm is the exchange of letters “m” and “w”, because there are many words in Finnish where either “w” (“v”) or “m” could be positioned. Therefore, the algorithm has hard time to determine if a word with one of these letters in some position is correct or not, or, in other words, the algorithm tends to overestimate the Finnishness of such words.

4.3. Frequency-based correction

A simple and straightforward approach for correction is to use the information that is available in the corpus itself. Tables 4, 5 and 6 demonstrate how the most frequent words tend to be correct. The lower the frequency of a word form, the more probable it is that the word contains even a large number of errors. Table 4 also demonstrates a very typical distribution that follows Zipf’s law.

Frequency	OCR form	Translation
23818195	ja	and
12473329	on	is
5737985	että	that
4003638	oli	was
3224891	ei	no
2501891	niin	so
2465708	hän	he, she
2457173	se	it
2447894	joka	that, which
2188373	sen	its

Table 4. The ten most frequent word forms in the Digi corpus.

When the number of word form instances is in thousands or hundreds, the words very frequently have errors even though they coincide frequencywise with rare lexical items or rare word forms. Finnish is a language with complex morphology. Every noun has approximately 2,000 different forms and every verb more than 10,000. Naturally, not all of these forms appear in a corpus but there are forms that are typically much more common than others (Kettunen 2014). However, the number of incorrect word forms can be vast. Each of the billions of correct word forms can further take potentially thousands of incorrect forms. This phenomenon follows the Anna Karenina principle (Tolstoy 1878). Rather than every unhappy family being unhappy in its own way, there are numerous ways of being incorrect. Arnold (2004) expresses the same by saying that in good situations a number of requirements must hold simultaneously, while in bad ones even one failure suffices. Detecting the difficult errors and correcting the most corrupted words would require contextual information and can be challenging also for humans. In this work, we do not attempt to use contextual information at the level of sentences.

Freq.	OCR form	Correct form	Edit distance	Translation
100	y**t**simi**c**lis**c**sti	yksimielisesti	3	unanimously
100	ys**läm**ällisesti	ystäwällisesti	2	kindly
100	todistuskappaleilla	todistuskappaleilla	0	with the pieces of evidence
100	peltikattovernissaa	peltikattovernissaa	0	tin roof varnish
100	**m**astaaminen	wastaaminen	1	answering
100	**l**y**f**ymy**lf**een	kysymykseen	3	into the question
100	k**n**stann**n**ksella	kustannuksella	2	at the expense of
100	glasgo**m**ista	glasgowista	1	from glasgow
100	annis**l**eluosa**l**eyhtiön	anniskeluosakeyhtiön	2	of the licensed limited liability company
100	a**m**ioliitoista	awioliitoista	1	of marriages

Table 5. Examples of word forms that appear one hundred times in the Digi corpus. The correct form is also shown as well as the edit distance between the original and the correct form.

The number of different word forms in the Digi corpus is 72,128,046 and 52,968,959 of them occur only once. A small selection of these unique word forms are shown in Table 6. Even though there are occasional correct word forms, a large majority of these forms are formed by the OCR process as incorrect recognitions. The edit distance between the original and the correct form is also typically high.

Freq.	OCR form	Correct form	Edit distance	Translation
1	zzhdysvautki	yhdyspankki	5	union bank
1	zzznuirypäleitä	wiinirypäleitä	4	grapes
1	wiljelystartaltutsessa	wiljelystarkoituksessa	4	in a cultivation purpose
1	urheilutarloinksiin	urheilutarkoituksiin	3	for sports purposes
1	uratkakupoissa	urakkakupoissa (urakkakaupoissa)	1	in contract jobs
1	taitanuiubcsta	taitawuudesta	4	of dexterity
1	taitamattomundestani	taitamattomuudestani	1	from my ineptitude
1	taiötelelutanteren	taistelutanteren	4	of the battlefield
1	taioafliftiutpn	tavallisuuden	8	of the usual
1	taimokkaisuudclllllln	tarmokkaisuudellaan	6	with his/her vigor

Table 6. An illustrative selection of word forms that appear once in the Digi corpus. The correct form is shown with the edit distance to the original form. One of the original forms in a misspelling in the newspaper.

For each correct form there are usually a number of incorrect forms. This relation can be used for our benefit. Namely, the correct form is in most cases more frequent than any of the incorrect ones. For instance, the form “zzznuirypäleitä” in Table 6 is an outcome of a short advertisement in the newspaper Keski-Suomi, published on 22nd of September, 1877[15].

If we use approximate string search on the corpus, we notice that with the edit distance four useful matches become available: “viinirypäleitä” (appears 426 times in the corpus), “wiinirypäleitä” (51) and “miinirypäleitä” (37). In modern Finnish, the letter “w” is not in use but historically it was used instead of “v”. In the third most frequent form, the word starts with “m” which is commonly mixed with “w” due to their similar shape.

[15] http://digi.kansalliskirjasto.fi/sanomalehti/binding/422857?term=zzznuiryp%C3%A4leit%C3%A4#?page=4

With this idea in mind, we can formulate the frequency-based correction algorithm:

```
1. Determine Finnishness value F of Word W
2. If F > Threshold T, accept W, else continue
3. Set EditDistance to 1
4. Repeat until (EditDistance > length(W)+2) or (EditDistance > 8) or solution found:
   4.1. Find Candidate words C in the corpus that are within EditDistance from W
   4.2. If C is an empty set, increment EditDistance by 1 and step back to 4.1.
   4.3. Return word within C that has highest frequency in the corpus and
        Finnishness value F of which is higher than Threshold T
   4.4. If no solutions are found, return word W.
```

The basic idea is to conduct approximate match string search until a high-frequency word form is found. The parameter for edit distance is incremented step by step. For approximate search, the computationally efficient *agrep* tool was used (Wu & Manber 1992).

4.4. Representation and correction based on letter shapes

The main problem in relying on edit distances in correcting words is the fact that all changes are considered equal. However, in the OCR process, recognition errors are related to the similarities of the letter shapes. It is much more probable that “w” is recognized as “m” or vice versa than either of them as “o”. It would be possible to construct a model of the transformations based on a data set where the original and corrected forms are side by side. This kind of approach will be presented in Chapter 4.5. Here we consider another option that does not require any availability of corrected texts.

In Chapter 4.3. we presented an algorithm for correction based on edit distances. We modified this algorithm by adopting a weighted edit distance scheme. In the OCR process, the errors made by the system are influenced by the shaped of the letters (Bhatti et al. 2014). We wanted to create a scheme in which the cost of replacing, for instance, “w” with “m” would be considered to be lower than with some other letter. We started by downloading the Fraktur typeface samples available in Wikipedia[16]. We cropped, normalized and resized the images with ImageMagick[17] to be in 10x6 pixel grayscale format. These matrices were flattened to be 600-dimensional vectors. The distance between each pair of vectors was calculated to provide us with a distance matrix. The upper left corner of this distance matrix is shown in Table 7. The values are between 0 (the same shape) and 100 (the largest distance found among the pairs). Integer values were used in order to increase computational efficiency of the subsequent steps.

[16] https://en.wikipedia.org/wiki/Fraktur

[17] http://www.imagemagick.org/

	a/A	ä/Ä	b/B	c/C	d/D	e/E
a/A	0	27	52	85	60	76
ä/Ä	27	0	55	79	58	71
b/B	52	55	0	74	41	58
c/C	85	79	74	0	81	38
d/D	60	58	41	81	0	67
e/E	76	71	58	38	67	0

Table 7. The upper left corner of a distance matrix that encodes similarities between letter shapes.

In order to ensure that the shape patterns were encoded in an appropriate manner, the distance matrix was given as an input to the Self-Organizing Map (SOM) algorithm (Kohonen 2001, Kohonen & Honkela 2007). The map clearly shows that the formation of the distance matrix has been successful as similarity of letter shape coincides with closeness on the map.

The edit (Levenshtein) distance calculation algorithm was modified to take into account letter-specific substitution cost[18]. The insertion and deletion costs remain the same as before. The weighted edit distance provides, at best, significant chance to improve the performance of the system especially if the distance between the original string and the candidates is high (cf., e.g., the cases shown in Table 6.

The evaluation results regarding the methods presented in Chapters 4.3. and 4.4. are still inconclusive. The methods improve the quality of the texts only in a subset of test corpora. Clear improvements are to be expected when the statistical approach is integrated with the natural language processing tools and resources.

[18] The C-language implementation of the weighted edit distance algorithm is available upon request from one of the authors (T.H.).

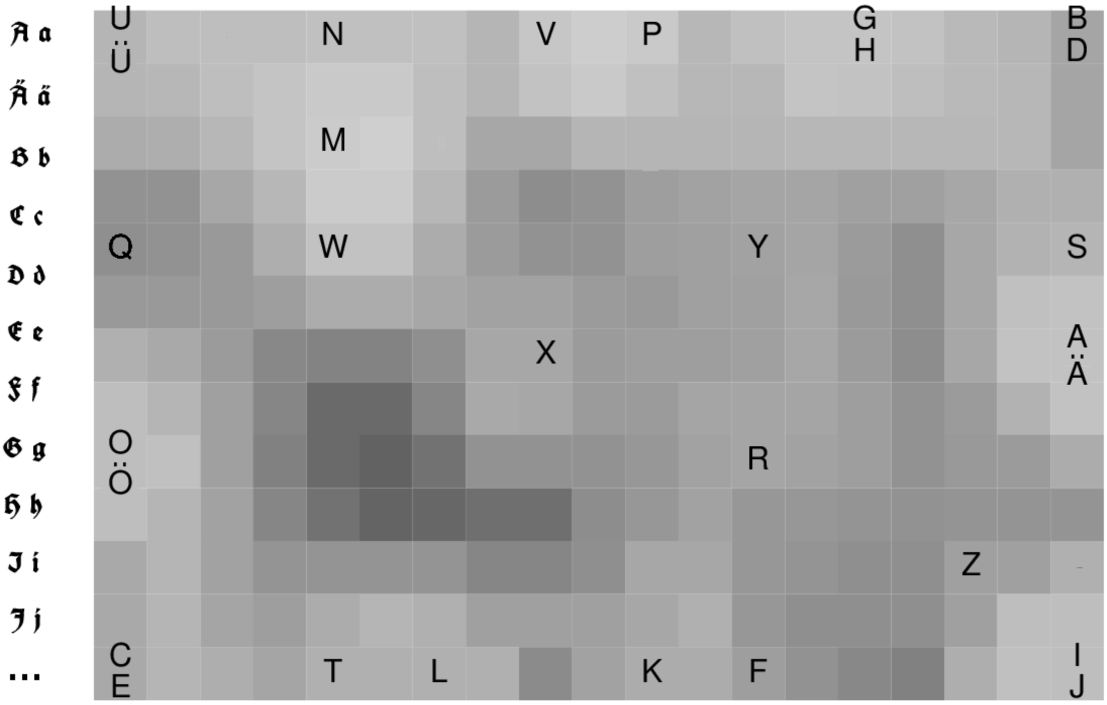


Fig. 6 A self-organizing map of Fraktur letter shapes. Two letters are close to each other on the map if their shape is relatively similar. The darker the shade of gray on the map, the larger the distance in the original, high-dimensional space.

4.5 Transformation-rule based correction

In our second main approach, the character substitution operations needed to correct the OCRed words are expressed as an instance of the general edit distance and implemented using weighted finite-state transducers. Correction candidates are generated from erroneous word forms by applying weighted context-sensitive parallel substitution rules. The method requires some training data consisting of OCRed words and their manually corrected target words.

In the first step, all possible character substitutions are extracted from the training data. This is achieved by aligning each OCRed string with its target string by using minimum edit distance (the character "0" symbolizes here an empty string):

kuroäl:kuwat → k:k u:u r:0 o:w a:a l:t

Sequences of successive character pairs or character pair string n-grams (in this case, unigrams, bigrams, trigrams, tetragrams) are then derived from this data.

The hash symbol (#) is used to pad the strings in order to match string-initial and string-final correspondences:

#:#		
k:k		
u:u		*#:# k:k u:u,*
r:0		*k:k u:u r:0,*
o:w	→	*u:u r:0 o:w,*
a:a		*r:0 o:w a:a,*
l:t		...
#:#		

In the next step, the pair string n-grams and their frequencies are converted into context-sensitive substitution rules. In each n-gram, the second pair string from the left represents the substitution operation, whereas the surrounding symbols serve as context. For instance, the pair string trigram *r:0 o:w a:a* maps the letter "o" to the letter "w" between the letters "r" and "a" and yields the rule *o* → *w* || *r* _ *a*. Each operation is assigned a logarithmic weight according to its probability in the context, i.e., the weight for the operation *X*→ *Y* || *A* _ *B* is calculated using the formula -log(F_x/F_y) where F_x=freq(*A:? X:Y B:?*) and F_y=freq(*A:? X:? B:?*). The symbol *?* denotes any symbol of the alphabet.

In order to eliminate the most unlikely correspondences as well as to keep the size of the rule set and the transducer within reasonable limits, the frequency of an operation in a given context must exceed a certain threshold in order for the operation to be included in the rule set. The symbol *T* will be used from here on for the minimum number of required occurrences. The value of *T* can be changed to modify the productivity and the permissiveness of the rule set.

The rules can then compiled into a finite-state transducer that may be paired up with a lexicon automaton to filter out non-word candidates when performing OCR correction.

A sample of 23,486 tokens from the collections presented in Chapter 3.1. together with their original OCRed versions was used training data for building the rule set. We used the HFST toolkit[19] to compile the rules into transducers. Preliminary evaluation of the method was done with test data consisting of 2,871 tokens with the WER of 18.5, which equals to 531 incorrect word forms. Three different kind of tests were performed, each with different values of *T*:

Test 1: Only non-word errors are corrected. If the input string is found in the lexicon, nothing is done. Otherwise, the string is passed on to the general edit distance transducer. The four output strings with the smallest weights that are found in the lexicon are checked to see whether the correct alternative is among them.

Test 2: All input strings are treated as potentially erroneous and are passed on to the general edit distance transducer. The four output strings with the smallest weights that are found in the lexicon are checked to see whether the correct alternative is among them.

[19] https://kitwiki.csc.fi/twiki/bin/view/KitWiki/HfstHome

Test 3: No lexicon is used at all. The input string is passed on directly to the general edit distance transducer. The four output strings with the smallest weights are checked to see whether the correct alternative is among them.

In tests 1 and 2, the OMorFi morphological analyzer[20] was used as the lexical acceptor. The analyzer was modified slightly to also accept some of the more archaic spelling variants such as those containing the letter "w" instead of "v" and "tz" instead of "ts".

The results of the preliminary tests are shown in Table 8. The top row shows the ranking of the correct alternative among the output strings. In test 1, the input strings that were found in lexicon and required no correction have the ranking of 0, whereas real-word errors that passed without being corrected have the ranking of -1.

***T* = 6**	**-1**	**0**	**1**	**2**	**3**	**4**	**> 4**	**total**
Test 1	131	1920	76	5	1	0	738	2871
Test 2	-	-	989	7	1	0	874	2871
Test 3	-	-	2403	121	25	5	317	2871

***T* = 4**	**-1**	**0**	**1**	**2**	**3**	**4**	**> 4**	**total**
Test 1	131	1920	76	7	1	0	736	2871
Test 2	-	-	1989	13	2	0	867	2871
Test 3	-	-	2403	125	13	22	308	2871

***T* = 2**	**-1**	**0**	**1**	**2**	**3**	**4**	**> 4**	**total**
Test 1	131	1920	76	9	1	1	733	2871
Test 2	-	-	903	17	4	0	947	2871
Test 3	-		2296	93	41	15	426	2871

Table 8. Results of the preliminary tests with different values of *T*, the minimum number of required occurrences

Predictably, the best results were achieved in when no lexicon was used. If the output string having the ranking of 1 was picked as the correction alternative in Test 3, the number of incorrect word forms in the test data would only be reduced by 11.9% at best. In all the other cases, no improvement would be achieved. Since an overwhelming majority of the existing manually corrected material dates from the early and mid-19th century, the use of OMorFi (which is primarily designed for modern Standard Finnish) as the lexicon for Early Modern Finnish is of little help. However, the OMorFi model could yield considerably better results

[20] http://code.google.com/p/omorfi/

when correcting Fraktur material from the early 20th century, since the standard language from that era is already very close to modern standard Finnish.

5 CONCLUSIONS AND DISCUSSION

In this paper, we have analyzed the properties of a large historical newspaper collection digitized by the National Library of Finland, and presented methods that can be used to improve the quality of the texts. In particular, we have considered challenges that are caused by the size of the collection and the fact that the complexity of Finnish morphology is high in comparison, for example, with most Indo-European languages. We have discussed the use of different approaches. In the future work, we will integrate these different approaches to a single system that can be used in improving the quality of the texts in the Digi production system and as a FIN-CLARIN corpus. The integration of different information sources may be easiest if the modules are defined in a common probabilistic framework. At that stage, also a more careful comparison with related methods and earlier results will be appropriate. The quality improvement task is by no means new, rather there is a substantial amount of relevant scientific and methodological literature in the field (cf., e.g., Boytsov 2011, Ford et al. 2011, Gotscharek et al. 2009, Kukich 1992, Ringlstetter et al. 2007, Schmitt 1991, Strohmaier et al. 2003).

The qualitative experience gained so far indicates that clear improvements can be obtained but a careful quantitative evaluation remains as a future task. Moreover, in order to achieve as high quality result as can be reasonably expected, some further considerations have to be taken into account. Maybe the most central issue is to enable sentence-level context-sensitive transformations. The computational complexity of such an approach is, however, high which may limit the practical implementability. Contextual information could include knowledge related to the usage of words, word forms and phrases in different historical time periods. Relevant contextual information could also be processed in computationally efficient manner using topic models (Blei et al, 2003, Honkela et al. 2010). A number of specific details can also be taken into account. For instance, occasionally a number of words have been concatenated in the OCR process. A good segmentation of long strings can be obtained using the Morfessor method (Creutz & Lagus 2007). Regarding morphological analysis, the OMorFi model can be extended to include lexical items and inflection rules of historical Finnish. In general, it can be foreseen that a better quality texts will promote the use of the collection both by laypersons looking for historical information as well as by researchers within different areas of digital humanities.

Acknowledgments

K.K., T.H. and T.P. are grateful for the financial support by the EU Commission through its European Regional Development Fund, and the program “Leverage from the EU 2007-2013”. K.L., M.S. and P.K. are grateful to FIN-CLARIN and the Academy of Finland for financial support. We warmly thank the Institute for the Languages of Finland that has kindly provided as with the corrected text corpus that was used for evaluation purposes. Maj-Lis Bremer-Laamanen, Kaj Ekholm, Kimmo Koskenniemi, Anna Mauranen, Arto Mustajoki and Esko Ukkonen are warmly thanked for their determined efforts in promoting digital humanities in a manner that has been constitutive for this work.

References

Arnold, V. I. 2004. Teoriya Katastrof (Catastrophe Theory), 4th ed. Moscow, Editorial-URSS.

Bagdanov, A.D., Rice, Stephen V. and Nartker, T.A. 1999. *The OCR Frontiers Toolkit*. Version 1.0. Information Science Research Institute. Available at https://code.google.com/p/isri-ocr-evaluation-tools/

Bhatti, Z., Waqas, A., Ismaili, Imdad Ali, Hakro, Dil Nawaz and Soomro, W.J. 2014. Phonetic based SoundEx & ShapeEx algorithm for Sindhi Spell Checker System. *arXiv preprint arXiv:1405.3033*.

Blei, D. M., Ng, A. Y. and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, pp. 993-1022.

Boytsov, L. 2011. Indexing methods for approximate dictionary searching: Comparative analysis. *Journal of Experimental Algorithmics (JEA)* 16:Article A, 92 pages.

Bremer-Laamanen, M.-L. 2001. A Nordic Digital Newspaper Library. *International Preservation News*, vol. 26, pp. 18-20.

Bremer-Laamanen, M.-L. 2005. Connecting to the past – newspaper digitisation in the Nordic Countries. *World Library and Information Congress: 71th IFLA General Conference and Council, "Libraries - A voyage of discovery"*, August 14th - 18th 2005, Oslo, Norway. Available at http://archive.ifla.org/IV/ifla71/papers/019e-Bremer-Laamanen.pdf

Crane, G. and Jones, V. 2006. The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection. *Proceedings of JCDL'06*, June 11–15, 2006, Chapel Hill, North Carolina, USA. Available at http://repository01.lib.tufts.edu:8080/fedora/get/tufts:PB.001.001.00007/Archival.pdf

Chrons, O. and Sundell, S. 2011. Digitalkoot: Making Old Archives Accessible Using Crowdsourcing. Human Computation, *Papers from the 2011 AAAI Workshop*. http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3813/4246.

Creutz, M. and Lagus, K. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, *4*(1): Article 3.

Denkowski, M. and Lavie, A. 2014. *Meteor Universal: Language Specific Translation Evaluation for Any Target Language*. Available at https://www.cs.cmu.edu/~alavie/METEOR/pdf/meteor-1.5.pdf

Ford, G., Hauser, S.E., Le, D.X. and Thoma, G.R. 2001. Pattern matching techniques for correcting low confidence OCR words in a known context. *Proceedings of SPIE*, vol. 4307, pp. 24-25.

Gotscharek, A., Reffle, U., Ringlstetter, C. and Schulz, K. U. 2009. On lexical resources for digitization of historical documents. *Proceedings of the 9th ACM symposium on Document engineering,* pp. 193-200, ACM.

Häkkinen, J., and Tian, J. 2001. N-gram and decision tree based language identification for written words. *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'01,* pp. 335-338.

Honkela, T., Hyvärinen, A. and Väyrynen, J.J. 2010. WordICA—Emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, *16*(03), pp. 277-308.

Jansen, B., Spink, A. and Sarasevic, T. 2000. Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. *Information Processing & Management* 36, pp. 207-227.

Kettunen, K. 2014. Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics*, *21*(3), 223-245.

Kohonen, T. 2001. *Self-organizing maps*. Springer.

Kohonen, T. and Honkela, T. 2007. Kohonen network. *Scholarpedia*, *2*(1), p. 1568.

Kukich, K. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, *24*(4), pp. 377-439.

Lindén, Krister, Pursula, Antti, and Onikki-Rantajääskö, Tiina 2013. Integrating Language Resources. A presentation at Sprogmode 2013. Available at http://islenskan.is/Sprogmode2013/forelaesninger/Krister-Linden-FIN-CLARIN%20-%20Integrating%20LRs%20overview.pdf

Lindén, K., Silfverberg, M. and Pirinen, T. 2009. HFST tools for morphology–an efficient open-source package for construction of morphological analyzers. *State of the Art in Computational Morphology*, pp. 28-47.

Paukkeri, M. S., Nieminen, I. T., Pöllä, M., & Honkela, T. 2008. A Language-Independent Approach to Keyphrase Extraction and Evaluation. In *Proceedings of COLING'08 (Posters),* pp. 83-86.

Raitanen, I. 2012. "Etsikäät hywää ja älläät pahaa." Tiedonhakumenetelmien tuloksellisuuden vertailu merkkivirheitä sisältävässä historiallisessa sanomalehtikokoelmassa. Master’s thesis, University of Tampere. Available at https://tampub.uta.fi/handle/10024/59337/browse?value=RAITANEN%2C+ISMO&type=author

Ringlstetter, C., Schulz, K.U. and Mihov, S. 2007. Adaptive text correction with Web-crawled domain-dependent dictionaries. *ACM Transactions on Speech and Language Processing (TSLP)*, *4*(4): Article 9.

Schmitt, J. C. 1991. *U.S. Patent No. 5,062,143*. Washington, DC: U.S. Patent and Trademark Office.

Strohmaier, C., Ringlstetter, C., Schulz, K.U. and Mihov, S. 2003. A visual and interactive tool for optimizing lexical postcorrection of OCR results. *Proceedings of CVPRW'03 , Computer Vision and Pattern Recognition Workshop*, vol. 3, pp. 32-32.

Tanner, S., Muñoz, T. and Ros, P.H. 2009. Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. D-Lib Magazine July/August http://www.dlib.org/dlib/july09/munoz/07munoz.html

Tolstoy, L. 1878. Анна Каренина (Anna Karenina), Moscow.

Torsello, A., Robles-Kelly, A., and Hancock, E. R. 2007. Discovering shape classes using tree edit-distance and pairwise clustering. *International Journal of Computer Vision*, *72*(3), 259-285.

Turian, J.P., Shen, L. and Melamed, I.D. 2003. Evaluation of Machine Translation and its Evaluation. *Proceedings of MT Summit IX.* New Orleans, USA, 23-28 September 2003 http://nlp.cs.nyu.edu/pubs/papers/turian-summit03eval.pdf

Vatanen, T., Väyrynen, J.J., and Virpioja, S. 2010. Language Identification of Short Text Segments with N-gram Models. In Proceedings of *LREC'01*.

Wu, S. and Manber, U. 1992. Agrep—A Fast Approximate Pattern-Matching Tool. *Usenix Winter 1992*, pp. 153-162.