Optical character recognition quality affects subjective user perception of historical newspaper clippings

Received 5 January 2023
Revised 24 April 2023
Accepted 28 April 2023

Kimmo Kettunen
Philosophical Faculty, School of Humanities, University of Eastern Finland, Joensuu, Finland

Heikki Keskustalo and Sanna Kumpulainen
Faculty of Information and Communication Sciences, Tampere University, Tampere, Finland, and

Tuula Pääkkönen and Juha Rautiainen
National Library of Finland, Mikkeli, Finland

Abstract

Purpose – This study aims to identify user perception of different qualities of optical character recognition (OCR) in texts. The purpose of this paper is to study the effect of different quality OCR on users' subjective perception through an interactive information retrieval task with a collection of one digitized historical Finnish newspaper.

Design/methodology/approach – This study is based on the simulated work task model used in interactive information retrieval. Thirty-two users made searches to an article collection of Finnish newspaper Uusi Suometar 1869–1918 which consists of ca. 1.45 million autosegmented articles. The article search database had two versions of each article with different quality OCR. Each user performed six pre-formulated and six self-formulated short queries and evaluated subjectively the top 10 results using a graded relevance scale of 0–3. Users were not informed about the OCR quality differences of the otherwise identical articles.

Findings – The main result of the study is that improved OCR quality affects subjective user perception of historical newspaper articles positively: higher relevance scores are given to better-quality texts.

Originality/value – To the best of the authors' knowledge, this simulated interactive work task experiment is the first one showing empirically that users' subjective relevance assessments are affected by a change in the quality of an optically read text.

Keywords Historical newspapers, OCR quality, Interactive information retrieval, Simulated work task, Evaluation, Finnish

Paper type Article

This work was part of the NewsEye project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 770299. Faculty of Information Technology and Communication Sciences of the Tampere University took part in the arrangement of the query sessions and evaluation of the results as part of the Project EVOLUZ (#326616) financed by the Academy of Finland.

The query environment was implemented by Evident Ltd. (https://evident.fi/).

Journal of Documentation
Vol. 79 No. 7, 2023
pp. 137-156
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-01-2023-0002

1. Introduction

Digitized versions of historical newspaper collections have been both produced and studied extensively during the last few decades in different parts of the world. In the year 2012, it was estimated that digitized newspaper collections in Europe consisted of ca. 129 M pages and 24,000 titles (Dunning, 2012), about 17% of the available content (Gooding, 2018). Since that several national libraries and other stakeholders, such as publishers, have produced and are currently producing more and more digitized historical content online out of their newspaper collections. In a recent publication describing in detail 10 different digitized historical newspaper collections, Beals and Bell (2020) state that

> over the past thirty years, national libraries, universities and commercial publishers around the world have made available hundreds of millions of pages of historical newspapers through mass digitisation and currently release over one million new pages per month worldwide. These have become vital resources not only for academics but for journalists, politicians, schools, and the general public.

It is evident that more collections and more data will be available in the future and usage of the collections will increase.

Several projects have studied and developed methods to access digitized newspaper collections. Out of the projects, only a few can be mentioned here. One of the first large European projects was *Europeana* (http://www.europeana-newspapers.eu/, Neudecker and Antonacopoulos, 2016). Europeana gathered a selection of digitized newspaper content produced in different European libraries together, enhanced searchability of the content, developed an evaluation and quality-assessment infrastructure for newspaper digitization and created best-practice recommendations for newspaper metadata, among other things. Later research projects like *NewsEye* (newseye.eu, Oberbichler *et al.*, 2021), *Impresso* (https://impresso-project.ch/app/) and *Oceanic Exchanges* (https://oceanicexchanges.org/) have studied how to access, develop and enrich the contents of digitized newspaper collections.

In Finland, the most important national project related to the study of digitized historical newspapers has been *Computational History and the Transformation of Public Discourse in Finland, 1640–1910*, which finished at the end of the year 2019. The project used historical newspapers published in Finland during the years 1771–1920 as one of its main research data. One part of the project studied metadata related to the newspapers (Mäkelä *et al.*, 2019; Marjanen *et al.*, 2019), another part of the project produced, e.g. a database out of which reuse of the published news stories could be detected and studied (Salmi *et al.*, 2020). The National Library of Finland (NLF) participated in the project by providing both data and working with improvement of the data and production of optical character recognition (OCR), named entity and article extraction training and evaluation collections (Kettunen and Koistinen, 2019; Kettunen *et al.*, 2019a, b).

In the current study, a collection of 49 years of Finnish newspaper *Uusi Suometar* is used in a user-oriented study of information retrieval and access. The study can be considered as a realistic approach to information retrieval: participants of the study make searches in an optically read historical newspaper collection to find newspaper articles related to given topics. This is what scholars and lay users of these collections do daily in different parts of the world with varying success. In our evaluation, users search the newspaper database both with pre-formulated queries and queries that they formulate on their own based on the topic descriptions. The database of *Uusi Suometar* has two different versions of the content: one with original OCR and one with improved, new OCR. The query engine always searches for the results of queries in the new OCR version of the database and ranks the results according to these. However, retrieved texts presented for reading are balloted in the two different optically read qualities of the same articles. Users of the query system were not aware of differences in the OCR quality when they used the query environment.

During the last decades, vast investments have been made in various national and international archival projects to digitize growing numbers of historical documents and grant end-user access to them. Yet, as recent empirical studies attest (e.g. Kumpulainen and Late, 2022; Late and Kumpulainen, 2021), the use of such collections is prone to be negatively affected by the noise present in digitized documents, caused by the errors made during the OCR process. However, the core question of how the variation in OCR quality affects the user has proved to be evasive. In this paper, we study this question experimentally. The key question is whether a change in the document's OCR quality affects subjective perception regarding the usefulness of the documents (perceived usefulness) when the test person is framed to be performing a simulated research work task.

2. Related research

Even if new ways of access to large, digitized text collections are under development, basic information retrieval tools are still the main entry point to these collections (Organisciak *et al.*, 2021). Users search collections in a query engine interface using keywords to describe their information needs. Information retrieval of historical newspapers has several challenges, among which are, e.g. OCR quality, spelling variation of historical language, lack of proper tools for natural language processing of older language and lack of structure in the optically read documents (Lopresti, 2009; Gotscharek *et al.*, 2011; Piotrowski, 2012; Järvelin *et al.*, 2016; Karlgren *et al.*, 2019; Pfanzelter *et al.*, 2021; Torget *et al.*, 2022). In their present state, historical newspapers are a hard task for information retrieval engines, and users of the collections, such as researchers of digital humanities, are many times not very satisfied with the search possibilities and may have low trust in the search results (Jarlbrink and Snickars, 2017; Pfanzelter *et al.*, 2021).

Textual properties of digitized historical newspapers, such as the quality of OCR and segmentation, are often studied in data-oriented scenarios, which pay attention to the statistical properties of text without consulting the user viewpoint. Effects of subquality OCR on efficiency of information retrieval have been studied to a fair extent, and the results include both simulations, where the quality of the text content has been tampered with artificially and original OCR text. Actual user studies in a controlled query environment, however, have been so far missing. Simulated research settings include, e.g. Taghva *et al.* (1996), Savoy and Naji (2011) and Bazzo *et al.* (2020), just to mention a few. The general result of these studies is that worse OCR quality lowers query results clearly. Most clearly, the effect of worse OCR is with short documents and queries of a few words: with these query, engine has less evidence for matching the document and the query words.

Järvelin *et al.* (2016) report the results of information retrieval in a laboratory-style collection of digitized Finnish newspapers. Their collection consisted of 180,468 documents (84,512 pages of newspapers), for which they had developed 56 search topics with graded relevance assessments. Results of the study show that low-level OCR quality of the collection lowered search results clearly, even if heavy fuzzy-matching methods were used in query expansions to improve the results.

If we broaden the scope and look at research outside information science, digital humanists have also paid attention to the problems of low-level OCR in digital historical newspaper collections. Jarlbrink and Snickars (2017), for example, show how one digital Swedish newspaper collection, Aftonbladet 1830–1862, "*contains extreme amounts of noise: millions of misinterpreted words generated by OCR, and millions of texts re-edited by the auto-segmentation tool*". Their main contribution is a discussion of low-quality OCR and its effects on using digitized newspapers as research data. Pfanzelter *et al.* (2021) describe user experiences and needs of digital humanities researchers with three digitized newspaper collections: Austrian ANNO (https://anno.onb.ac.at/), Finnish Digi

(digi.kansalliskirjasto.fi) and French Gallica (https://gallica.bnf.fr/) and Retronews (https://gallica.bnf.fr/edit/und/retronews-0). Although their main concern in the paper is related to general functionality demands for interfaces of digitized newspaper collections, they report also experiences related to the searchability of the collections. One of their general findings is that "*in some cases, the OCR quality is still very low. After identifying some major issues in this regard, the DH team's reliance on (and trust in) some search results was very low*". Also, slightly differing opinions have been stated by digital humanities researchers. Strange *et al.* (2014), for example, state that "*The cleaning was thus desirable but not essential*" referring to cleaning to correction of OCR errors in the digitized texts they were studying.

van Strien *et al.* (2020) suggest caution in trusting retrieval results of optically read text. They show that both rankings of articles and a number of returned articles from the query engine are affected by the text quality. Traub *et al.* (2018) show that better data quality decreases so-called retrievability bias, which tends to bring certain documents as a search result more often than others (Azzopardi and Vinay, 2008). Chiron *et al.* (2017) show, with respect to the French Gallica collection, that low-frequency query words that contain frequent optical character error patterns have a higher risk to result in poor query results. Results of Hill and Hengchen (2019) on the Eighteenth Century Collections Online (ECCO), on the other hand, show that the effects of poor OCR quality vary with respect to different higher level NLP tasks. They studied topic modelling, collocation analysis, authorial attribution and vector space modelling in a data-oriented setting.

Traub *et al.* (2015) interviewed historians to get an impression of their usage of digital archives and their awareness of possible problems related to them. Researchers were usually aware of OCR quality problems of digitized collections and the possible bias caused by quality problems, but they could not quantify, how the problems could affect their research. Traub *et al.* show also that problems of OCR quality affect different types of research settings differently. They concluded "*that the current knowledge situation on the users' side as well as on the tool makers' and data providers' side is insufficient and needs to be improved*" with respect to the data quality. In the interviews of Korkeamäki and Kumpulainen (2019; Late and Kumpulainen, 2021), historians were asked of their task-based information interaction in digital environments. Some of the interviewed researchers worked a lot with historical optically read texts and some mentioned problems with the quality of data, especially in the analysis phase of the data: "*Especially when using big data, historians had to evaluate the impact of the noisy OCR on the results and if the analysis is worth doing*" (Korkeamäki and Kumpulainen, 2019). Some of the interviewed researchers did not trust that search in the optically read collections would retrieve all the relevant hits as results (Late and Kumpulainen, 2021).

3. Research questions

In this interactive information retrieval study, we look answers for the following research questions.

(1) Do the search results of pre-formulated queries differ from self-formulated queries of the users?

(2) Does different quality of the OCR (old versus new) affect the perceived usefulness of the newspaper clippings? The subquestions are these two:

- What happens with perceived usefulness in the case of pre-formulated queries?
- What happens with perceived usefulness in the case self-formulated queries?

4. The research setting

4.1 Research method

Interactive information retrieval, in general, is a user-centred approach for doing information retrieval research (e.g. Ingwersen and Järvelin, 2005, p. 191). Our research belongs specifically to the tradition of interactive information retrieval, where we use a simulated work task situation approach with its information needs (Borlund, 2000; Ingwersen and Järvelin, 2005, pp. 251–254). Our interactive information retrieval setting uses the three main requirements for this kind of task, as described in Borlund (2000): (1) potential users as test persons, (2) application of dynamic and individual information needs and (3) use of multidimensional and dynamic relevance judgements. Our interactive information retrieval environment developed for the task is the tool that triggers the simulated information needs of the users (cf. Section 4.5).

Interactive information retrieval approach has the following four main advantages (cf., Borlund, 2000; Borlund and Ingwersen, 1998): firstly, the usage of cover stories triggers information needs provoked by simulated work tasks. Secondly, the setting allows individual test persons to assess the usefulness of the newspaper clippings with respect to their own interpretation. Thirdly, using graded and multidimensional relevance assessments instead of binary and topical ones facilitates both control and repeatability during experimentation by combining the use of both static queries and allowing free-form queries. And finally, the setting enables the use of a realistic search interface with actual data.

4.2 Participants of the study and their task

To perform the study, we recruited 32 participants for the evaluation task. The student users for the evaluation task were recruited from the courses *Information Retrieval and Language Technology* and *Information Retrieval Methods* at the Tampere University, Faculty of Information Technology and Communication Sciences. Three teachers of information science also participated in the evaluation task. The choice of the participants was based mainly on the ease of getting a large enough group to perform the tasks, and the group can be considered as a convenience sample (Kelly, 2009, p. 67). A large enough group of real historians would have been harder to gather, but on the other hand, university-level education is an asset for the participants. We did not collect detailed information about the participants, only the information on whether they were students or teachers of information science (cf. Figure 4). Thus, we do not perform any analysis of results based on different user qualities.

The participants, students and teachers of information science were informed in background that they use the information retrieval system of digitized newspaper clippings to write an article on historical events in Finland or the world in the time span of 1869–1918. Participants needed to evaluate the results of the search in relation to this information need.

Participants were given a one-page instruction leaflet which described the information retrieval task. The leaflet described the general idea of the task and informed the participants about two sessions, gave them the backgrounding simulated work task story and explained the graded evaluation scale of 0–3. Participants were guided to perform six queries in both sessions with different topics that the query environment balloted one at a time after the user had logged in to the system. Due to COVID-19 users made searches alone using remote net access to the system, and no common session was arranged. Users were instructed to perform the searches on their own schedule in a timeframe of 10 days beginning from March 8, 2021.

Instructions for the search sessions were given to the participants as one A4 page. The instructions gave the participants a background story and formulated the evaluation scale to use in relation to the simulated work task as described in Table 1.

Table 1.
The background story and evaluation instructions given to the participants

The background story
Imagine that you are writing an article related to topics in history of Finland or world history at the end of 19th century or the beginning of 20th century. Evaluate quality of the clippings you get as search results. Evaluate the quality of each clipping from the viewpoint, how it helps you to proceed with your article writing
Evaluation of the search results (graded relevance scale of 0–3)
3. The clipping deals with the topic very broadly and its information content corresponds well with the task. The clipping helps well in accomplishing your task
2. The clipping deals partially with the task or touches it. The content of the clipping helps to some extent in accomplishing your task
1. The clipping does not deal with the actual topic but helps to find better search terms and to limit the topic somehow. It helps indirectly in accomplishing your task
0. The clipping is wholly off topic and does not even help to formulate new queries. This clipping brings no benefit in accomplishing your task

Source(s): Table by authors

4.3 Topic creation for the study

The topics of the searches were created using history timelines from two popular history encyclopedias: *Suomen historian pikkujättiläinen* ("A small encyclopedia of Finnish history", Zetterberg, 1989) and *Maailmanhistorian pikkujättiläinen ("A small encyclopedia of world history", Zetterberg, 1988)*. After finding suitable topics from the timeline, searches to the newspaper database at digi.kansalliskirjasto.fi were performed to confirm that the database had enough hits related to the topic. During the final creation of the query environment, many original topics were abandoned, and new ones were created due to too few hits in the final article extraction database. Final topic descriptions were based on Finnish Wikipedia articles related to the topics. Out of the topics we created readymade short queries for the task.

The topics cover the time frame of the historical collection of Uusi Suometar, beginning from 1870s and ending in 1918. First mentioned year in the topic descriptions is 1871, last 1918. Topics cover both domestic and foreign news, the share of domestic news being 21, and foreign 9. Demarcation line between foreign and domestic news is not always sharp, some topics could be classified as both.

Participants performed two separate query sessions: one with pre-formulated queries and another one where they could formulate the queries freely based on the topic descriptions, which were available on the user-interface of the search environment. In the second session, participants could use, e.g. Google searches to gather background information about the topics. For each query, only the top 10 results were shown to the users. Pre-formulated Finnish queries and their translations are listed in Appendix.

This study uses a relatively small, digitized collection of one newspaper's 49-year history to study users' access to the collection with means of article extraction on the pages of the collection. Article extraction – many times also called segmentation – on historical newspaper collections is not a pervasive property, as good quality separation of articles has been hard to produce automatically (cf. Clausner *et al.*, 2017; 2019; Dengel *et al.*, 2014; Kise *et al.*, 2014; Jarlbrink and Snickars, 2017). Recent evaluation studies reported in International Conference on Document Analysis and Recognition 2017 and 2019 (Clausner *et al.*, 2017, 2019) show that especially state-of-the-art systems (commercial and open source) used many times by libraries in data production do not perform very well in page segmentation and region classification of complex layout document pages. Among the 10-digitized historical newspaper collections gathered in Beals and Bell (2020), some do have article extraction, but many do not have it. Article extraction in our study collection has been produced automatically based on a machine learning model and the quality of the result has not been assessed on a large scale, only with a small evaluation collection (cf. section *Newspaper data – Optical Character Recognition quality and article structure*).

Our data has two main challenges for the user, as will be described in more detail later. Shortly put, the OCR quality of the data is not optimal even if it has been improved, and the results of article extraction will make an evaluation of query results hard – the texts shown to the users may contain irrelevant content from the surrounding context of the newspaper page. These kinds of problems are, however, many times a reality with digitized historical newspapers, as testified, for example, in digital humanities pilot studies of three different digitized newspaper collections in Pfanzelter *et al.* (2021).

We conduct the study in the framework of interactive information retrieval, but our results are also relevant for the expanding field of digital humanities, which increasingly uses digitized historical newspapers as its research data. We thus try to include some implications for digital humanities in our discussion of the findings of our research.

4.4 Newspaper data – optical character recognition quality and article structure

Our search collection consists of the whole history of *Uusi Suometar* 1869–1918, ca. 86,000 pages and 306.8 million words (Kettunen and Koistinen, 2019). Uusi Suometar was at the time of its publication one of the most important Finnish language newspapers in Finland, where newspapers were published in two languages, Finnish and Swedish (Hynynen, 2019). The newspaper data has been scanned and optically read as part of the digitization of Finnish newspapers and journals at the NLF and is part of the digitized content offered by the NLF (https://digi.kansalliskirjasto.fi/etusivu?set_language=en).

The original OCR for *Uusi Suometar* was performed using a line of ABBYY FineReader® products. The quality of the digitization of the whole collection of Finnish newspapers from 1771 to 1910 has been estimated by Kettunen and Pääkkönen (2016). They conclude that ca 70–75% of the words in the Finnish language 2.4-billion-word index database could be recognized by using automatic morphological analysers. In general, this level of word-level quality can be considered quite typical for older digitizations of historical newspaper collections (cf. e.g. Tanner *et al.*, 2009).

Improved OCR for the whole history of *Uusi Suometar* was achieved with Tesseract v.3.0.4.01. Improvement to the earlier quality in recognition of words is approximately 15% units as a mean over the whole period. On average 83% of the words in the newspaper were recognized with automatic morphological analysers, and the recognition rate varied from ca 78–88% over the 49 years. For the old OCR mean word recognition rate was 68.2% (Kettunen and Koistinen, 2019). Even if the improvement in OCR quality is considerable, the improved quality can still be challenging for information retrieval engines, especially with short queries and articles, where the information retrieval engine has less evidence for matching the query words and collection data in the engine's index (Järvelin *et al.*, 2016; Mittendorf and Schäuble, 2000). In a recent study, Bazzo *et al.* (2020), for example, found that statistically significant degradation of search results begins already at the word error rate of 5%, when Portuguese pdf texts with artificially induced errors were sought for with state-of-the-art modern query engine and algorithms. The same kind of results has been achieved in most of the current studies, as a current survey paper of Nguyen *et al.* (2021) shows. It should also be emphasized that even if substantially better-quality OCR for collections can be achieved with the latest tools like Transkribus (Muehlberger *et al.*, 2019), new OCR of large old historical collections is tedious and time-consuming – thus the problem of low-level optically read text will stay for some time.

Newspaper data at the NLF was originally scanned and recognized page by page without article structure information besides the basic layout of the pages. For this study, we used articles that were extracted automatically from the pages of *Uusi Suometar* using a trained machine learning model with the software PIVAJ (Hebert *et al.*, 2014a, b; Kettunen *et al.*, 2019a). Training of the PIVAJ model was based on 168 pages of manually marked data that had a different number of columns (varying from 3 to 9). Kettunen *et al.* (2019a) reported success percentages of 67.9, 76.1 and 92.2 for an evaluation dataset of 56 pages in three different evaluation scenarios based on Clausner *et al.* (2011) using layout evaluation software from PRImA (https://www.primaresearch.org/). In the automatic segmentation process, the collection of *Uusi Suometar* was divided into 1,459,068 articles with PIVAJ. Figure 1 shows an example of PIVAJ's graphical output for the article structure. Different articles on the page are shown with different colours.

In the article extraction of the whole history of Uusi Suometar, article separation is far from optimal, and articles are perhaps best called automatically extracted clippings with varying lengths. In the search evaluation task, these clippings are documents that users search and evaluate. It should be emphasized that the article segmentation that was producible for the whole history of Uusi Suometar is experimental and its quality will bring one layer of difficulty to the evaluation of search results. As Jarlbrink and Snickars (2017) formulate it, autosegmentation tools create random texts, and borders of text snippets are fuzzy. This feature was informed to the users in the instructions of the search task.

4.5 The query interface

Participants of the evaluation task performed their task using the query engine *Elastic search* (https://www.elastic.co/), version 7.3.2, which is the background engine of the library's presentation system. Queries were performed in AND mode, where every query term is sought for in the documents. Hits of the search engine shown for the users needed to be at least 500 characters long to avoid very short text passages which would be hard to evaluate. The index of the newspaper collection's database is lemmatized, i.e. it contains base forms of the words, which is crucial as Finnish is a highly inflected language (Järvelin *et al.*, 2016). Search environment and database out of the Uusi Suometar article extraction results were developed by Evident Solutions Ltd (https://evident.fi/).

Figure 2 shows the search environment we created for the evaluation of the search results. Figure 3 depicts the query interface the participants used.

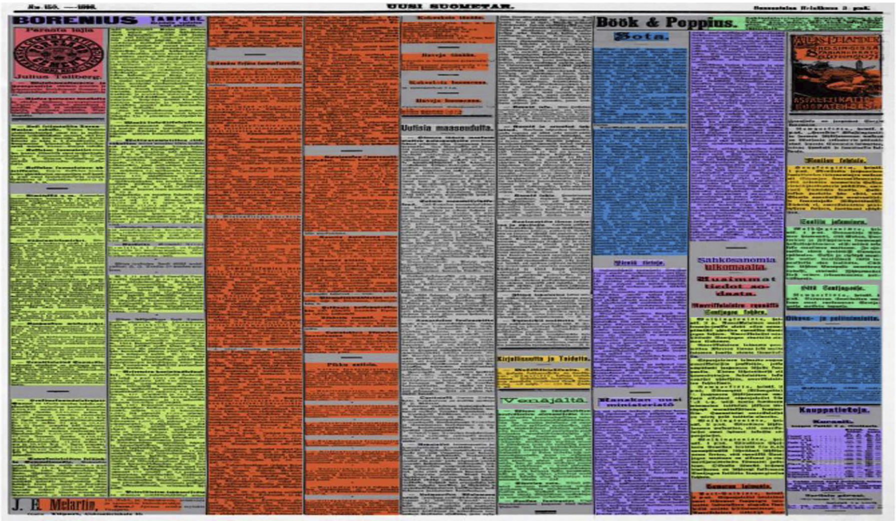
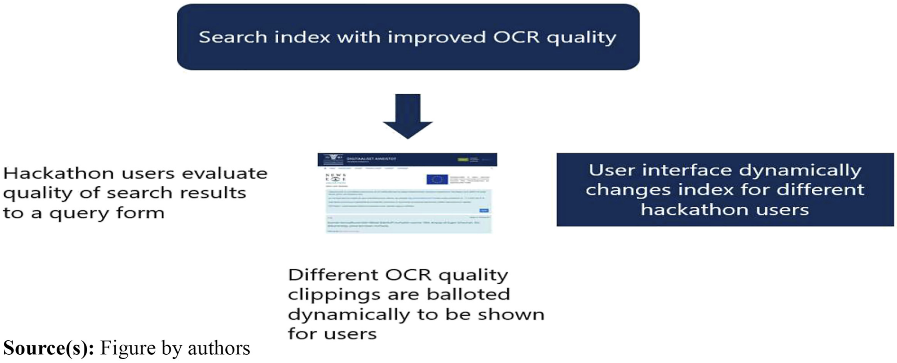


Figure 1. Automatically produced article structure of one page from Uusi Suometar from the graphical output of PIVAJ

Source(s): Figure by authors

Figure 3 shows the query interface after a pre-formulated query has been performed and 10 results retrieved. Text on the blue background on the top describes the topic and shows the pre-formulated query beneath in pink. The light purple rectangle below shows the beginning of the first query result. Grading buttons are on the right side of the rectangle. Underneath the text snippet of the result on the left is the button for opening the clipping in its whole. The button also shows the character length of the clipping. Matches of the query words are highlighted in the snippet view and in the actual clipping view which the participants used for evaluating the relevance of the clippings.



Search index with improved OCR quality

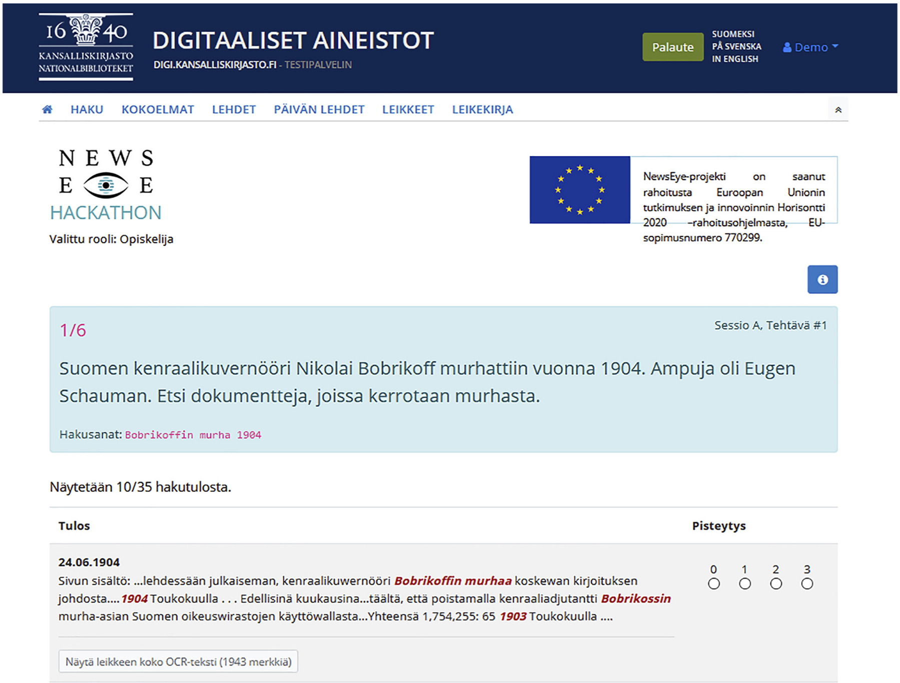
Hackathon users evaluate quality of search results to a query form

User interface dynamically changes index for different hackathon users

Different OCR quality clippings are balloted dynamically to be shown for users

Source(s): Figure by authors

Figure 2.
Index side view: improved Optical Character Recognition version of text is always used in the retrieval phase



DIGITAALISET AINEISTOT

DIGI.KANSALLISKIRJASTO.FI - TESTIPALVELIN

Valittu rooli: Opiskelija

NewsEye-projekti on saanut rahoitusta Euroopan Unionin tutkimuksen ja innovoinnin Horisontti 2020 -rahoitusohjelmasta, EU-sopimusnumero 770299.

1/6

Sessio A, Tehtävä #1

Suomen kenraalikuvernööri Nikolai Bobrikoff murhattiin vuonna 1904. Ampuja oli Eugen Schauman. Etsi dokumentteja, joissa kerrotaan murhasta.

Hakusanat: `Bobrikoffin murha 1904`

Näytetään 10/35 hakutulosta.

Tulos	Pisteytys
24.06.1904 Sivun sisältö: ...lehdessään julkaiseman, kenraalikuwernööri ***Bobrikoffin murhaa*** koskewan kirjoituksen johdosta.... ***1904*** Toukokuulla . . . Edellisinä kuukausina...täältä, että poistamalla kenraaliadjutantti ***Bobrikossin*** murha-asian Suomen oikeuswirastojen käyttöwallasta...Yhteensä 1,754,255: 65 ***1903*** Toukokuulla .... Näytä leikkeen koko OCR-teksti (1943 merkkiä)	0 ○ 1 ○ 2 ○ 3 ○

Source(s): Figure by authors

Figure 3.
Screenshot of the query interface

4.6 Collection and analysis of the query results

Query results of the user sessions were collected for analysis in a query log, which is shown in Figure 4.

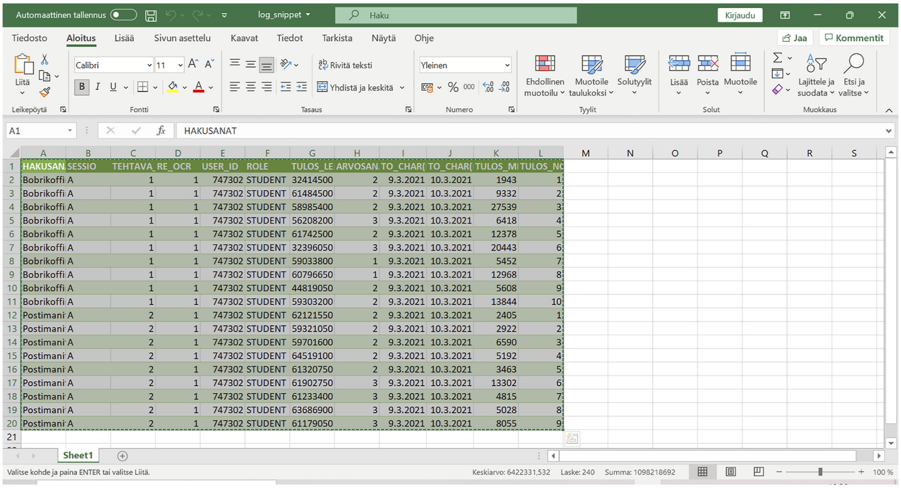
The columns in the query log indicate the following data beginning from the left: (1) query words; (2) session: pre-formulated or self-formulated queries, A and B, respectively; (3) number of the topic; (4) OCR quality in the results (0 for the old and 1 for the new); (5) user id; (6) role of the user (student or teacher); (7) id number of the result clipping; (8) user-given evaluation result on the scale of 0–3; (9) date and time of the performance; (10) possible change for the evaluation: time stamp; (11) size of the clipping in characters; and (12) rank (1–10) of the result clipping in the result list.

The interactive information retrieval system balloted the topics for each user so that out of the 32 users' work we obtained 3,893 evaluations. There were 1,861 evaluations of pre-formulated queries and 2,032 evaluations of self-formulated queries. Differences in these numbers are due to the fact that part of the users did not perform all their tasks, and some did more than that required: if 32 users had each performed 120 evaluations (2 sessions and 10 evaluations for 6 queries), the log should contain 3,840 evaluations evenly divided to the 2 different sessions.

Analysis of the self-formulated queries showed that users formulated slightly longer queries than the pre-formulated queries were. The mean length of the pre-formulated queries is 2.87 words, and the mean length of the self-formulated queries is 3.15 words, which is a 9.8% increase. In general, the queries are short and web-like (cf. Jansen *et al.*, 2000). The clippings the users evaluated were of varying lengths. We had set a minimum length of 500 characters for the results to be shown for users, but no maximum length. The mean length of the clippings in all the results of the two sessions was 6,116 characters.

Some type of query structuring was used very little in users' self-formulated queries: 111 queries used AND connector and 90 queries connector OR. Combined AND and OR was used in 20 queries. Truncation operators ? and * were both used 43 times in 33 queries.

We assumed that the user interface would take care of the number of queries and evaluations each user finished. However, some of the users did not finish all the queries or evaluations in the pre-formulated queries' session because the possibility of a user's



A1 HAKUSANAT

	A	B	C	D	E	F	G	H	I	J	K	L
1	HAKUSAN	SESSIO	TEHTAVA_	RE_OCR	USER_ID	ROLE	TULOS_LE	ARVOSAN	TO_CHAR(	TO_CHAR(	TULOS_MI	TULOS_NC
2	Bobrikoffi	A	1	1	747302	STUDENT	32414500	2	9.3.2021	10.3.2021	1943	1
3	Bobrikoffi	A	1	1	747302	STUDENT	61484500	2	9.3.2021	10.3.2021	9332	2
4	Bobrikoffi	A	1	1	747302	STUDENT	58985400	2	9.3.2021	10.3.2021	27539	3
5	Bobrikoffi	A	1	1	747302	STUDENT	56208200	3	9.3.2021	10.3.2021	6418	4
6	Bobrikoffi	A	1	1	747302	STUDENT	61742500	2	9.3.2021	10.3.2021	12378	5
7	Bobrikoffi	A	1	1	747302	STUDENT	32396050	3	9.3.2021	10.3.2021	20443	6
8	Bobrikoffi	A	1	1	747302	STUDENT	59033800	1	9.3.2021	10.3.2021	5452	7
9	Bobrikoffi	A	1	1	747302	STUDENT	60796650	1	9.3.2021	10.3.2021	12968	8
10	Bobrikoffi	A	1	1	747302	STUDENT	44819050	2	9.3.2021	10.3.2021	5608	9
11	Bobrikoffi	A	1	1	747302	STUDENT	59303200	2	9.3.2021	10.3.2021	13844	10
12	Postimani	A	2	1	747302	STUDENT	62121550	2	9.3.2021	10.3.2021	2405	1
13	Postimani	A	2	1	747302	STUDENT	59321050	2	9.3.2021	10.3.2021	2922	2
14	Postimani	A	2	1	747302	STUDENT	59701600	2	9.3.2021	10.3.2021	6590	3
15	Postimani	A	2	1	747302	STUDENT	64519100	2	9.3.2021	10.3.2021	5192	4
16	Postimani	A	2	1	747302	STUDENT	61320750	2	9.3.2021	10.3.2021	3463	5
17	Postimani	A	2	1	747302	STUDENT	61902750	3	9.3.2021	10.3.2021	13302	6
18	Postimani	A	2	1	747302	STUDENT	61233400	3	9.3.2021	10.3.2021	4815	7
19	Postimani	A	2	1	747302	STUDENT	63686900	3	9.3.2021	10.3.2021	5028	8
20	Postimani	A	2	1	747302	STUDENT	61179050	3	9.3.2021	10.3.2021	8055	9

Valitse kohde ja paina ENTER tai valitse Liitä.

Keskiarvo: 6422331,532 Laske: 240 Summa: 1098218692

Source(s): Figure by authors

Figure 4.
Screenshot of the query log

premature quitting was not taken care of in the system. In the self-formulated queries' session, the users could edit their queries and resubmit them after they had performed them once, and the query log stored all the results for the same query with possible query variations. These issues were unexpected and should be taken care of in the user interface and instructions.

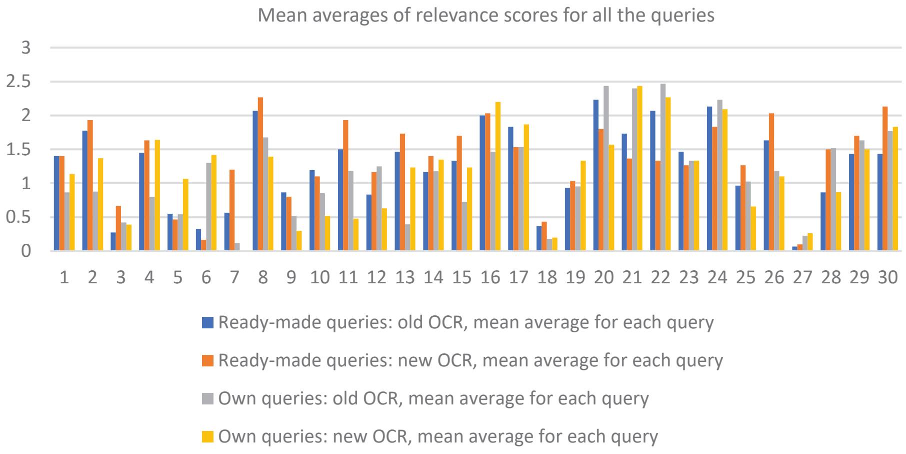
5. Results

As the example of the log shows, the result set is multifaceted and could be analysed in many ways. In this study, we concentrate on the analysis of the main results of the query sessions to be able to answer our research questions stated in the second section of the study.

Figure 5 shows mean averages for evaluations of the individual queries in both sessions with different quality OCR.

Mean averages for the relevance scores over the whole query set are shown in Table 2.

From Figure 5 and Table 2, we can see that especially pre-formulated queries benefited from improved OCR. The mean average evaluation score for the improved OCR in the pre-formulated queries is 7.93% higher than with the old OCR. In self-formulated queries, the difference is clearly smaller: 1.71%. Three queries, #3, #18 and #27, obtained low evaluations in all the scenarios. Two queries, #6 and #7, obtained low evaluations in part of the scenarios: #6 with pre-formulated queries with both OCR qualities and #7 with self-formulated queries with both OCR qualities.



Source(s): Figure by authors

Figure 5. Mean averages of relevance scores for the top-10 clippings retrieved for all topics: graded relevance scale of 0–3 was used

Pre-formulated queries, old optical character recognition: Mean average for evaluations of the query set	Pre-formulated queries, new optical character recognition: Mean average for evaluations of the query set	Self-formulated queries, old optical character recognition: Mean average for evaluations of the query set	Self-formulated queries, new optical character recognition: Mean average for evaluations of the query set
1.26	1.36	1.17	1.19

Source(s): Table by authors

Table 2. Mean averages for the relevance scores over the whole query set: graded relevance scale of 0–3 was used

Inspection of query-per-query results shows that pre-formulated queries gain better mean evaluation scores in 19 cases out of 30. There is one tie and 10 queries, where evaluations of old OCR gain better mean evaluation scores. With self-formulated queries, the result is more even: new OCR results gain better mean evaluation scores in 15 cases, and old OCR results gain better mean evaluation scores in 14 cases; there is one tie. This is depicted in detail in Figure 6.

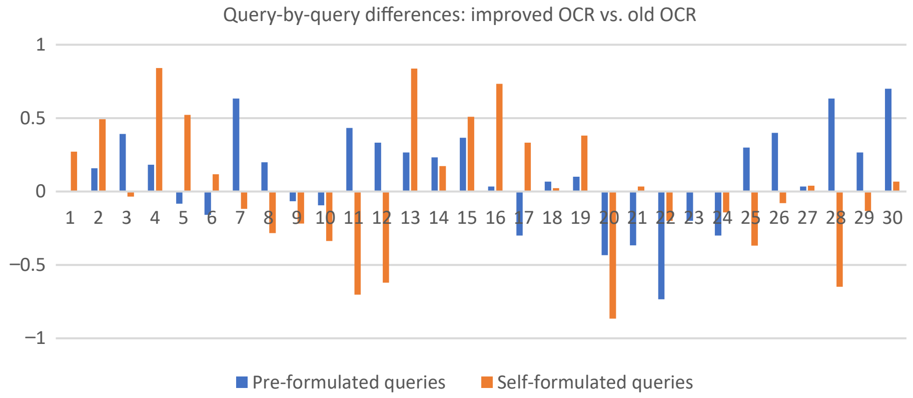
The mean length of the clippings retrieved by the search engine in the two sessions differs quite a lot. With pre-formulated queries the mean length of the clippings was 5,467 characters, and with self-formulated queries 6,711 characters, which is a 22.75% difference. It is thus possible that the longer result clippings of the self-formulated queries were harder to evaluate for users and this is reflected in the lower evaluation results, as was seen in Figure 5 and Table 2. With longer clippings, the users may get tired or frustrated, which may lower their evaluations. The longer clippings may also be fuzzier than the shorter ones and contain text from adjacent text segments. This aspect would need further study.

Wilcoxon's signed rank test (Croft *et al.*, 2010) showed that, in the case of the pre-formulated queries, the improved OCR quality resulted in statistically significantly higher relevance judgments scores compared to the lower quality OCR at the level of assessed individual document pairs ($p = 0.002$). However, when two OCR quality situations were compared at the level of topics (N = 30) via mean average of cumulated gain values for top-10 documents retrieved, the difference was not statistically significant in Wilcoxon's signed rank test ($p = 0.10$).

In the case of the self-formulated queries' session, statistically significant differences were not observed. Self-formulated queries retrieved almost four times larger set of unique documents for all user searchers than pre-formulate queries; these queries also differed to some extent from the pre-defined queries (they were slightly longer), and they returned longer documents. However, we did not study these differences and effects deeper.

6. Discussion

Before discussing our results further, we would like to emphasize key issues of the work and developments that made the present study possible. As Oberbichler *et al.* (2021) indicate, it is vital to successfully negotiate both the challenges and opportunities present in typical interdisciplinary research settings. In the case of our research setting, several theoretical and pragmatic (resource-related) obstacles were present, and advanced work with the collection



Source(s): Figure by authors

Figure 6.
Query-by-query differences of relevance scores' mean averages for the top-10 clippings: graded relevance scale of 0–3 was used

was not self-evident. In general, we have utilized NLF's document presentation system, different OCR software, an experimental page segmentation software and a search engine. The developed realistic web query interface and user logging needed to be adapted to this environment. Only the cumulation of much previous work made this study possible.

Firstly, experienced OCR quality was the main target of the study instead of studying effects of data-oriented OCR quality, which has been repeated probably tens of times in different collections and languages (e.g. Järvelin *et al.*, 2016; Bazzo *et al.*, 2020). As Kumpulainen and Late (2022) show empirically, noise in OCR quality disturbs researchers of historical newspaper collections both in the searching and selection phases of their research. In general, bad OCR quality concerns all users of the collections, although the quality may manifest differently to different user groups. Secondly, simulated interactive work task method introduced by Borlund (2000) was used for the gathering of evaluation data. The method is crucial in gaining insight into the measurement of experienced effects. Thirdly, many years of earlier background work in newspaper digitization and development work at the NLF enabled us to even think of the possibility of this kind of work. Fourthly, we needed to set up the user interface for the query system, design the collection of feedback data and recruit participants for the simulated work task. Out of these pieces, we created a whole that can be used as a general model for similar future studies.

To the best of our knowledge, this is the first study showing empirically, based on simulated work task situations, that the subjective relevance assessments of the test persons were affected by the change of quality of the optically read text presented to them. Earlier studies on the effects of OCR quality have been performed in data-oriented settings, often using laboratory-style tests and artificially tampered data or describing subjective experiences of users regarding the effects of OCR quality on their work. Our study presents a unique approach and a re-useable methodology, if parallel collections of clearly different OCR quality are available for empirical user evaluation.

The answer to our first research question was that pre-formulated and self-formulated queries differ in their results. Self-formulated queries of users were slightly longer than the pre-formulated queries that had been created for the topics. The result documents of queries in the two different query sessions differed clearly in length and number. It is possible that the longer result documents of the self-formulated queries were harder to evaluate for users, and they could also contain more text from the adjacent text segments, which would make their evaluation more difficult. This aspect would need a more detailed comparison between the result clippings of the two sessions.

The answer to our second research question and its first subquestion is clearly positive. With pre-formulated queries, the improved OCR quality clippings gained a 7.93% better mean relevance score, and the difference to the basic level OCR quality evaluations was statistically significant. The answer to the second subquestion was not as clear. With self-formulated queries, we found a small difference of 1.71% in the evaluations in favour of improved OCR. However, this difference was not statistically significant.

Our query environment implementation for the evaluation of two optically read text qualities is the first version of the system. As such it works well, but experience from user sessions showed that it has features that could be developed.

A clear limitation of our research is the assumption that the user interface would take care of the number of queries and evaluations each user finished. However, some of the users did not finish all the queries or evaluations in the pre-formulated queries' session because the possibility of a user's premature quitting was not taken care of in the system. Another limitation is that in the self-formulated queries' session, the users could edit their queries after they had performed them, and the query log stored all the results for the same query with possible query variations. These user behaviours were unexpected and should be taken care of in developing the user interface and instructions. In Kettunen *et al.* (2022), we have analysed our query environment and its components from the point of view of re-usability and suggest improvements to the user-interface's functionality.

Another limitation in the research was that we did not consider how display qualities affected reading or compare reading or proofreading speed between computer screen and paper, which have been studied by, e.g. Gould *et al.* (1986, 1987) and Dillon (1992). Current studies like Köpper *et al.* (2016), Ocal *et al.* (2022) and a recent literature review by Vitello (2022) show that there is no real difference in reading speed or error correction when modern TFT-LCD screens and paper are compared. COVID-19 made it impossible to arrange the query sessions in a computer class, and all the participants performed their assessments using their own computers and screens. It would be advisable that evaluation sessions are arranged in a computer class where equipment is standard, and especially computer screens would be identical.

Our user group in the study was also not optimal, as it consisted mainly of students. A large enough group of historians, however, would have been hard to recruit as, e.g. experience of Kumpulainen and Late (2022) shows. Kumpulainen and Late examined how researchers used digital newspaper collections in their work and what were obstacles they encountered using the collections. They were able to find 13 participants who use Finnish historical newspapers in their work. The usage of university-level students in our evaluation task can be considered motivated: an evaluation task which includes historical information needs a higher level educational background. But as Kelly emphasizes (2009, p. 69), we cannot over-generalize our results based on the convenience sample.

We also had differences in the two query types. In the pre-formulated queries' session, the search was run in Elastic's AND mode, where all the query words were sought for. In the self-formulated queries' session, users could use query operators AND, OR and NOT, if they wished. Users' self-formulated queries were also slightly longer than pre-formulated queries. The effect of these differences on the evaluation results is hard to establish conclusively. Evidently, it would be more consistent to instruct users not to use any operators in the self-formulated queries' session and instead run the search in AND mode.

The clippings the users evaluated were of varying lengths. The mean length of the clippings in all the results of the two sessions was 6,116 characters. With pre-formulated queries, the mean length of the clippings was 5,467 characters, and with self-formulated queries 6,711 characters – a 22.75% difference. It is thus possible that the longer clipping results of the self-formulated queries were harder to evaluate for users and this is reflected in the lower evaluation results. With longer clippings, the users may get tired or frustrated, which may lower their evaluations. Longer clippings may also be fuzzier than shorter ones in their content.

One possible development issue for the evaluation could be the evaluation of the clippings' overall textual and segmentation quality by the users. Our article segmentation for the collection is experimental, and many of the clippings may be quite hard to read due to fuzzy boundaries: the clippings may contain text from adjacent segments, which affects evaluations. A combination of relevance evaluation scores and users' scores for the quality of the clipping boundaries might bring new insights to the evaluation and reveal the usability of the optically read texts better than the usage of traditional relevance assessments only.

The well-known simulated work tasks used in interactive information retrieval have been used in this study to answer the question of optical character recognition quality's effect on the relevance evaluation of retrieval results in a Finnish historical newspaper collection. We have shown that improvement in OCR quality of documents leads to higher mean relevance evaluation scores in a simulated work task scenario. This means that the perceived usefulness of historical newspaper clippings increases with better OCR quality. Our results were achieved with one language, one specific collection and with one user group, but our method is generalizable to any language and can be evaluated with further users and different collections. As our results are achieved in one collection and one language, further studies with different collections and languages should be conducted to strengthen our results. So far studies like Kumpulainen and Late (2022) and Late and Kumpulainen (2021) have raised subjective concerns of digital humanities researchers about the OCR quality of the historical newspaper collections and its effect on their work progress.

Our results should be seen both in the context of information retrieval and the requirements of digital humanities scholars and lay users of the collections. These results bring more weight to both higher quality document needs of digital humanists and efforts of improving the quality of optical character recognition with new developments in software. Better quality of optically read historical documents should be strived for both for the sake of research and lay users.

References

Azzopardi, L. and Vinay, V. (2008), "Retrievability: an evaluation measure for higher order information access tasks", *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*, New York, NY, USA, Association for Computing Machinery, pp. 561-570, doi: 10.1145/1458082.1458157.

Bazzo, G.T., Lorentz, G.A., Suarez Vargas, D. and Moreira, V.P. (2020), "Assessing the impact of OCR errors in information retrieval", in Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J and Martins, F. (Eds), *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science*, Springer, Cham, Vol. 12036, pp. 102-109.

Beals, M.H. and Bell, E., with contributions by Ryan Cordell, Paul Fyfe, Isabel Galina Russell, Tessa Hauswedell, Clemens Neudecker, Julianne Nyhan, Sebastian Padó, Miriam Peña Pimentel, Mila Oiva, Lara Rose, Hannu Salmi, Melissa Terras, and Lorella Viola (2020), "The Atlas of digitised newspapers and metadata: reports from oceanic Exchanges. Loughborough: 2020", doi: 10.6084/m9.figshare.11560059.

Borlund, P. (2000), "Experimental components for the evaluation of interactive information retrieval systems", *Journal of Documentation*, Vol. 56 No. 1, pp. 71-90.

Borlund, P. and Ingwersen, P. (1998), "Measures of relative relevance and ranked half-life: performance indicators for interactive IR", *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval August 1998*, pp. 324-331, doi: 10.1145/290941.291019.

Chiron, G., Doucet, A., Coustaty, M., Visani, M. and Moreux, J. (2017), "Impact of OCR errors on the use of digital libraries: towards a better access to information", *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Toronto, ON, 2017, pp. 1-4, doi: 10.1109/JCDL.2017.7991582.

Clausner, C., Pletshacher, S. and Antonacopoulos, A. (2011), "Scenario driven in-depth performance evaluation of document layout analysis methods", *2011 International Conference on Document Analysis and Recognition (ICDAR)*, doi: 10.1109/ICDAR.2011.282.

Clausner, C., Pletshacher, S. and Antonacopoulos, A. (2017), "ICDAR2017 competition on recognition of documents with complex layouts – RDCL2017", available at: https://ieeexplore.ieee.org/document/8270160

Clausner, C., Antonacopoulos, A. and Pletschacher, S. (2019), "ICDAR2019 competition on recognition of documents with complex layouts – RDCL2019", *Proceedings of the 15th International Conference on Document Analysis and Recognition (ICDAR2019)*, Sydney, Australia, 2019, September, pp. 1521-1526.

Croft, W.B., Metzler, D. and Strohman, T. (2010), *Search Engines. Information Retrieval in Practice*, Pearson, Boston.

Dengel, A. and Shafait, F. (2014), "Analysis of the logical layout of documents", in Doerman, D. and Tombre, K. (Eds), *Handbook of Document Image Processing and Recognition*, Springer, London, pp. 177-222.

Dillon, A. (1992), "Reading from paper versus screens: a critical review of the empirical literature", *Ergonomics*, Vol. 35 No. 10, pp. 1297-1326, doi: 10.1080/00140139208967394.

Dunning, A. (2012), "European newspaper survey report", available at: http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf (accessed 15 December 2022).

Gooding, P. (2018), *Historic Newspapers in the Digital Age. Search All about it!*, Routledge, New York.

Gotscharek, A., Reffle, U., Ringlstetter, C., Schulz, K. and Neumann, A. (2011), "Towards information retrieval on historical document collections: the role of matching procedures and special lexica", *International Journal on Document Analysis and Recognition*, Vol. 14, pp. 159-171, doi: 10.1007/s10032-010-0132-6.

Gould, J.D., Alfaro, L., Finn, R., Haupt, B. and Minuto, A. (1986), "Why reading was slower from CRT displays than from paper", *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface (CHI '87)*, New York, NY, USA, Association for Computing Machinery, pp. 7-11, doi: 10.1145/29933.30853.

Gould, J.D., Alfaro, L., Finn, R., Haupt, B. and Minuto, A. (1987), "Reading from CRT displays can be as fast as reading from paper", *Human Factors*, Vol. 29 No. 5, pp. 497-517.

Hebert, D., Palfray, T., Nicolas, T., Tranouez, P. and Paquet, T. (2014a), "PIVAJ: displaying and augmenting digitized newspapers on the web experimental feedback from the "Journal de Rouen" collection", *Proceeding DATeCH 2014 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, New York, NY, USA, Association for Computing Machinery, pp. 173-178, doi: 10.1145/2595188.2595217.

Hebert, D., Palfray, T., Nicolas, T., Tranouez, P. and Paquet, T. (2014b), "Automatic article extraction in old newspapers digitized collections", *Proceeding DATeCH 2014 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, New York, NY, USA, Association for Computing Machinery, pp. 3-8, doi: 10.1145/2595188.2595195.

Hill, M.J. and Hengchen, S. (2019), "Quantifying the impact of dirty OCR on historical text analysis: eighteenth Century Collections Online as a case study", *Digital Scholarship in the Humanities*, Vol. 34 No. 4, pp. 825-843, doi: 10.1093/llc/fqz024.

Hynynen, M.-L. (2019), "Building a bilingual nation", available at: https://www.newseye.eu/blog/news/building-a-bilingual-nation/ (accessed 15 December 2022).

Ingwersen, P. and Järvelin, K. (2005), *The Turn. Integration of Information Seeking and Retrieval in Context*, Springer, Dordrecht.

Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M. and Kettunen, K. (2016), "Information retrieval from historical newspaper collections in highly inflectional languages: a query expansion approach", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 12, pp. 2928-2946.

Jansen, B.J., Spink, A. and Saracevic, T. (2000), "Real life, real users, and real needs: a study and analysis of user queries on the Web", *Information Processing and Management*, Vol. 36 No. 2, pp. 207-227, doi: 10.1016/S0306-4573(99)00056-4.

Jarlbrink, J. and Snickars, P. (2017), "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive", *Journal of Documentation*, Vol. 73 No. 6, pp. 1228-1243, doi: 10.1108/JD-09-2016-0106.

Karlgren, J., Hedlund, T., Järvelin, K., Keskustalo, H. and Kettunen, K. (2019), "The challenges of language variation in information access", in Ferro, N. and Peters, C. (Eds), *From Multilingual to Multimodal: The Evolution of CLEF over Two Decades. Lessons Learned from 20 Years of CLEF*, Springer, Switzerland, pp. 201-216.

Kelly, D. (2009), "Methods for evaluating interactive information retrieval systems with users", *Foundations and Trends® in Information Retrieval*, Vol. 3 Nos 1-2, pp. 1-224, doi: 10.1561/1500000012.

Kettunen, K. and Koistinen, M. (2019), "Open source Tesseract in Re-OCR of Finnish Fraktur from 19th and early 20th century newspapers and journals – collected notes on quality improvement", DHN2019, available at: https://ceur-ws.org/Vol-2364/25_paper.pdf

Kettunen, K. and Pääkkönen, T. (2016), "Measuring lexical quality of a historical Finnish newspaper collection – analysis of garbled OCR data with basic language technology tools and means. LREC 2016", *Tenth International Conference on Language Resources and Evaluation*, available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf

Kettunen, K., Ruokolainen, T., Liukkonen, E., Tranouez, P., Anthelme, D. and Paquet, T. (2019a), "Detecting articles in a digitized Finnish historical newspaper collection 1771-1929: early results using the PIVAJ software", in *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019)*, Association for Computing Machinery, New York, NY, pp. 59-64, doi: 10.1145/3322905.3322911.

Kettunen, K., Pääkkönen, T. and Liukkonen, E. (2019b), "Clipping the page – automatic article detection and marking software in production of newspaper clippings of a digitized historical journalistic collection", in Doucet, A., Isaac, A., Golub, K., Aalberg, T. and Jatowt, A. (Eds), *TPDL 2019, LNCS 11799*, Springer Cham, Switzerland, pp. 356-360, doi: 10.1007/978-3-030-30760-8.

Kettunen, K., Keskustalo, H., Larsen, B., Pääkkönen, T. and Rautiainen, J. (2022), "Reusing the model and components of an IIR study for perceived effects of OCR quality change. BIIRRR 2022", *Third Workshop on Building towards Information Interaction and Retrieval Resources Re-use*, doi: 10.5281/zenodo.6513586.

Kise, K. (2014), "Page segmentation techniques in document analysis", in Doerman, D. and Tombre, K. (Eds), *Handbook of Document Image Processing and Recognition*, Springer, London, pp. 135-175.

Köpper, M., Mayr, S. and Buchner, A. (2016), "Reading from computer screen versus reading from paper: does it still make a difference?", *Ergonomics*, Vol. 59 No. 5, pp. 615-632, doi: 10.1080/00140139.2015.1100757.

Korkeamäki, L. and Kumpulainen, S. (2019), "Interacting with digital documents: a real life study of historians' task processes, actions and goals", *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*, New York, NY, USA, Association for Computing Machinery, pp. 35-43, doi: 10.1145/3295750.3298931.

Kumpulainen, S. and Late, E. (2022), "Struggling with digitized historical newspapers: contextual barriers to information interaction in history research activities", *Journal of the Association for Information Science and Technology*, Vol. 73 No. 7, pp. 1012-1024, doi: 10.1002/asi.24608.

Late, E. and Kumpulainen, S. (2021), "Interacting with digitised historical newspapers: understanding the use of digital surrogates as primary sources", *Journal of Documentation*, Vol. 78 No. 7, pp. 106-124, doi: 10.1108/JD-04-2021-0078.

Lopresti, D. (2009), "Optical character recognition errors and their effects on natural language processing", *International Journal on Document Analysis and Recognition*, Vol. 12, pp. 141-151, doi: 10.1007/s10032-009-0094-8.

Mäkelä, E., Tolonen, M., Marjanen, J., Kanner, A., Vaara, V. and Lahti, L. (2019), "Interdisciplinary collaboration in studying newspaper materiality", in Krauwer, S. and Fišer, D. (Eds), *Proceedings of the Twin Talks Workshop, co-located with Digital Humanities in the Nordic Countries (DHN 2019), Aachen: CEUR Workshop Proceedings*, Vol. 2365, pp. 55-66, available at: http://ceur-ws.org/Vol-2365/07-TwinTalks-DHN2019_paper_7.pdf

Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L. and Tolonen, M. (2019), "A national public sphere? Analyzing the language, location, and form of newspapers in Finland, 1771-1917", *Journal of European Periodical Studies*, Vol. 4 No. 1, pp. 54-77, doi: 10.21825/jeps.v4i1.10483.

Mittendorf, E. and Schäuble, P. (2000), "Information retrieval can cope with many errors", *Information Retrieval*, Vol. 3 No. 3, pp. 189-216, doi: 10.1023/A:1026564708926.

Muehlberger, G., Seaward, L., Terras, M., Oliveira, S.A., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S. and Gatos, B. (2019), "Transforming scholarship in the archives through handwritten text recognition: transkribus as a case study", *Journal of Documentation*, Vol. 75 No. 5, pp. 954-976.

Neudecker, C. and Antonacopoulos, A. (2016), "Making europe's historical newspapers searchable", *12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Greece, pp. 405-410, 2016, doi: 10.1109/DAS.2016.83.

Nguyen, T., Jatowt, A., Coustaty, M. and Doucet, A. (2021), "Survey of post-OCR processing approaches", *ACM Computing Survey*, Vol. 54 No. 6, Article 124 (July 2021), p. 37, doi: 10.1145/3453476.

Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H. and Tolonen, M. (2021), "Integrated interdisciplinary workflows for research on historical newspapers: perspectives from humanities scholars, computer scientists, and librarians", *Journal of the Association for Information Science and Technology*, Vol. 73 No. 2, pp. 225-239, doi: 10.1002/asi.24565.

Ocal, T., Durgunoglu, A. and Twite, L. (2022), "Reading from screen vs reading from paper: does it really matter?", *Journal of College Reading and Learning*, Vol. 52 No. 2, pp. 130-148, doi: 10.1080/10790195.2022.2028593.

Organisciak, P., Schmidt, B.M. and Downie, J.S. (2021), "Giving shape to large digital libraries through exploratory data analysis", *Journal of the Association for Information Science and Technology*, Vol. 73 No. 2, pp. 317-332, doi: 10.1002/asi.24547.

Pfanzelter, E., Oberbichler, S., Marjanen, J., Langlais, P.-C. and Hechl, S. (2021), "Digital interfaces of historical newspapers: opportunities, restrictions and recommendations", *Journal of Data Mining and Digital Humanities, January*, Vol. 11, p. 2021, HistoInformatics-, doi: 10.46298/jdmdh.6121.

Piotrowski, M. (2012), *Natural Language Processing for Historical Texts*, Morgan & Claypool Publishers, San Rafael, California.

Salmi, H., Paju, P., Rantala, H., Nivala, A., Vesanto, A. and Ginter, F. (2020), "The reuse of texts in Finnish newspapers and journals, 1771-1920: a digital humanities perspective", *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Vol. 54 No. 1, pp. 14-28, doi: 10.1080/01615440.2020.1803166.

Savoy, J. and Naji, N. (2011), "Comparative information retrieval evaluation for scanned documents", *Proceedings of the 15th WSEAS international conference on Computers*, Wisconsin, USA, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, pp. 527-534, doi: 10.5555/2028299.2028394.

Strange, C., McNamara, D., Wodak, J. and Wood, I. (2014), "Mining for the meanings of a murder: the impact of OCR quality on the use of digitized historical newspapers", *Digital Humanitites Quarterly*, Vol. 8 No. 1, available at: http://digitalhumanities.org/dhq/vol/8/1/000168/000168.html

Taghva, K., Borsack, J. and Condit, A. (1996), "Evaluation of model-based retrieval effectiveness with OCR text", *ACM Transactions on Information Systems*, Vol. 14 No. 1, pp. 64-93, doi: 10.1145/214174.214180.

Tanner, S., Munoz, 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*, Vol. 15 Nos 7/8, doi: 10.1045/july2009-munoz.

Torget, A.J. (2022), "Mapping texts: examining the effects of OCR noise on historical newspaper collections", in Bunout, E., Ehrmann, M. and Clavert, F. (Eds), *Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology*, De Gruyter Oldenbourg, Berlin, Boston, pp. 47-66, 2023, doi: 10.1515/9783110729214-003.

Traub, M.C., van Ossenbruggen, J. and Hardman, L. (2015), "Impact analysis of OCR quality on research tasks in digital archives", in Kapidakis, S., Mazurek, C. and Werla, M. (Eds), *Research and Advanced Technology for Digital Libraries. TPDL 2015. Lecture Notes in Computer Science*, Vol. 9316, Springer, Cham, doi: 10.1007/978-3-319-24592-8_19.

Traub, M.C., van Ossenbruggen, J.R., Samar, T. and Hardman, L. (2018), "Impact of crowdsourcing OCR improvements on retrievability bias", *Proceedings of the 18th ACM/IEEE on Joint*

Conference on Digital Libraries (JCDL '18), New York, NY, USA, Association for Computing Machinery, pp. 29-36, doi: 10.1145/3197026.3197046.

van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B. and Colavizza, G. (2020), "Assessing the impact of OCR quality on downstream NLP tasks", *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, Vol. 1, ARTIDIGH, pp. 484-496, ISBN 978-989-758-395-7, doi: 10.5220/0009169004840496.

Vitello, S. (2022), *What Impacts Success in Proofreading? A Literature Review of Proofreading on Screen vs on Paper*, Cambridge University Press & Assessment, Cambridge.

Zetterberg, S. (1988) (Ed.), *Maailmanhistorian pikkujättiläinen*, WSOY, Helsinki.

Zetterberg, S. (1989) (Ed.), *Suomen historian pikkujättiläinen*, WSOY, Helsinki.

Further reading

Kantor, P.B. and Voorhees, E.M. (2000), "The TREC-5 confusion track: comparing retrieval methods for scanned text", *Information Retrieval*, Vol. 2 No. 2, pp. 165-176.

(The Appendix follows overleaf)

Appendix
Pre-formulated queries

Table A1.
List of the pre-formulated queries

ID	Query in Finnish	Rough translation
1	Bobrikoffin murha 1904	Murder of (Nikolai) Bobrikoff in 1904
2	Postimanifesti 1890	Postal manifest in 1890
3	Nuorsuomalaisen puolueen perustaminen vuonna 1894	Founding of the young Finns' party in 1894
4	Helmikuun manifesti 1899	The February manifest in 1899
5	Eduskuntavaalit 1907	Parliamentary elections in 1907
6	Hannes Kolehmainen Tukholman olympialaisissa 1912	Hannes Kolehmainen at the Stockholm Olympics in 1912
7	Maailmansodan rauha 1918	Peace of the First World War in 1918
8	Nansenin matka pohjoisnavalle	Nansen's expedition to the North Pole
9	Lokakuun vallankumous Venäjällä 1917	October revolution in Russia year 1917
10	Saksan keisarikunta 1871	The German Empire 1871
11	Norjan itsenäisyys 1905	Independence of Norway in 1905
12	Tampereen valloitus 1918	Conquest of Tampere in 1918
13	Suomen kuningas Friedrich Karl	Karl Friedrich, the King of Finland
14	Tokoin senaatti 1917	The senate of (Oskari) Tokoi
15	Tukholman olympialaiset 1912	The Olympic games of Stockholm in 1912
16	Maamieskoulu	Agricultural school
17	Laukon torpparilakko	Sharecroppers' strike in Laukko
18	Helsingin valtaus 1918	Occupation of Helsinki in 1918
19	Suomen itsenäisyys 1917	Independence of Finland in 1917
20	Espanjantauti	The Spanish flu
21	Viaporin kapina 1906	Rebellion in Viapori in 1906
22	Laulaja Aino Ackte	Singer Aino Ackte
23	Suomen laulu kuoro	The choir of Finnish song
24	Suomen Naisyhdistys	The Finnish Womens' association
25	Lontoon olympialaiset 1908	London Olympics 1908
26	Raitiotie Helsingissä	Tramway in Helsinki
27	J. L. Runebergin kuolema 1877	Death of J.L. Runeberg in 1877
28	Mannerheim valtionhoitajana 1918	(General) Mannerheim as a regent in 1918
29	Torpparilaki	The sharecropper law
30	Elinkeinovapaus 1879	Freedom of livelihood in 1879

Corresponding author
Kimmo Kettunen can be contacted at: kimmo.kettunen@uef.fi