Routledge
Taylor & Francis Group

Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/vhim20

The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective

Hannu Salmi , Petri Paju , Heli Rantala , Asko Nivala , Aleksi Vesanto & Filip Ginter

To cite this article: Hannu Salmi , Petri Paju , Heli Rantala , Asko Nivala , Aleksi Vesanto & Filip Ginter (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, DOI: 10.1080/01615440.2020.1803166

To link to this article: https://doi.org/10.1080/01615440.2020.1803166

Published online: 15 Sep 2020.

HISTORICAL METHODS: A JOURNAL OF QUANTITATIVE AND INTERDISCIPLINARY HISTORY
https://doi.org/10.1080/01615440.2020.1803166

The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective

Hannu Salmi[a], Petri Paju[a], Heli Rantala[a], Asko Nivala[a], Aleksi Vesanto[b], and Filip Ginter[b]

[a]Department of Cultural History, Faculty of Humanities, University of Turku, Finland; [b]Department of Future Technologies, Faculty of Science and Engineering, University of Turku, Finland

ABSTRACT

The digital collections of newspapers have given rise to a growing interest in studying them with computational methods. This article contributes to this discussion by presenting a method for detecting text reuse in a large corpus of digitized texts. Empirically, the article is based on the corpus of newspapers and journals from the collection of the National Library of Finland. Often, digitized repositories offer only partial views of what actually was published in printed form. The Finnish collection is unique, however, since it covers all published issues up to the year 1920. This article has a two-fold objective: methodologically, it explores how computational methods can be developed so that text reuse can be effectively identified; empirically, the article concentrates on how the circulation of texts developed in Finland from the late eighteenth century to the early twentieth century and what this reveals about the transformation of public discourse in Finland. According to our results, the reuse of texts was an integral part of the press throughout the studied period, which, on the other hand, was part of a wider transnational practice.

KEYWORDS

Digital history; digital humanities; newspaper history; computational methods

In the nineteenth century, newspapers were highlighted as "the chronicle of civilisation, the common reservoir into which every stream pours its living waters, and at which everyman may come and drink" (*The Bradford Observer* May 31, 1838). These words by Sir Edward Bulwer Lytton were originally presented in the opening address of the Lincoln Tradesmen's Newsroom in the 1830s, but they became an epitome of, and a catchphrase for, the nineteenth-century press in general. Newspapers were the big data of the century, a "vast accumulation of facts," as Bulwer Lytton expressed it, and an expanding industry of knowledge production. Newspapers were also a network, the participants of which were keen to share one another's content. In fact, Bulwer-Lytton's words became a viral phenomenon themselves and throughout the century were copied from paper to paper in the English-language press, not only in Britain but also in the United States and Australia.

The contemporaries characterized the habit of borrowing textual contents from various sources as "labour with scissors and paste" (*The Westmorland Gazette* March 29, 1823). This was also the case in the newspaper business, where scissors-and-paste journalism became a dominant form of reusing and recycling texts (Beals 2017; see also Gruber Garvey 2013). In essence, the practice continues today, faster and easier than ever before, on the internet. Today, there are excellent possibilities of exploring text reuse, since a tremendous number of historical newspapers has been digitized worldwide, starting in the late 1990s, and more are constantly being transformed into digital formats.

These digital collections have given rise to a growing interest in studying them with computational methods. This article contributes to this discussion by presenting a new method for detecting text reuse in a large corpus of digitized texts. Empirically, the article is based on the digitized corpus of newspapers and journals from the collection of the National Library of Finland. Often, digitized repositories offer only partial views of what actually was published in printed form. The Finnish collection is unique, however, since it covers all published issues up to 1920.

This article has a two-fold objective: methodologically, it explores how computational methods can be developed so that text reuse can be effectively

CONTACT Hannu Salmi hansalmi@utu.fi Department of Cultural History, Faculty of Humanities, University of Turku, Turku FI-20014, Finland.

identified, for the benefit of historical research on large digital corpora; empirically, the article concentrates on how the circulation of texts developed in Finland from the late eighteenth century to the early twentieth century and what this reveals about the transformation of public discourse in Finland. The empirical dimension aims at showcasing the potential that exists in the development of computational methods for historical research.

The study of text reuse is not a new invention as such. Several projects have investigated this phenomenon, not only in newspapers but also across different literary genres. As an idea, text reuse refers to repetition of information, including exact quotations but also intentional or unintentional borrowing and paraphrasing. The study of text reuse has paid attention to morphological, linguistic, syntactic and semantic similarities as well as variations in the chain of repetition (Gaizauskas et al. 2001; Clough et al. 2002; Lee 2007; Büchler et al. 2014; Franzini, Franzini, and Büchler 2016). For example, Lincoln Mullen (2016) analyzed how American newspapers in the nineteenth century quoted the Bible, and Büchler et al. (2012) explored how Homeric quotations appeared in other ancient texts. Melodee Beals (2017, 2018) experimented on how to identify the reuse of text in the British press. Text reuse has also been studied in different language contexts. For example, the project *Knowledge, Information Technology and the Arabic Book* analyzed text reuse in the corpus of Arabic texts from the period of 700–1500. Our work on newspapers has been inspired by the *Viral Texts* project, led by Ryan Cordell and David A. Smith, analyzing text reuse in the nineteenth-century American press (Smith, Cordell, and Maddock Dillon 2013; Cordell 2015). In the *Oceanic Exchanges* project, the study of text reuse enlarged to a global scale and brought together researchers from Finland, Germany, Mexico, the Netherlands, the United Kingdom, and the United States to explore transnational and transcontinental information flows (Oceanic Exchanges Project Team 2017).

As a background for our case, it is important to note that major changes occurred in the Finnish press and society during the study period. Finland had belonged to the Swedish kingdom for centuries, but in 1809, Sweden had to cede the Finnish regions to Russia. Within the Russian Empire, Finland was granted the autonomous status of a grand duchy, directly overseen by the tsar. The publication of newspapers started in the 1770s in the city of Turku (Åbo in Swedish), from where it first expanded to Helsinki, Vyborg, and Oulu in the 1820s, and gradually into the inland towns during the second half of the nineteenth century. From 1829 onwards, Finland had its own degree of censorship, supervised for the most part by the Finnish officials. Throughout the nineteenth century, there were different adjustments made to the degree. (Tommila 1988, 175–176, 178; Landgren 1988, 277–279; Leino-Kaukiainen 1988, 440–442.) In the midst of political tensions, Finns were, especially from the 1850s onwards, allowed and able to develop new national and state institutions for their country, and the evolving press played an important role in this development. By the beginning of the twentieth century, newspapers were published in nearly 40 locations. In 1917, toward the end of our study period, Finland gained independence from Russia and waged a short but brutal civil war in 1918. In 1920, 136 different newspaper titles were published in the country. A huge expansion of the press took place at the turn of the century.

Our article is based on the research project *Computational History and the Transformation of Public Discourse in Finland* (COMHIS), wherein we have studied text reuse in the Finnish newspapers and journals from the first published paper in Finland, *Tidningar Utgifne af et Sällskap i Åbo*, translated as "News Published by a Society in Turku," in 1771 up to the year 1920, the last year released for data mining. In the project, we were interested in verbatim text reuse, exact passages of text that had been reprinted within the press (an earlier version of this article and the main results of the project in Finnish, see Rantala et al. 2019, 53–67). The corpus included practically all the published issues from that timeframe in Finland, in sum, 5.1 million pages. In addition to newspapers, this corpus includes journals and magazines that were published, for example, on a weekly or monthly basis. In this article, we use the words "journal" and "magazine" interchangeably (Prior-Miller 2015, 22–50).

Since a substantial amount of Finnish material was originally printed in Gothic typeface, difficult for the optical character recognition (OCR) software to interpret, we had to develop a new method for detecting text reuse. Therefore, this article includes both a discussion on the method that was constructed and also an experiment on how it could be employed in the study of early newspaper history, in our case, the history of the Finnish press, which has had its distinctive pathway. The Finnish press has been published mainly in two languages, Finnish and Swedish, and the latter dominated until the 1880s. During the project's whole timespan, 62 percent of the material was in Finnish, and 37 percent was in Swedish. These figures show

how rapidly the proportion of Finnish language rose at the end of the nineteenth century and in the early twentieth century. In general, the volume of the press was rather small until the 1860s when the real growth started.

In Finland, there is a strong tradition of press historical research. Works published in this field include *Opinionens tryck* (Weight of Opinion, 1985, in Swedish) by Clas Zilliacus and Henrik Knif, the multivolume *Suomen lehdistön historia* (History of the Press in Finland, 1985–1992, in Finnish) edited by Päiviö Tommila, and later, *Sanomia kaikille* (News for All, 1998, in Finnish) by Tommila and Raimo Salokangas. However, these studies were published before the digitization of Finnish newspapers and other periodical publications. In particular, Tommila's project includes a detailed investigation of publication practices and the varying contents of the press. The project also implemented a focused case study of news circulating from one paper to another. Due to the great amount of manual work involved in this kind of study, the period of examination was limited to two months of newspaper publishing in 1848. It is thus understandable that the reuse of texts could not have been systematically scrutinized before the digital age.

The quantitative approach as such is no news: it has a long tradition in the examination of newspapers. Following similar research agendas and projects abroad, the researchers in Finland discussed the most suitable quantitative methods for press history in the 1970s. In particular, Viljo Rasila presented various possibilities for using mathematical analysis models in press research (Rasila 1973; see Paju 2019). Still, for a long time, press material had to be studied on paper or read from a microfilm. This changed in 2001, when the National Library of Finland opened its collection of digitized newspapers, at first consisting of 36,000 pages. In 2020, the National Library's digital corpus covers more than 13 million pages openly available online. This corpus provides excellent opportunities for researchers interested in the contents of historical newspapers. Furthermore, it enables a bird's-eye view of the Finnish publishing trends, such as the thorough analysis of the metadata records of newspapers published in Finland (Marjanen et al. 2019).

In this article, we seek to answer the following questions: to what extent did the press in Finland copy and circulate texts in the years between 1771 and 1920, which kinds of texts were reprinted, how did these publication processes change, and what variations took place in the rhythms of repetition, such as in the virality of texts? Further, we situate these developments with the concurrent transformations such as advancing transport and communication technologies. While answering these questions we also demonstrate the ways in which the material of text reuse can be employed for detailed historical research, combining distant and close reading. These examples also serve to confirm the reliability of the results from our computational processing.

According to our results, the reuse of texts was an integral part of the press, and its major practice, throughout the studied period. Paying attention to the copying and borrowing of texts opens up a new perspective on the central form of Finnish publishing culture, newspapers and journals, which, on the other hand, were part of a wider transnational practice. Overall, we have found that reuse was very diverse. It consisted of the repetition of various permanent materials such as announcements and advertisements, as well as the borrowing of news, stories, and anecdotes from one publication to another, either in the short or very long term. Significantly, the new computational method introduced here, titled *text-reuse-BLAST*, the generated database, and its online interface enable fresh insights both into the press history and, when using newspapers as sources, into the study of history at large.

In the following chapters, we first describe and discuss in more detail the method by which the five million pages available were processed and analyzed. We then look more closely at reuse thematically, first through information movements as a whole and then by focusing on the virality and long-term repetition of texts. In each section, we examine, test, and complement our analysis through close reading and by qualitatively interpreting case studies based on original digitized newspapers as well as other sources.

Text-reuse-BLAST for detecting textual overlaps

The digitized newspapers and journals of the National Library of Finland have served as source material for the project. The project draws on the OCR corpus, originally created to enable full-text searches from a digital newspaper archive (Pääkkönen et al. 2016). The pages of the scanned newspapers and magazines have been converted to text with OCR. There are still flaws in OCR technology and, as a result, some of the characters have been misidentified, causing a considerable amount of noise in the material (Koistinen, Kettunen, and Pääkkönen 2017).

Multa t\ä@tä fyNlkÄsiii kchtalostu ,ct , Abouil Asi,3 wic!lä ticiun't>t ,mitää>«, »vaalii luiftti iloista M,m<iä Tshiragauissa, ©elä fi:föf3>i'öi että uiUfatfpäim -uhkaisiloui i Hviarat, miinto fu^tiaani 'fatifefi- fuffotai» lÄuja THi roinin, puutarhassa ja, ipici'ilitsi hwi'tt<iiöii fmmiamcrk^iUi ja anoo» »imilyMla,

Mutta tästä synkästä kohtalosta ei Abbul Asib »ielä tiennyt mitään, vaan »ietti iloista elämää TshiraganiSsa. Sekä sis»Stä «ttä ulkoapäin uhkasivat «aarat. mutta sulttaani katseli lukkotaisteluja Tfhiiaaanin puutarhassa ja palkitsi voittajan lunnicnnerleillä ja ar° vonimityksillä.

Figure 1. A real example of two equivalent passages in the data, with severe OCR noise.

There have been numerous prior computational approaches to text reuse detection. These have drawn on a range of techniques: text fingerprinting with word n-grams (Citron and Ginsparg 2015), the use of substring matching algorithms (Clough et al. 2002), or a full string alignment as in the well-known *Passim* software developed by David A. Smith in the *Viral Texts* project. The Passim tool has been successfully used e.g., in a study of American newspapers (Smith, Cordell, and Maddock Dillon 2013; Cordell 2015). Passim applies word n-grams to recognize duplicate texts within a certain threshold, even if the texts differ due to editorial work or OCR issues (Smith, Cordell, and Mullen 2015). A common feature of the prior approaches is their reliance on word n-grams as sufficiently unique anchors in text. We tested Passim on the Finnish material, but the Finnish corpus' Gothic typeset, along with the problems in scanning and the peculiarities of the Finnish language, sometimes made the OCR accuracy very low. Passim is based on the recognition of so-called seed overlaps, which must consist of a sequence of more than one whole word, a sequence of five words by default. These kinds of overlaps are, unfortunately, quite rare in our material, in which, on average, 23–30 percent of the words can be misidentified, depending on the time of publication (Kettunen, Pääkkönen, and Koistinen 2016). An admittedly somewhat extreme, but real case from the data is illustrated in Figure 1. Here we see two texts which are in fact the same text (as established from the actual scanned newspaper pages), yet their correspondence is far from obvious among the OCR noise.

With such a low word recognition rate, it became quickly obvious that a character-level solution is more prudent, despite the necessarily higher computational cost. Therefore, we decided to base our own solution for detecting text reuse on the National Center for Biotechnology Information Basic Local Alignment Search Tool (NCBI BLAST). BLAST is a software originally created for comparing and aligning gene and protein sequences and is therefore well applicable to material that includes a substantial amount of noise. The BLAST algorithm is rather similar to Passim, striving to efficiently arrive at the alignment of two strings, and even uses many similar ideas such as seed overlaps. The most important difference is that BLAST is purely character-based and not word-based, comparing texts as sequences of characters. This substantially improves the chance of finding an alignment in extremely noisy text, since even parts of words can contribute to it. The main disadvantage, on the other hand, is that, naturally, the same text is much longer in terms of characters than it is in terms of words. This causes a computational penalty, as the complexity of string alignment increases quadratically in terms of the sequence length. BLAST, however, is developed to align sequences against vast databases and builds on highly optimized algorithms and their implementations.

As a by-product of these optimizations, BLAST is hard-wired to the small alphabet of biological sequences, which means that its application requires preprocessing the material so that the letters are encoded into the alphabet of 23 amino acids. These are not enough to convert all the letters of the Finnish alphabet, and therefore only the 23 most used letters were used. The missing letters are used so rarely compared to others that they did not significantly affect the BLAST results. Also, numbers and special characters were removed for the analysis (Vesanto, Nivala, Salakoski, et al. 2017; for source code, see https://github.com/avjves/textreuse-blast). The process has been described in Figure 2.

As a result of the BLAST run, we got a database where the hits, or the strings of text that sufficiently resemble one another, were grouped in the same cluster. After analyzing five million pages of newspapers and magazines from the years 1771–1920, we got almost 61 million hits, which were divided into 13.8 million clusters. Our results do not directly refer to the number of texts shared however. We did not study BLAST's recall, that is, how many reused texts BLAST can find, as this would have required the manual assembly of a reference corpus, which would have been very laborious. One further study would be to compare the BLAST results, for example, to those of previous studies, where the reprinted news were

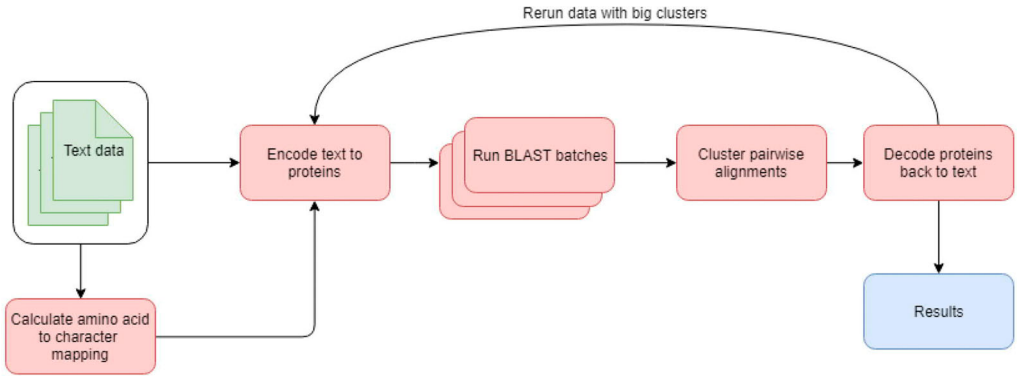


Figure 2. The process of text reuse detection (Vesanto 2019).

collected by human readers. One possible case would be the study conducted in the newspaper project by Päiviö Tommila (1988, 205): they carefully explored how newspapers copied one another's news in January and February 1848. However, the OCR quality varies so much decade by decade that a single case can only give indicative results. Further, the National Library's OCR Corpus is segmented into pages, not into articles. As a result, it is possible that one article may have been split into several clusters due to page or column borders. Furthermore, all found cases of similarity are not necessarily copies from other papers since overlaps in the corpus might also have an external source and be published independently. Still, these overlaps tell on the role of textual similarities in the corpus. BLAST may also have bundled together train timetables, market fare price lists, and other commercials and notifications if the number of characters repeated in them exceeded 300. These problems cannot be avoided, however, if we want text-reuse-BLAST to recognize as many copied texts as possible. From the outset, we did not take this as a problem, since we were interested in all textual contents that had traveled within the press.

It is clear that browsing this kind of a database requires fairly advanced search and filtering functions. Our search portal (Figure 3), which is open to all users at http://comhis.fi, is implemented by Apache Solr indexing. The search functions allow the user to search for both single hits and multiple text clusters. Each cluster is numbered with a unique identifier. Solr filters allow the user to simultaneously target searches to a specific individual newspaper or journal, periodicals of a specific language, either to newspapers or journals only, to a specific time frame, or geographic location. Each hit also includes a link to that page on the National Library's portal of digital newspapers and journals, allowing the user to access and read the same text as an original scan if they want. Solr is also capable of fuzzy search, which searches for similar terms in addition to the search term. This is a useful feature because of OCR errors. In addition, it is possible to save search results in tab-separated values (TSV) format, enabling the processing of those results by other digital humanities methods, such as named-entity recognition (NER), geospatial analysis, topic modeling, or network analysis. Moreover, our search portal contains some analytical tools that allow the user to see the temporal as well as the geographic dimensions of Solr search results (for further details on the construction of the database, see Vesanto, Nivala, Rantala et al. 2017).

For each cluster, we calculated a value that describes its viral character and called this the *virality score*. We will use this analysis tool later in the "Tracking Down Virality" section. For this value, we counted the number of unique newspapers/journals, the locations where the news were printed, and how many days it took. The value is obtained by multiplying the number of newspapers/journals and locations by the inverse of the elapsed time. This penalizes the value if the news has not spread geographically wide, or if the spreading took time (on the code, see https://github.com/avjves/cluster-viral-score).

Also, we left out those hits that clearly differ from the dates of other hits so that the calculated value is not distorted by, for example, single outliers. This can happen, for example, in a situation where the news has spread very quickly during a short timeframe, but a single text was published long after. In this case, the cluster span is long but, in fact, the cluster spread has happened quickly. Finally, the values of all the clusters were normalized between 0–100 for clarity. In the interpretation of virality scores, it is important to consider the difference between newspapers and magazines. The material's metadata derives from the National Library of Finland, and its quality is, in general, very good. Errors may occur, however. The magazines' metadata does not indicate the publication date as accurately as in the case of newspapers, which

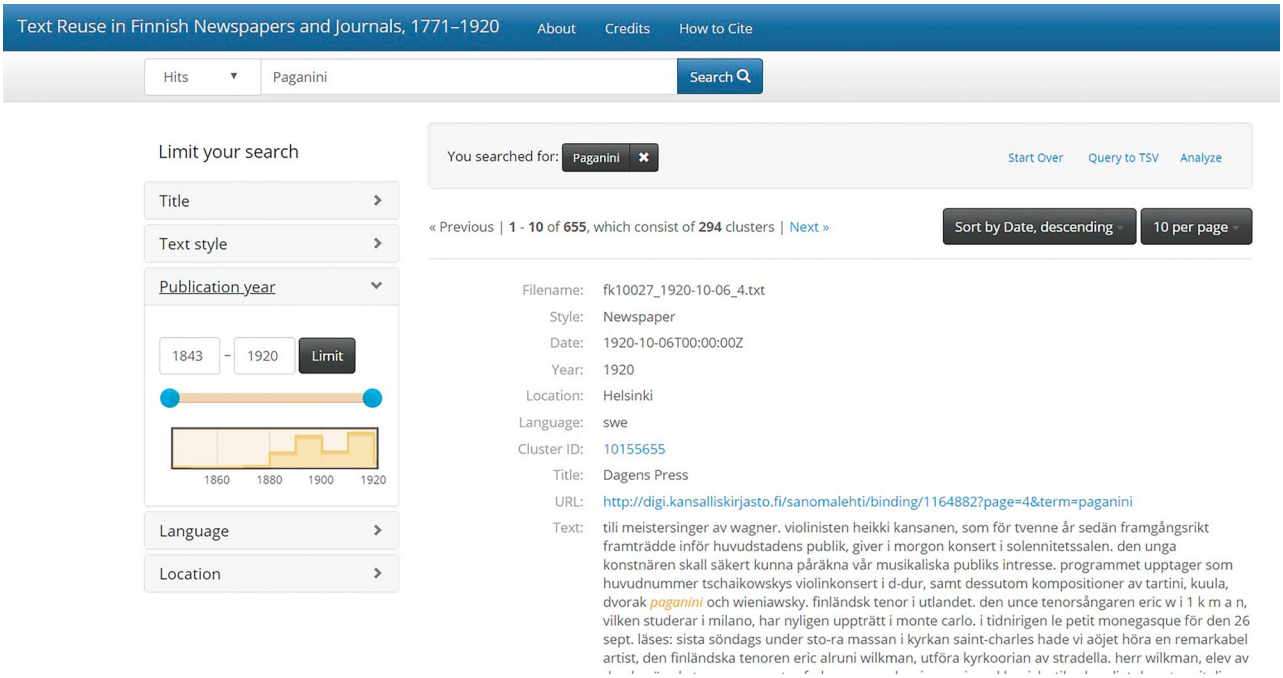


Figure 3. The user interface of the text reuse database.

affects the results. The overall picture is still reliable since the most viral texts can be found from the newspapers, the publication rate of which was much more intensive. In the case of a single cluster, however, this must be taken into account. In the future, it might be illuminating to develop more refined strategies of measuring virality. It is possible to count virality also in relation to the prevailing capacity of the press, i.e., how many newspapers and journals were published during the time span of a repetition chain.

Information movements based on the reuse of texts

In total, the research project found around 13.8 million clusters of repetitive texts or text snippets in the corpus of newspapers and journals using the text-reuse-BLAST software. Figure 4 shows the evolution of the number of clusters by describing how many clusters were launched annually between 1771 and 1920. In Figure 5, the number of clusters is proportional to the number of characters published in newspapers and magazines, thus offering us an idea of how general a feature text reuse was in the press. In the National Library's digitization project, they stored the material one page at a time, but during the study period, the page sizes grew, and fonts changed dramatically (see Marjanen et al. 2019, 63, 67.) If one looks at the publications in terms of the amount of information, it is clear that their "volume" changed so much over time that proportioning to the number of pages does not make sense. Figure 5 shows that while the number of repeated texts increased almost exponentially toward the end of the nineteenth century and the beginning of the twentieth century (cf. Figure 4), in proportion to the number of characters, the reuse of text increased more evenly. Significantly, this indicates that the reuse of texts was an integral part of the press throughout the period. Because text-reuse-BLAST recognizes similarities despite fluctuations in the accuracy of the OCR result level, and because the page breaks result in text clusters breaking into multiple clusters throughout the period, we consider the image in Figure 5 sufficiently reliable. Clearly, there are more textual repetitions that have been grouped into several text clusters toward the end of the period, when the volume of the press increased significantly.

The cluster data we created contains repetition of all kinds, including similarities in repeated texts such as in various bulletins, advertisements, and price lists. Overall, the amount of repetition in the press is enormous, and discovering it and making it visible adds a new dimension to press history and the study of Finnish publishing culture. This new dimension offers us multiple implications. Through the recycling of texts, it is possible to examine, for example, how information has moved from one place to another, how newspapers in different locations have followed one another's news, and what kind of communication network the press has formed in Finland. Our database includes a mapping tool that enables one to visualize the geographic extent and movement of text

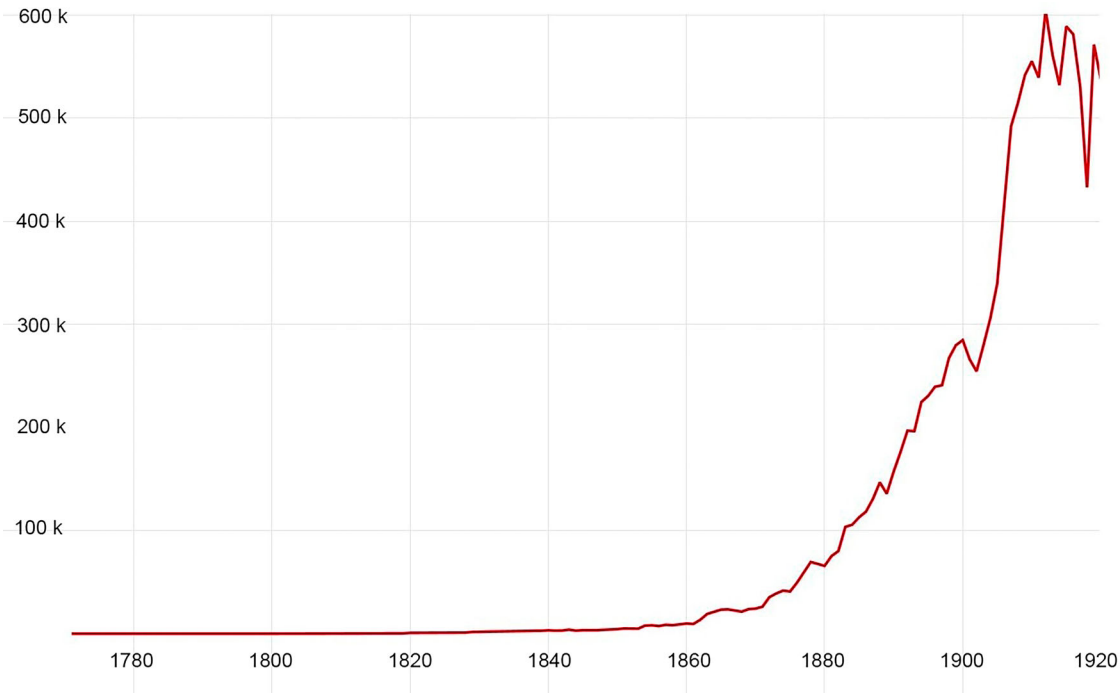


Figure 4. The total amount of text reuse clusters per year, 1771–1920.

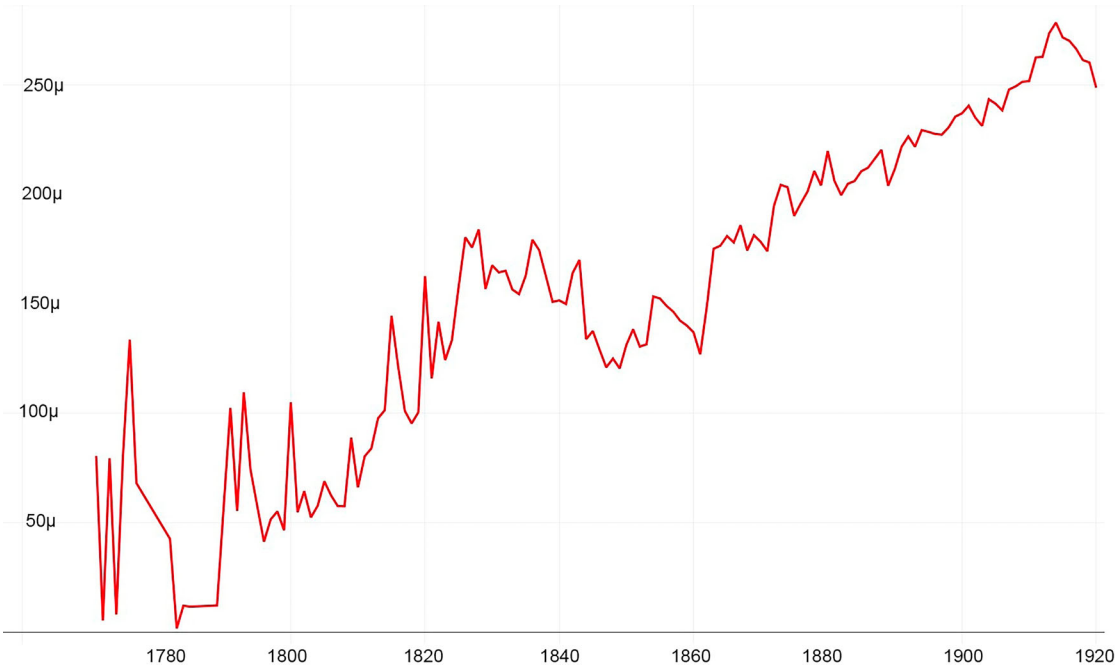


Figure 5. Text reuse clusters per year in relation to the volume of the press (the amount of published characters).

clusters. The user interface allows a researcher to see how newspapers and journals shared content, and s/he can view the flow of information on an annual or longer-term basis. At the same time, one can also narrow the results by language (Swedish or Finnish).

The database shows, for example, that in 1877, the newspapers in Finland launched 59,581 clusters of Finnish and Swedish texts. If we look at the first place of the clusters' publication, the three strongest centers of recycled information in the year 1877 were Helsinki (69 percent), Turku (14 percent), and Vyborg (6 percent). Still, newspapers in smaller towns also copied content from one another, thus bypassing the major centers and suggesting a topic of interest that someone could further investigate in the future. Correspondingly, during the whole studied period, Helsinki was by far the largest cluster originator: approximately 5.9 million clusters began from

Helsinki-based newspapers. The next largest locations starting clusters were Turku (about 1.8 million clusters) and Vyborg (about 1.2 million clusters). For example, Tampere's clearly lower presence (about 760,000 clusters) is understandable since the first newspaper in the area, *Tampereen Sanomat*, began to appear in 1866. Importantly, these results correspond well to the earlier examination by the Finnish Press History project concerning the Finnish publishing hubs, thus confirming the overall results of our computational analysis (Landgren 1988, 284–285, 389–390. See also Marjanen et al. 2019).

Previous research has focused on the Finnish newspapers' usage of international news material and various news channels. The connections in our database appear to be "internal" in Finland because text-reuse-BLAST recognizes reuse within the processed corpus. In fact, the editors scissored foreign news from newspapers published abroad, and foreign publications remained an important source of news even in the latter half of the century, despite technological advances (Apunen 1970, 30; Tommila 1988, 110–111, 389). Although the database only contains the repetition occurring inside the material from the National Library of Finland, we have examined selected reprinting cases to confirm that some of the copied news was part of a wider international chain of republication: information traveled quickly from Central Europe to Finland, for example, on health-threatening topics such as cholera or news related to warfaring (Salmi et al. 2018, 71–73). Texts traveling far beyond European countries have recently been explored in a large-scale international project *Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914*, led by Ryan Cordell, which searches for the reprinting of global news streams by cross-mining newspaper databases in different countries. The project has, for example, studied how the news on the assassination of Nikolay Bobrikov, the Governor-General of Finland, on June 16, 1904, spread on a global level, drawing on newspaper repositories in Australia, Austria, Denmark, Finland, France, Germany, Mexico, Sweden, the Netherlands, and the United States (see Oiva et al. 2019).

Reprinting information and news has been a very strong practice internationally and is a widely accepted and customary way of compiling news releases and regular news publications (Silberstein-Loeb 2012, 467–468). Copying texts gave rise to reprint chains that could also branch and transform, as when editors commented or shortened texts. Repetition chains continued across language boundaries. We have noticed that it was rather typical for a Swedish-language news story to be published in Finnish in an abridged form and sometimes in a very concise length. In the case of foreign news, this is easy to understand: Swedish-language newspapers in Finland could speed up the reporting of overseas events by copying a large, descriptive text from publications printed in neighboring Sweden, but reprinting any foreign news into a Finnish-language newspaper required translation. In such a case, the appropriate use of time often shortened news reporting in Finnish (see and cf. Tommila 1988, 205–207).

In the last decades of the nineteenth century, many technological changes accelerated the movement of information. In 1870, the railway line from the Russian Empire's capital, St. Petersburg, to nearby Helsinki, Finland, was opened. The railway connection enhanced the transport of mail, such as newspapers, and thus simultaneously boosted the spreading of news. According to earlier research, the most important news channel for editors in Finland in the 1860s and 1870s was still other newspapers. Building railways continued in Finland for decades to come, all the while improving communications both within and beyond the national borders. Simultaneously, printing technology progressed, and the number of newspapers began to grow rapidly. Concurrent societal changes, particularly favorable economic development, which increased the number of advertisements and classified ads, supported and amplified the movement of texts (Suomalainen 1979, 103–140; Landgren 1988, 267–420, 280–283, and Passim.) In fact, our database provides an unforeseen opportunity to explore the volume, distribution, and recycling of advertisements in Finland. For example, there are 99 clusters in the database for the word "sale" ("alennusmyynti" in Finnish). The lengthiest of these clusters was a book sale ad by Arvi A. Karisto's publishing house. In August 1917, they had their book advertisement reproduced altogether 27 times in 16 localities (cluster no. 15082100). A very different example was when the Finnish Cooperative Society (SOK) announced in November 1917 that they wanted to buy reindeer lichen to process it into animal fodder, thus saving the usual fodder for human consumption. Telling of severe food supply problems during times of upheaval in Russia, the lichen announcement was printed 125 times (cluster no. 11925897) in 70 Finnish newspapers and journals in 32 locations during one month.

These simultaneous, parallel changes also included the electrical telegraph or, more specifically, its spread into ordinary use in the news business. The eighteenth-

century innovation of the optical telegraph did little for recycling texts. The administration first built the electrical telegraph line to Finland from the direction of St. Petersburg in 1855. Celebrated as the "electric wire," it attracted attention as a symbol of a new era and connections but, in practice, it was primarily used for military purposes. The electrical telegraph had an indirect impact on the press in Finland, as the telegraph network abroad accelerated the flow of information from elsewhere in the world to St. Petersburg and Sweden, and at the same time, from their newspapers to Finland. The electrical telegraph's influence on news coverage became more visible in the newspapers in the late 1870s. They referred to electrical messages when, for example, the news of the death of Johan Ludvig Runeberg, regarded as the national poet of Finland, reached the newspapers in 1877: "This message, which has been flying around Finland through the power of electricity, is a source of sadness and longing everywhere" (*Länsi-Suomi* May 12, 1877, 1; Suomalainen 1979, 105.) Nevertheless, telegrams remained expensive in Finland, and their use was restricted.

The first Finnish news agency, Suomen Sähkösanomatoimisto (the Telegram Agency of Finland), the predecessor of the Finnish News Agency (STT), started in 1887 with the aim of providing news especially for the countryside, where many local newspapers had been recently established. Thus, the use of electrical telegraphy began to grow more widespread at the same time as the use of the telephone, which advanced editing offices had introduced as early as 1882. (Kaskinen 1978, 76–93; Rantanen 1987, 20–23; Immonen 2002, esp. 64–66.) In our database, one can see these aspects of the increasing speed and decreasing costs of information transmission in the late nineteenth and early twentieth century as continued and significant growth in the number of text reuse clusters.

When taking an example year from the end of the nineteenth century, we see that in 1897, there appeared 240,933 reuse clusters, which is four times the number in 1877. The mapping tool of the database also shows threads that go beyond its frames (see the image in Rantala et al. 2019, 63). This is because the material includes magazines founded and published by Finnish immigrants in North America and, as can be deduced from the cluster data, many news stories from 1890s Finland were reprinted and circulated by the editors of immigrant newspapers across the Atlantic.

In the years to come, the growth of the press continued in terms of both the number of titles and capacity, and the reuse clusters reflect all this. In 1907, the Finnish press started 492,157 clusters and in 1917, 530,843. As stated, these numbers do not directly refer to the number of texts published in the news, as the processing has divided reprinted texts into several separated clusters. The figures are, however, indicative as such and display the geographical movements of the news and the increasing communication frequency.

Tracking down virality

In the media culture of the 2000s, the concept of virality has become common parlance in referring to "an image, video, piece of information, etc. that is circulated rapidly and widely on the Internet" (Oxford Living Dictionaries 2019). Although virality is strongly present in the study of present-day social media, it may be argued that the necessary condition for virality, the rapid spread of information and messages, already happened during the press' expansion in the nineteenth century (Salmi 2018, 71–79). In Finland, however, this kind of viral character of news circulation can hardly be identified before the end of the century when the volume of the press multiplied in an accelerating rhythm. As an example, let us consider the circular launched by the Finnish Literature Society in August 1864 to promote the project of erasing a statue for the famous academic Henrik Gabriel Porthan, hailed by the nationalist movement in Finland. The circular was finally published in 18 newspapers, that is, in all the newspapers published in the country in August 1864. The letter by the Society thus received maximal attention by the press, but this text cannot be regarded as a particularly viral phenomenon if the absolute volume of the spread is regarded as a criterium. In addition to volume, the qualities of virality also include rapidity: the diffusion of this text took 15 days, although most of the publications came out within only five days (Salmi 2018, 75–76).

The volume of the Finnish press grew rapidly from the 1860s onwards. In the end of our timeframe, in 1920, 136 different newspaper titles were published in Finland. According to the National Library of Finland's digital collection, 412 journal or magazine titles were published in 1920. In the database, constructed for our project, it is possible to browse reuse clusters on the basis of both the count and what we have called the *virality score*. The text with the highest virality score is not journalistic content but an advertisement, through which the Finnish tobacco industry urged the audience to favor domestic cigarettes in its fight "against the invasion of an American tobacco trust" (*Kotimainen työ* 3 [1916], 43, see also cluster no. 11592519). This advertisement, or announcement, was

published 75 times in 45 different newspapers or magazines in 26 different locations, and it was a direct continuation of the fight against the products of foreign companies, which had already been discussed in the public *en masse*. In other words, this was not a text that would have originated from within the press as a media ecosystem; it came from an external source. The purpose was to make a matter, regarded as of high importance, to circulate through the press as quickly and as widely as possible. The announcement was first published in the daily newspapers *Keski-Savo* (Savonlinna), *Tampereen Sanomat* (Tampere), and *Suupohjan Kaiku* (Kristiinankaupunki) on March 14, 1916, and thereafter as a fast-paced series until the end of March 1916.

Many texts with high virality scores were advertisements or announcements. There are, in sum, 81 clusters with virality scores higher than 50. Of these, clearly the most widely circulated journalistic text is an essay on Finnish literature by Yrjö Koskelainen (1885–1951), published in *Käkisalmen Suomalainen* and *Kajaanin Lehti* on October 9, 1911, the eve of the author Aleksis Kivi's commemoration day (cluster no. 13611711). Again, the text was deliberately distributed to the press for the widest possible publicity. At the same time, the spread of the text indicates how the press was an active contributor to Aleksis Kivi's reputation and to the construction of Finnish culture. Koskelainen, who belonged to the so-called Young Finns, wrote regularly to the magazines *Valvoja* and *Aika* and became the editor of the Turku-based *Uusi Aura* the following year (Muiluvuori 2000). Koskelainen's writing was published 64 times between October 9 and 14 and additionally once a year later. Because the publication frequency was so fast, it is clear that the spread of the text was not really about copying it from one publication to another; rather, the text was sent to several papers simultaneously. Still, it was a very effective recycling of textual content.

International media events were also covered by the press in a viral manner. The disaster of the passenger liner *Titanic* during the night between April 14 and 15, 1912, was widely discussed in the Finnish press. The destiny of the *Titanic* was an intermedial event in the sense that the news itself was circulated through different media, from wireless telegraph to early cinema. In Finland, the most viral news was not actually the first story of the catastrophe but the news on the gathering of the Finnish Parliament on April 19. The Parliament President, P. E. Svinhufvud, had offered his regrets to the relatives of all the victims around the world, including the families of those many Finns that had been on board (cluster no. 12897295). The opening words of the speech were published by 53 papers in 23 locations between April 20 and 26, the virality score being 15.24. The news was actually simultaneously published by 19 papers on April 20, which shows that they had all gotten the news from the same parliamentary documents. It seems likely, too, that the cluster's last hits were copied from other papers that had already printed the text.

The database shows how the press was like a sounding board, or an echo base, for news, announcements, advertisements, and propaganda. The press offered widespread publicity, and it was possible to reach readers from all over Finland. Of course, there were also texts that spread from one magazine to another with a slower rhythm. Such was the case of, for example, the "Winter Notes for Compatriots" article circulated in 1911, which was copied from a Swedish source and was intended to provide instructions for day-to-day life during wintertime (cluster no. 10634989). The virality score is only 2.07, which is explained by the news' slower procession because it was specifically copied from one publication to another. The text first appeared in Finland in *Suomalainen Kansa* on February 1, 1911, and in *Etelä-Suomi* only three days later. In the end, this Swedish text was released 30 times in 19 locations within 21 days.

The virality of the press can also be exemplified by a piece of educational news from 1905. On the last day of May, the newspaper *Helsingfors-Posten* mentioned Ms. Jenny Markelin, a newly graduated engineer, saying that she was "Finland's first woman engineer." After many female architects, Markelin was the first woman graduate at the Department of Engineering. The next day, the largest Swedish- and Finnish-language papers in the Helsinki area repeated the news on the first woman engineer word by word. Soon, the news was copied for publication in newspapers and journals across the country in places like Sortavala and Kuopio. A magazine published in the latter town, *Pohjois-Savo*, repeated the Finnish version of the news as the last sequel of the repetition chain on June 5, 1905. In general, the latest news coverage was published in Swedish in the newspaper *Kotka Nyheter* on June 10, 1905. However, for instance, the newspaper *Uusi Aura* in Turku reprinted only part of the news on June 1, 1905. The case also highlights some of the limitations of our database, since the reuse of texts under 300 characters could not be recognized. The National Library of Finland's digital newspaper collection contains a number of shorter, partial repetitions that have not ended in our

database. While the clusters of our database include a total of 16 news reprints in two languages on the first woman engineer, the National Library's digitized newspapers also include around ten other cases, shorter than 300 characters, from the same period. All in all, the news on Finland's first female engineer reached most of the Grand Duchy of Finland during the first week of June in 1905. After the newspapers, the news was repeated, with slight editorial changes, by magazines such as *Palvelijatarlehti*, *Suomen Teollisuuslehti*, and *Nutid*, which was the supporter of the Women's Union. This example also reminds us that our database's reuse clusters should be qualitatively analyzed in parallel with the original digitized material (Paju 2018, 5–24).

Toward the end of the nineteenth century, newspapers and magazines circulated texts with increasing speed. The recycling of media content fulfilled the virality characteristics, and news, announcements, and other texts were distributed at a rapid rate over a wide geographical area. Texts were fed and provided to the press for recycling but, at the same time, the newspapers and magazines were copying texts from one another. Our database allows the researcher to analyze how particular historical phenomena, such as the news on changing gender roles, were amplified in late nineteenth-century and early twentieth-century Finland.

Long-term reuse of texts

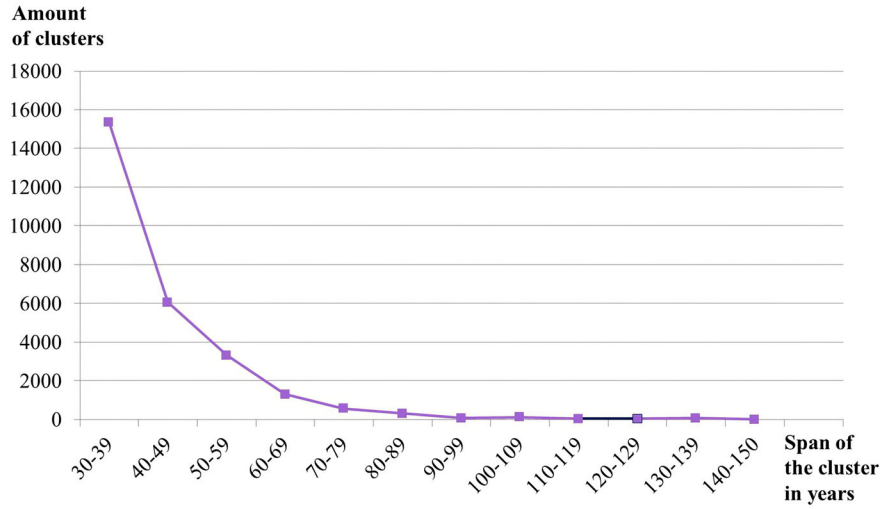
Besides the rapid circulation of texts, we have been able to trace the opposite of virality, the reuse of texts that has taken place over a very long period of time. Actually, the longest chains of repetition cover almost the whole time scale of the project, over 140 years. This means that there were texts being republished over 100 years after their first publication. We call this feature of text recycling *long-term reuse*. This type of reuse was characteristic of the Finnish press, but it is probable that it was also common practice in other countries and regions. In the *Viral Texts* project, it was discovered that some reused texts traveled slowly but, in this case, the time lag was counted in years, not in decades, as in our project. It must be noted that in the *Viral Texts* project, the text corpus covers "only" 30 years (Smith, Cordell, and Maddock Dillon 2013, 93). The discovery of very long chains of reuse highlights the strength of our method, as well as the benefit of being able to operate with a corpus that covers a long time span. It would be very hard, even impossible, to trace this kind of circulation of texts with close reading methods and/or by concentrating on just a few decades of newspaper publishing. In Finland, some press historians have been able to trace some individual cases of this kind of long-term republication (Teperi 1977, 83–4, 154; Pietilä 2008, 330; footnotes 1257–59), but with text-reuse-BLAST, we can discover this phenomenon in its full scale.

As such, the long-term reuse of texts is not in any way a dominant feature in our corpus. According to the database, 85 percent of all the chains of reuse are shorter than one year in their total length. The remaining 15 percent—still over 2 million clusters—includes very different time spans. When examining long-term reuse cases, it is useful to distinguish the difference between the span and gap of the clusters, both offering valuable information. Here, "span" refers to the temporal distance between the first and last hit in a cluster, while "gap," for its part, refers to the break between different hits in a cluster. When a cluster consists of more than just two hits, the span and gap differ from each other. Moreover, one cluster can have several gaps in it.

Figure 6 shows the number of those clusters that remained active for at least 30 years. In the figure, one can clearly see the diminishing trend of clusters: there are 27,299 clusters spanning at least 30 years, but only 5,888 clusters that spanned for 50 years or more. Still, 289 clusters have a span that stretched to 100 years. The longest span and gap values in the database cover over 140 years, meaning that there were texts in the Finnish press that traveled in time from the late eighteenth century to the early twentieth century.

In practice, the cases with the highest span or gap (this value being 100 years or more) are texts taken from the very first papers published in the 1770s and republished in early twentieth-century newspapers or journals. In most cases, the source was the first Finnish-language paper, *Suomenkieliset Tieto-Sanomat*, which was published only for a short period of time in 1776. One sample issue of this newspaper had already come out in 1775, and its content was also reused. (On these first papers' reuse, see Salmi et al. 2019, 538.) As such, it is not surprising that the ever-growing Finnish-language press was interested in republishing extracts from the very first paper published in the Finnish language. By circulating the contents of these early papers, the press maintained awareness of the history of Finnish newspaper publishing practices.

In some cases, the old text was republished only once. For example, in 1910, a journal called *Frisk Bris*, published by a yachtsmen association, republished a text from 1771, which originally appeared in *Tidningar*



Span in years	Number of clusters
30–39	15351
40–49	6060
50–59	3329
60–69	1312
70–79	571
80–89	320
90–99	67
100–109	128
110–119	33
120–129	53
130–139	64
140–150	11
Total	27299

Figure 6. Text reuse clusters that spanned for 30 years or more.

Utgifne af et Sällskap i Åbo and discussed the flooding of the sea in the Finnish archipelago (cluster no. 3699979). The gap between the original text and its reuse is as long as 139 years. Similarly, the musical journal *Finska musikrevy* was interested in the same paper's musical contents. In 1906, it republished extracts from *Tidningar*'s issue from 1773, which dealt with the early concert life in Finland (cluster no. 4932546). In these cases, the republishing becomes motivated through the topic area of the periodical in question. However, there are also cases where the old text has been republished in several cycles by many newspapers. In 1837, the Finnish-language journal *Mehiläinen* published a text dealing with the dangers of spirits and drunkenness, which was considered a common problem among the peasantry. *Mehiläinen* was published by Elias Lönnrot, a Finnish scholar who was trained as a physician but better known as a collector of folk poetry and the author of the *Kalevala*. This text by Lönnrot was recycled by the later Finnish-language press in several cycles: first in 1889 by eight papers, then again in 1891 and 1892 by several papers and, finally, in 1902 by 15 different papers (cluster no. 7928699).

We have discovered that some of the long-term reuse cases circulating in the Finnish press have been anecdotes or news copied and translated from foreign papers (Salmi et al. 2019, 540). Moreover, among other reuse cases with long time lags, there are many anonymous texts—minor news, stories, and anecdotes—that lived in the press through text circulation. In the case of nineteenth-century American press, Ryan Cordell (2015) has argued that many circulated texts were different kinds of anonymous stories, snippets, and notices—texts with minor editorial value and thus also ignored in press history. In the case of circulating anecdotes or stories, the question of authorship is of secondary importance. An important motivation behind the practice of republishing these kinds of snippets seems to be the entertainment value and a need to fill newspaper columns. Without this entertainment value, for example, the following story would not have interested the early twentieth-century press the way it did. In 1851, the Finnish-language newspaper *Suometar* published a story of a tailor from southern Finland who had a special quality of being able to run fast. Sixty-two years later, this story was rediscovered by the newspapers, and in 1913, more specifically during a couple of months, 29 newspapers throughout the country circulated the old story of a light-footed tailor (cluster no. 14169335). A parallel reading of the original digitized material reveals that *Suometar* had actually borrowed and translated the story from the Swedish-language paper *Borgå Tidning* (May 21, 1851). This reuse has remained unnoticed in our database, however, since BLAST is unable to depict the similarity between different languages. In this reprinting case, however, there was a contemporary connection. In Finland, success in long-distance running had become an issue of national pride in the Stockholm Olympic Games of 1912, where the Finns won several gold medals, and the text reuse in 1913 referred to Tatu Kolehmainen ("Tatu from the last century") who was the leading Finnish long-distance runner that summer.

The long-term reuse of texts is a form of repetition that has been discovered in its full scale only by computational methods. Without text-reuse-BLAST, this feature of the Finnish press would have remained hidden. With the help of our database, it is now possible to further examine the volume and contents of long-

term reuse. It is important to add that the database of reuse detection can help the researcher to focus his/her close reading on the topics of his/her own choice, which again contributes to the understanding of the variety and complexity of text reuse.

Conclusion

In this article, we have discussed the volume and nature of text reuse in the Finnish press from 1771 to 1920. Text-reuse-BLAST proved highly effective in aligning similar passages in the OCR corpus. This also succeeded in the case of early Finnish newspapers from the late eighteenth century, where the OCR accuracy is especially poor.

Our study shows that borrowing and recycling have not been connected only to certain time frames in the history of the Finnish press; rather, it characterizes the whole period. As the total volume of the press increased over the course of the nineteenth century, the number of circulated texts also grew respectively. A special feature which, to the best of our knowledge, has not been analyzed earlier, is the long-term repetition: the press was an archive from which it was possible to draw contents for new publications and, at the same time, it formed a channel and a platform for cultural memory. We also found the viral circulation of texts in the press. Virality was manifested when the same text spread rapidly across the network, often within a few days or weeks. Because there were millions of shared texts, we decided to publish the material as a database so that the results of our study can more widely benefit researchers and the general public.

Recycled material led us to think about the importance of agency. In the nineteenth century, the press was a forum not only for many types of texts but also for a variety of interests. The plethora of advertisements and notices emphasizes the material's polyphonic nature and the shared and mixed agency: businesses and communities, parishes and cities, and authorities and citizens spoke through the press. The significance of journalistic writing and the efforts to influence public opinion grew over the period under study, but the nature of the papers remained heterogeneous. Ryan Cordell, a researcher of the nineteenth-century press, used the "network author" concept to describe the communal and shared authorship evident in the printed material. In Finland, too, the press could be interpreted through networking: of course, editorial and other opinion leaders influenced the writing, but so did the structural ways in which texts were created. This shared agency consisted of focal points and centers of text circulation, routes of information flow and information access, and many republication practices.

As a method, the detection of text reuse is, as our results show, a fruitful way of studying the movements and routes of information. The method itself is not language-specific but can be employed in the study of any other digitized newspaper repository. However, the method we have proposed also has limitations, such as the fact that translation, that is, reuse across a language border, could not be automatically detected. This is clearly something that should be investigated in the future: when the possibilities of machine translation improve, could it be possible to recognize information flows across linguistic borders? The same news spread in both Finnish and Swedish, as shown by the example of Finland's first female engineer in 1905. Currently, the poor quality of OCR'd text makes it very difficult to draw on machine translations. The identification of translated text passages would also be important for understanding how the Finnish press, in the end, participated in the international exchange of news and information. Evidently, Finland was not only a receiving area, as the news also traveled from Finland to other areas in Europe and to other continents. If it were possible to combine digitized newspaper corpora more extensively, the Finnish press would appear as part of a wider transnational information network. The methods for text reuse detection can be further developed and exploited on both a national and an international level to shed more light on the history of communication.

Text reuse database

Vesanto, A., F. Ginter, H. Salmi, A. Nivala, R. Sippola, H. Rantala, and P. Paju. 2018. *Text Reuse in Finnish Newspapers and Journals, 1771–1920*, http://comhis.fi/clusters.

Bibliography

Apunen, O. 1970. *Hallituksen sanansaattaja: Virallinen lehti – Officiella tidningen 1819–1969 (The messenger of the government: The official newspaper 1819–1969, in Finnish)* (pp. 1–5). Helsinki, Finland: Valtion painatuskeskus.

Beals, M. H. 2017. Scissors and paste: The georgian reprints, 1800–1837. *Journal of Open Humanities Data* 3. https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.8/. doi: 10.5334/johd.8.

Beals, M. H. 2018. Close readings of big data: Triangulating patterns of textual reappearance and attribution in the

Caledonian Mercury, 1820–40. *Victorian Periodicals Review* 51 (4):616–39. doi: 10.1353/vpr.2018.0046.

Büchler, M., G. Crane, M. Moritz, and A. Babeu. 2012. Increasing recall for text re-use in historical documents to support research in the humanities. Proceedings Second International Conference on Theory and Practice of Digital Libraries, 7489: 95–100. doi: 10.1007/978-3-642-33290-6_11.

Büchler, M., P. R. Burns, M. Müller, E. Franzini, and G. Franzini. 2014. Towards a historical text re-use detection. In *Text mining. Theory and applications of natural language processing*, ed. C. Biemann and A. Mehler. Cham: Springer. doi: 10.1007/978-3-319-12655-5_11.

Citron, D. T., and P. Ginsparg. 2015. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences of the United States of America* 112 (1):25–30. doi: 10.1073/pnas.1415135111.

Clough, P. D., R. Gaizauskas, S. L. Piao, and Y. Wilks. 2002. Measuring text reuse. Proceedings of Association for Computational Linguistics (ACL2002), Philadelphia, PA, 152–159.

Cordell, R. 2015. Reprinting, circulation, and the network author in antebellum newspapers. *American Literary History* 27 (3):417–45. doi: 10.1093/alh/ajv028.

Franzini, G., E. Franzini, and M. Büchler. 2016. Historical Text Reuse: What Is It? http://www.etrap.eu/historical-text-re-use/

Gaizauskas, R., J. Foster, Y. Wilks, J. Arundel, P. Clough, and S. L. Piao. 2001. The METER corpus: A corpus for analysing journalistic text reuse. Proceedings of Corpus Linguistics 2001, Lancaster, UK, 214–223.

Gruber Garvey, E. 2013. *Writing with Scissors. American Scrapbooks from the Civil War to the Harlem Renaissance*. New York: Oxford University Press.

Immonen, K. 2002. *Sillat sielujen ja ihmismietteen: Suomalaisen puhelimen kulttuurihistoriaa keskusneideistä tekstiviesteihin (The bridges of souls and thoughts: The cultural history of the telephone in Finland, in Finnish*). Helsinki, Finland: Edita.

Kaskinen, T. 1978. Lennätin ja radio (The telegraph and the radio, in Finnish). *Tietoliikenne Suomessa 1860–1939*, 76–93. Helsinki, Finland: Suomen sanomalehdistön historia -projektin julkaisuja.

Kettunen, K.,. T. Pääkkönen, and M. Koistinen. 2016. Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910. Sanatason laatu, kokoelmien käyttö ja laadun parantaminen (The digitized historical newspapers of the National Library of Finland, 1771–1910: The word-level quality, the use of the collection and the improvement of quality, in Finnish). *Informaatiotutkimus* 35 (3):3–14.

Koistinen, M., K. Kettunen, and T. Pääkkönen. 2017. Improving optical character recognition of finnish historical newspapers with a combination of Fraktur & Antiqua models and image preprocessing. In *Nordic conference on computational linguistics, NoDaLiDa 2017 Gothenburg, Sweden. Linköping electronic conference proceedings* 2017, 277–283. https://ep.liu.se/ecp/131/038/ecp17131038.pdf

Landgren, L. 1988. Kieli ja aate – politisoituva sanomalehdistö 1860–1889 (Language and thought: The politicization of the press in 1860–1889, in Finnish). In *Suomen lehdistön historia 1: Sanomalehdistön vaiheet vuoteen 1905*, 269–420. Kuopio, Finland: Kustannuskiila.

Lee, J. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th annual meeting of the Association of Computational Linguistics (ACL2007), Prague, Czech Republic*, 472–479. https://www.aclweb.org/anthology/P07-1060

Leino-Kaukiainen, P. 1988. Kasvava sanomalehdistö sensuurin kahleissa 1890–1905 (The growing press at the times of censorship, 1890–1905, in Finnish). In *Suomen lehdistön historia. Sanomalehdistön vaiheet vuoteen 1905*, 421–632. Kuopio, Finland: Kustannuskiila.

Marjanen, J., V. Vaara, A. Kanner, H. Roivainen, E. Mäkelä, L. Lahti, and M. Tolonen. 2019. A national public sphere? Analyzing the language, location, and form of newspapers in Finland, 1771–1917. *Journal of European Periodical Studies* 4 (1):54–77. doi: 10.21825/jeps.v4i1.10483.

Muiluvuori, J. 2000. Yrjö Koskelainen. *Kansallisbiografia-verkkojulkaisu*. Studia Biographica 4. Suomalaisen Kirjallisuuden Seura. August 25, 2000. Accessed December 11, 2019. https://kansallisbiografia.fi/kansallis-biografia/henkilo/1625.

Mullen, L. 2016. America's public bible: Biblical quotations in U.S. newspapers, including website, code, and datasets. Accessed December 12, 2019. http://americaspublicbible.org/.

Oceanic Exchanges Project Team. 2017. *Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840–1914*. http://Osf.io/wa94s.

Oiva, M., A. Nivala, H. Salmi, O. Latva, M. Jalava, J. Keck, L. Martínez Domínguez, and J. Parker. 2019. Spreading News in 1904: The Media Coverage of Nikolay Bobrikov's Shooting. *Media History* 25: 1–17. doi: 10.1080/13688804.2019.1652090.

Oxford Living Dictionaries. 2019. Oxford University Press. Accessed December 12, 2019. https://en.oxforddictionaries.com/definition/viral.

Pääkkönen, T., J. Kervinen, A. Nivala, K. Kettunen, and E. Mäkelä. 2016. Exporting Finnish digitized historical newspaper contents for offline use. *D-Lib Magazine* 22 (7/8): 1–9. http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html. doi: 10.1045/july2016-paakkonen.

Paju, P. 2018. Ensimmäiset naiset insinöörien ja arkkitehtien yhdistyksissä (The first women in the societies of engineers and architects, in Finnish). *Tekniikan Waiheita* 36 (1):5–24.

Paju, P. 2019. International collaboration and Finland in the early years of computer-assisted history research: Combining influences from Nordic and Soviet Baltic historians. Proceedings of the 4th Digital Humanities in the Nordic Countries 2019. Copenhagen, Denmark, March 5–8, 2019. http://ceur-ws.org/Vol-2364/.

Pietilä, J. 2008. *Kirjoitus, juttu, tekstielementti. Suomalainen sanomalehtijournalismi juttutyyppien kehityksen valossa printtimedian vuosina 1771–2000 (A Writing, A Story, A Textual Element: The Finnish newspaper journalism in the light of print media, 1771–2000, in Finnish)*. Jyväskylä, Finland: Jyväskylä studies in humanities 111.

Prior-Miller, M. R. 2015. Magazine typology: Using operational classification theory. In *The routledge handbook of magazine research: The future of the magazine form*, ed.

David Abrahamson and Marcia R. Prior-Miller, 22–50. London: Routledge.

Rantala, H., H. Salmi, A. Nivala, P. Paju, R. Sippola, A. Vesanto, and F. Ginter. 2019. Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920 – Digitaalisten ihmistieteiden näkökulma (Text Reuse in Finnish Newspapers and Magazines, 1771–1920: A digital humanities approach, in Finnish). *Historiallinen Aikakauskirja* 1:53–67.

Rantanen, T. 1987. "STT:n uutisia" sadan vuoden varrelta ("STT News" – One hundred years, in Finnish). Helsinki, Finland: Weilin & Göös.

Rasila, V. 1973. Tilastomatemaattisten analyysien käyttömahdollisuudet lehdistötutkimuksessa (The possibilities of using mathematical statistics analysis in press research, in Finnish). In *Lehdistöntutkijain seminaari 1973 – alustukset ja keskustelut*, ed. Kristiina Ritari, 112–28. Helsinki, Finland: Helsingin yliopiston historian laitoksen julkaisuja 1.

Salmi, H. 2018. Viraalisuus – kulttuurihistoriallinen näkökulma (Virality: A cultural-historical approach, in Finnish). *Niin & Näin* 1 (2018):71–9.

Salmi, H., A. Nivala, H. Rantala, R. Sippola, A. Vesanto, and F. Ginter. 2018. Återanvändningen av text i den finska tidningspressen 1771–1853. *Historisk Tidskrift För Finland* 1:46–76.

Salmi, H., H. Rantala, A. Vesanto, and F. Ginter. 2019. The long-term reuse of texts in the Finnish Press 1771–1920. Proceedings of the 4th Digital Humanities in the Nordic Countries 2019, Copenhagen, Denmark, March 5–8, 2019. http://ceur-ws.org/Vol-2364/.

Silberstein-Loeb, J. 2012. Exclusivity and cooperation in the supply of news: The example of the associated press, 1893–1945. *Journal of Policy History* 24 (3):466–8. doi: 10.1017/S0898030612000140.

Smith, D. A., R. Cordell, and A. Mullen. 2015. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History* 27 (3): E1–E15. doi: 10.1093/alh/ajv029.

Smith, D. A., R. Cordell, and E. Maddock Dillon. 2013. Infectious texts: Modelling text reuse in nineteenth-century newspapers. Proceedings of the Workshop on Big Humanities, IEEE Computer Society Press, 86–94.

Suomalainen, M.-L. 1979. Sanomalehtien uutiskanavista 1860- ja 1870-luvuilla (The newspapers' channels of news in the 1860s and 1870s, in Finnish). *Lehdistöhistoriallisia tutkimuksia 1*. Helsinki, Finland: Suomen sanomalehdistö historia -projektin julkaisuja.

Teperi, J. 1977. *Sudet Suomen rintamaiden ihmisten uhkana 1800-luvulla (Wolves as a human threat in rural Finland during the nineteenth century, in Finnish).* Helsinki, Finland: Suomen Historiallinen Seura.

Tommila, P. 1988. Yhdestä lehdestä sanomalehdistöksi 1809–1859 (From one newspaper into a press, 1809–1859, in Finnish). In *Suomen lehdistön historia. Sanomalehdistön vaiheet vuoteen 1905*, 77–265. Kuopio, Finland: Kustannuskiila.

Vesanto, A. 2019. Detecting and analyzing text reuse with BLAST. MA thesis, Computer Science. University of Turku, Turku.

Vesanto, A., A. Nivala, H. Rantala, T. Salakoski, H. Salmi, and F. Ginter. 2017. Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910. Proceedings of the 21st Nordic Conference of Computational Linguistics, Gothenburg, Sweden, May 23–24, 2017. Linköping Electronic Conference Proceedings, 54–58. http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf

Vesanto, A., A. Nivala, T. Salakoski, H. Salmi, and F. Ginter. 2017. A system for identifying and exploring text repetition in large historical document corpora. Proceedings of the 21st Nordic Conference of Computational Linguistics, Gothenburg, Sweden, May 23–24, 2017. Linköping Electronic Conference Proceedings, 330–333. http://www.ep.liu.se/ecp/131/049/ecp17131049.pdf.

Zilliacus, C., and H. Knif. 1985. *Opinionens tryck. En studie över pressens bildningsskede i Finland (Weight of Opinion: A study of the construction phase of the press in Finland).* Helsingfors: Svenska litteratursällskapet i Finland.