# The Long-Term Reuse of Text in the Finnish Press, 1771–1920

Hannu Salmi<sup>1,3</sup>, Heli Rantala<sup>1,3</sup>, Aleksi Vesanto<sup>2,3</sup> and Filip Ginter<sup>2,3</sup>

<sup>1</sup> University of Turku, Dept of Cultural History
<sup>2</sup> University of Turku, Dept of Future Technologies
<sup>3</sup> Turku Group for Digital History

hannu.salmi@utu.fi, heli.rantala@utu.fi, aleksi.vesanto@utu.fi, filip.ginter@utu.fi

Abstract. This paper is based on the study of text reuse in the Finnish press from 1771-1920. In the Computational History and the Transformation of Public Discourse in Finland (COMHIS) project, we found 61 million occurrences of similarity, which formed 13.8 million clusters of reuse. This material also included strikingly slow processes of repetition, and the longest reuse cases were almost as long as the time span of the project. In sum, 2.03 million clusters, 15 per cent of the total amount, were longer than 12 months. As well, 76,259 clusters spanned over 20 years or more. The longest span was 146 years. The paper explores the volume and nature of this long-term text reuse in the Finnish press and analyses three distinctive features of slow repetition: newspapers as a site of memory, newspapers as an archive and the political ramifications of reuse. The paper argues that the habit of reprinting old texts aimed to bridge the gap between past and present, emphasising the continuity between old and new. On the other hand, there were cases where past texts were activated precisely for the opposite purposes, to obscure the past and to show how different the bygone world was.

**Keywords:** text reuse, digitised newspapers, long-term reuse, nineteenthcentury newspapers, text mining.

### 1 Introduction

In the project *Computational History and the Transformation of Public Discourse in Finland* (COMHIS) [1], we studied text reuse in the Finnish press. Our work is based on the digitised corpus of newspapers and magazines from the collection of the National Library of Finland. Our time span covers 149 years, starting from the publication of the first issue of the *Tidningar Utgifne af et Sällskap i Åbo* in 1771 and ending in 1920, which is the last year available for text mining so far. The corpus includes practically all published issues from that timeframe in Finland, in sum, 5.1 million pages [2].

There has been growing interest in the study of text reuse or, more generally, the overlaps between texts and text corpora. Lincoln Mullen analysed biblical quotations in American newspapers [3]; Marco Bühler et al. studied, for example, Homeric quotations in ancient texts [4]; the project *Knowledge, Information Technology and the* 

*Arabic Book* (KITAB) explored text reuse in the corpus of Arabic texts from the period 700–1500 [5]; and, in their *Viral Texts* project, Ryan Cordell and David A. Smith analysed text reuse in the American press in the nineteenth century [6]. The present Trans-Atlantic Platform project *Oceanic Exchanges*, led by Cordell, aims to study textual migration on a global scale and bring together researchers from six countries: Finland, Germany, Mexico, the Netherlands, the United Kingdom and the United States [7].

The culture of copying has many historical ramifications. In the nineteenth century, newspapers could circulate content that was, if we might say so, 'born journalistic', that is, written by the editors and authors of newspapers themselves. Papers could of course also reuse other kinds of textualities, from biblical quotations to telegrams, from paid advertisements to public announcements and from circulars to readers letters. Our COMHIS project has concentrated on those computational methods that can identify similar passages within the corpus of newspapers and magazines to allow for an analysis of different forms of text reuse within the Finnish press. In this article, we concentrate on one particular feature that has become evident in the project: the long-term reuse of texts. In nineteenth- and early twentieth-century Finland, viral news spread quite rapidly across and throughout the newspaper network, but there were also strikingly slow processes of repetition. The longest text reuse cases in the material are almost as long as the time span of the project. Our questions are therefore: What has been the significance of the long-term reuse of content, what were its main characteristics and why were particular texts republished over a long period? We start by discussing the overall character of this form of reuse; thereafter, we present three historically important features of long-term reuse.

# 2 Detecting Long-Term Reuse with BLAST

In the COMHIS project, we developed a special solution for detecting text reuse, which is based on NCBI BLAST (National Center for Biotechnology Information, Basic Local Alignment Search Tool), a software originally created for comparing and aligning biomedical sequences. In our method, we encoded the data into protein sequences, which then could be read by BLAST to identify regions in the sequences that overlap. These pairs were clustered based on their offset values so that overlapping passages were interpreted to be part of the same cluster. A cluster then contains all found occurrences of a particular reuse. These may be fragmented, i.e. one article is separated into different clusters, due to such issues as the original pagination of texts, a lack of accurate article segmentation or too-substantial OCR (optical character recognition) noise, causing BLAST to detect only a shorter part of the reuse.

The method has been described in an article in 2017 [8], and the software code, as well as the necessary guidelines, is openly available online [9]. This method proved highly effective, even in cases where OCR noise makes the text almost unreadable: according to our evaluation of the recall, 60 per cent of the characters could be wrong and the similarity could still be recognised [10]. In total, we found 61 million occurrences longer than 300 characters from the years 1771–1920, which formed 13.8 million clusters of reuse. It is important to note that the Finnish press has been published mostly in Finnish and Swedish, and using this method, text reuse could not be detect-

ed across languages. To be able to explore the results further, we combined our findings with a Solr search and indexing engine to be able to provide a web interface to browse the clusters of copied texts easily [11]. The open database of clusters is now online at comhis.fi. It is possible to search and to arrange the results, for example, based on date, length, count, span and gap.

From the perspective of long-term text reuse, the span of the cluster is especially important. In our database, span simply means the temporal distance between the first and last hit of a cluster. Gap, in turn, refers to a break in the repetition chain. In a case with only two hits, span and gap are equivalent. There can however be clusters with several gaps. To give an example, our database includes a cluster with 48 hits (ID 14099508). The span of this cluster is 100 years, from 1820 to 1920, 36,546 days to be exact. The original Finnish-language text was published in the *Turun Wiikko-Sanomat* on 8 January 1820, and it was written to promote reading among Finns. In our corpus, the last hit of the text is from the newspaper *Huittinen* from 30 January 1920. Over time, the text was printed in 40 different newspapers in 22 unique locations. The cluster included long gaps in its repetition chain. After its first publication, the text was published in 1851, after a gap of 31 years. It was again published 10 times in 1895 after a silence of 44 years, which is the maximum gap of the cluster. Thereafter, the text was republished several times in 1904, 1909, 1912, 1915 and 1920 (Fig. 1).



Fig. 1. The hit frequency of cluster 14099508 in the database *Text Reuse in Finnish Newspapers and Journals*, 1771–1920 (Source: comhis.fi)

In this example, text reuse unfolded over a period of 100 years, which can obviously be thought of as 'long-term' repetition [12]. However, what does 'long term' mean in our corpus? If we count the span and the gap of all clusters, we notice that 85 per cent of all clusters are shorter than one year. The major part of reuse clusters can thus be seen as fast repetition. In previous research, Smith, Cordell and Dillon (2013) divided the American dataset of reuse clusters into fast and slow texts. In their fast set, the median lag time was under one year [13]. In our analysis, the clusters whose span was under 12 months comprise the majority of all material. In addition to these fast chains, there were however clusters whose span varied significantly, up to over 140 years. Figure 2 shows the amount of clusters when span and gap are counted in 10-year timeframes.



Fig. 2. Text reuse clusters in relation to span and gap (Source: comhis.fi)

In total, 2.03 million clusters lasted longer than 12 months. While this on the other hand supports the idea that text reuse predominantly took place rather quickly, sometimes in a viral manner, it also shows that a substantial amount of clusters travelled longer over time. If we extract clusters whose span or gap varies between one and nine years, the amount is still counted in millions. However, after a gap or span of 10 years, the amount starts to diminish. In 259,757 clusters, the span is 10 years or more. In turn, 76,259 clusters spanned over 20 years or more, and 5,888 clusters spanned over 50 years. Furthermore, 289 texts travelled over a century. The longest span in the database is 146 years, and the longest gap is 145 years.

We will continue by discussing three historically distinctive features of long-term reuse in the Finnish press. These are not categories of the most common cases but, in our view, historically relevant aspects of long-term reuse that need to be considered when analysing the cultural ramifications of the press in the nineteenth century. Furthermore, they exemplify the potential of computational methods in assessing the role of the press over a long period.

#### **3** Site of Memory

The volume and capacity of the Finnish press expanded towards the end of the nineteenth century. Information circulated with an ever-accelerating rhythm, which becomes obvious in our database of reuse. The majority of reuse clusters started at the end of the period. There were, as already pointed out, also texts that seem to have floated in time, and, evidently, the newspapers not only republished texts that had been recently printed, but they also drew on the volumes of previous years and decades. This is reminiscent of the notion of *lieux de mémoire*, proposed by the French historian Pierre Nora: there are sites of memory that participate in constructing a sense of community. These sites are not necessarily physical places but also symbolic sites that bridge the gap between past and present [14]. Although newspapers, or 'dai-lies', can be interpreted as epitomes of the present tense, the now, they have also become sites of memory. Newspapers filtered themes from the past to the present and participated in constructing cultural memory. At the same time, papers were not only a repository for remembering issues outside the press; the history of newspapers themselves also became an object of memory.

Already, the first papers of the late eighteenth century offered texts for later reuse. A poem by two professors of the Academy of Turku, Pehr Adrian Gadd (1727–1797) and Henrik Gabriel Porthan (1739–1804), was published in the first issue of the *Tidningar Utgifne af et Sällskap i Åbo* on 15 January 1771. This poem was republished in six papers between 1853 and 1871. When the *Litteraturblad för allmän medborgerlig bildning* printed the poem in 1853, it was contextualised with a longer article on the emergence of the Finnish press in 1771. On 9 November 1861, the *Helsingfors Tidningar* framed the poem by telling of Henrik Gabriel Porthan and his impact on the Finnish nation. In this case, the publication was motivated by the fact that 1861 marked the 30th anniversary of the Finnish Literature Society, which had been founded in 1831. Gadd and Porthan's poem was again reprinted in 1871 to commemorate the centenary of Finnish newspaper publishing, and the poem was offered to readers by the *Hufvudstadsbladet* on 15 January 1871.

The editors of the newspapers were conscious of the history of the medium. Another example is offered by the first Finnish-language paper *Suomenkieliset Tieto-Sanomat*. Its sample issue was published on 1 September 1775, including an article on the nature of the Finnish language and the importance of using it for publishing activities. This text was the first hit of cluster 11250221 that included 21 occurrences (Fig. 3). The span of the cluster was 145 years.



Fig. 3. The hit frequency of cluster 11250221 in the database *Text Reuse in Finnish Newspapers and Journals*, 1771–1920 (Source: comhis.fi)

As the previous examples show, newspapers participated in commemoration practices by revitalising issues from the past, for example, after 50 or 100 years. The habit of remembering the past in cycles is not a self-evident cultural practice. It seems that in newspapers, these cycles became increasingly important during the nineteenth century.

Gaps within clusters can be seen as indicative of these commemoration practices, although it must be noted that there can be several gaps in one cluster. Still, the cycles stand out clearly in the material when the amount of clusters is counted for each gap in years. For example, there were 4,216 clusters with a gap of 25 years, but only 2,214 clusters with a gap of 26 years. Figure 4 shows the amount of clusters according to the length of the gap, from 39 years to 105 years. Gaps of 40 and 50 years are the highest peaks, and there is a shorter peak indicating a gap of 100 years. These cases of commemoration included various kinds of materials, such as articles with topics ranging from agriculture to rare natural phenomena. Some news items were also later marked as historical events. In the *Turun Wiikko-Sanomat*, published on 15 January 1820, a short notice reported that Napoleon Bonaparte had been imprisoned on the island of St. Helena. The same story was republished on the centenary of this historic moment in 1920.



**Fig. 4.** The amounts of clusters per gap (gap =  $39 \rightarrow \text{gap} = 105$ ) (Source: comhis.fi)

# 4 Archive of Anecdotes

In press history, or more widely in the study of print culture, the focus of study has often been on individual authors or genres of writing [15]. However, nineteenth-century newspapers published a mixture of content, including different kinds of notices, snippets, anecdotes and anonymous stories. Cordell (2015) has argued that these forms of 'every day reading and writing' have been an ignored genre in press history. According to Cordell, the circulation of different kinds of anonymous texts was an important part of the American nineteenth-century pre-Civil War press [16]. One can find this similar practice—the reprinting of anonymous stories and anecdotes—in the Finnish system of text reuse, including in long-term cases.

Various kinds of stories and anecdotes travelled through time within the Finnish press. It seems the editors of newspapers around 1900 used earlier papers as an archive of possible material to be employed and reprinted. One motivation for this kind of selection of old stories to be circulated was likely the need to fill the columns of the

539

paper, as the capacity of the press increased rapidly. On the other hand, the editors had no doubt been looking for stories that would have entertainment value among readers.

Among the long-term reuse clusters, several cases of both Swedish- and Finnishlanguage anonymous stories have been reprinted various times in different temporal rhythms. For example, late nineteenth-century papers showed interest in a prayer that was published in the *Åbo Tidningar* in February 1791. The speciality of this prayer was the fact that it was reported to be a prayer for married women. Its content included passages where a wife was affirming her obedience to her husband. This prayer was republished by seven Swedish-language newspapers, first in 1862 and 1877 and then three times during the 1880s and finally in 1890. The original context of the prayer, according to the *Åbo Tidningar*, was a divorce case in Germany during the 1780s, in which the wife had complained that she was forced to pray each night in the presence of her husband. The prayer was apparently copied and translated from a German magazine, as the same prayer was published in the *Berlinische Monatsschrift* in 1784 [17]. In our database, the original story has been dispersed into three overlapping clusters (6317218, 1694951 and 3722889).

In addition, Finnish language anecdotes and stories were shared over a long period. On 2 January 1830, the *Oulun Wiikko-Sanomia* published a story about the spontaneous meeting between a former prisoner and his ex-guard. In this story, the guard is afraid of the possible revenge of the prisoner but finds that the ex-convict only wanted to buy him a beer. Seventy-four years later, this entertaining story was rediscovered and republished by several newspapers. The first reprint was in the Oulu-based *Kaiku* on 10 February 1904. After the first reprint, six other papers circulated the same story during February and March.

In some cases, the publishing of old stories from past papers created distance between the past and present. For instance, on 12 February 1820, the Finnish-language Turun Wiikko-Sanomat published a news clip reportedly taken from a 'paper in St. Petersburg', which in turn had used a German newspaper as a source. This newsprint was about large sea serpents that had surrounded a ship somewhere in the open sea. According to the story, the mariners had shot one of the creatures. The editor of the Turun Wiikko-Sanomat did not want to take any stance concerning the truth-value of this report. He only stated that this is not the only newspaper report on sea serpents. Seventy-eight years later, on 5 November 1898, the Keski-Suomi reprinted this old text and introduced it as an example of a canard from the past. The next time the same piece of news was found in 1920. Within a couple of weeks, 12 different papers published the old story about sea serpents. The case shows, we argue, that in 1820, the story about giant sea snakes seen in distant waters was still within the possibilities of the (un)known world [18]. Towards the end of the century, however, the situation changed, and stories about sea serpents were categorised as imaginary stories from the past. Here, long-term reuse aimed to create distance from the past.

# 5 Political Dimensions of Reuse

Old newspaper material was republished for political reasons as well. We have found examples in which newspapers and periodicals around the turn of the nineteenth century benefited from the material of older papers when commenting on the state of affairs in their contemporary world.

In 1776, the *Suomenkieliset Tieto-Sanomat* published an announcement by the Royal Swedish Academy of Sciences on a prize awarded to a peasant who seeded the largest amount of root crops in one field. The motivation behind the prize was to promote innovations in agriculture, including crop versatility (instead of only grain crop). This announcement was found 122 years after the first publication, and the span of this cluster of reprints is quite high at 142 years (cluster ID 6317820, Fig. 5). In most reprints (total count value is 19), the old text was simply copied without any further comments. In 1908, however, the working-class paper *Sosialisti* added its own interpretation when publishing the old newspaper clip. According to the *Sosialisti*, this case—the awarding of a prize to a hard-working farmer—only showed how 'Counts' and 'Barons' took advantage of the work of the peasantry and how the Crown wanted to expand its lands by making the peasants clear (and burn over) new areas of woodland for cultivation. This way, an eighteenth-century prize call, motivated by the Enlightenment ideals of the advancement of knowledge, became a part of socialist interpretation of the historical class struggle.



Fig. 5. The hit frequency of cluster 6317820 in the database *Text Reuse in Finnish Newspapers* and *Journals*, 1771–1920 (Source: comhis.fi)

Another example shows how the 'words' of Emperor Alexander I were being reprinted and interpreted by late nineteenth- and early twentieth-century newspapers. In this case, the source was the *Åbo Allmänna Tidning* from 21 April 1810, which published an official notice from the Emperor directed to the people of Finland. In this notice, Alexander assured that he would rule the land by honouring its freedom, constitution and rights. This formulation was quite near to the words Alexander directed to the participants of the Diet of Porvoo in 1809, gathered after Finland's annexation to Russia [19]. During the second half of the nineteenth century, the *Helsingfors Tidningar* reprinted passages of this notice in 1861 (20 August) as a part of a polemic the paper was having on the interpretations of the Diet of Porvoo with another newspaper. At this point, the Finnish Diet had not been summoned for over 50 years, and the discussion of Alexander's words was a way to comment on this situation. Later on, pieces of the same text were reprinted between 1871 and 1913 by several papers (cluster ID 5351089: total number of unique titles 14) in different connections.

541

#### 6 Conclusion

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

We have discussed three features of long-term text reuse. Newspapers were a site of memory, where past phenomena were activated for present readers through the process of reprinting. The press also participated in consolidating mnemonic operations that were cyclic in nature: texts were reprinted 25, 50 or even 100 years after their original publication. The second feature is that newspapers served as an archive for the editors. They offered raw material for the expanding press that simply needed texts to be published. Old papers were a repository of stories, anecdotes and other ephemera often neglected by newspaper historians. This material lived long, as our examples indicate. The third feature deals with the political ramifications of reuse. This was particularly important in Finland, which was a part of the Russian Empire until 1917, and the editors often tried to find ways of circumventing censorship regulations. Sometimes, the original political intentions were turned upside down, and the quotations could be used to criticise past conditions.

The cases of long-term text reuse could further be explored by concentrating on the strategies of bringing past news to the present. Based on our research, it is obvious that the habit of reprinting old texts aimed to bridge the gap between past and present, emphasising the continuity between old and new. On the other hand, there were also cases where past texts were activated precisely for the opposite purpose, to obscure the past and to show how different the bygone world was. This tension reveals the ambiguous nature of the press as a vehicle of modernisation in the nineteenth century.

#### Acknowledgments

The work was supported by the research consortium *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (293239) and the research project *Viral Culture in Early Nineteenth-Century Europe* (302433), funded by the Academy of Finland. Computational resources were provided by the CSC–IT Centre for Science, Espoo, Finland, and the data by the National Library of Finland.

## References

- The consortium *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (2016–2019) is funded by the Academy of Finland programme on Digital Humanities. The consortium is based on the cooperation of four partners: The Faculty of Humanities at the University of Helsinki, the Departments of Cultural History and Future Technologies at the University of Turku and the Centre for Preservation and Digitisation of the National Library of Finland. For further details, see comhis.fi. Previous publications of the project include, for example, Salmi H, Nivala A, Rantala H, Sippola R, Vesanro A, Ginter F (2018) Återanvändningen av text i den finska tidningspressen 1771–1853. Historisk Tidskrift för Finland 1/2018: 46–76; Nivala A, Salmi H, Sarjala J (2018) History and virtual topology: The nineteenth-century press as material flow. Historein 17(2). doi: http://dx.doi.org/10.12681/historein.14612.
- 2. On the corpus, see Oiva M, Salmi H, Nivala A (2018) Digitized newspapers at the National Library of Finland, Suomen kansalliskirjaston digitoidut sanomalehdet, Oceanic Exchanges, Data Reports. https://oceanicexchanges.org/2018-02-20-data-reports-finland/. It must be noted that in 1771, Finland was part of the Kingdom of Sweden, and the first Finnish paper was launched during the expansion of the Swedish press. Furthermore, the corpus also includes magazines published outside of Finland, such as Finnish-American journals, but it does not include Finnish-American newspapers.
- 3. Mullen L (2016) America's public bible: Biblical quotations in U.S. newspapers. http://americaspublicbible.org/.
- 4. Büchler M, Crane G, Moritz M, Babeu A (2012) Increasing recall for text re-use in historical documents to support research in the humanities. In: (Proceedings) Second International Conference on Theory and Practice of Digital Libraries, vol 7489, pp 95–100.doi: 10.1007/978-3-642-33290-6\_11. On text reuse, see also, Franzini G, Franzini E, Büchler M (2016) Historical text reuse: What is it? http://www.etrap.eu/historical-text-reuse/.
- 5. The KITAB project. http://kitab-project.org/text-reuse-methods/.
- Smith DA, Cordell R, Maddock Dillon E (2013) Infectious texts: Modeling text reuse in nineteenth-century newspapers. www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf.
- For further details, see Oceanic exchanges: Tracing global information networks in historical newspaper repositories, 1840–1914. https://oceanicexchanges.org/.
- Vesanto A, Nivala A, Rantala H, Salakoski T, Salmi H, Ginter F (2017) Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771–1910. In: Proceedings of the 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017, pp 54–58. http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.
- 9. Text reuse detection with BLAST: A software to detect text reuse with BLAST. https://github.com/avjves/textreuse-blast.
- Vesanto A, Nivala A, Rantala H, Salakoski T, Salmi H, Ginter F (2017) Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771–1910. In: Proceedings of the 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017, p 56. http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.
- Vesanto A, Nivala A, Rantala H, Salakoski T, Salmi H, Ginter F (2017) Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771–1910. In: Proceedings of the 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23–24 May 2017, pp 330–333. http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf.
- 12. In this case, the chain of repetition ends in 1920, as this is the last year in our corpus. A quick look at the scanned newspapers after 1920 indicates that the same text continued to be copied after that year. The original text from 1820 can be found, for example, in *Sa*-

543

takunnan Kansa from 15 January 1926, Karjalainen from 15 September 1928, Etelä-Suomen Sanomat from 19 March 1958 and Länsi-Savo from 20 December 1963.

- Smith DA, Cordell R, Maddock Dillon E (2013) Infectious texts: Modeling text reuse in nineteenth-century newspapers, p 8. www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf.
- 14. Nora P (1997) General introduction: Between memory and history. In: Nora P (ed) The realms of memory: Rethinking the French part, Vol. 1: Conflicts and divisions. Columbia University Press, pp 1–21.
- Cordell R (2015) Reprinting, circulation, and the network author in antebellum newspapers. American Literary History 27(3): 417–445; Jarlbrink J, Snickars P, Colliander C (2016) Maskinläsning: om massdigitalisering, digitala metoder och svenska dagspress. Nordicom Information 38(3): 27–40.
- 16. Cordell R (2015) Reprinting, circulation, and the network author in antebellum newspapers. American Literary History 27(3): 417–445.
- 17. Berlinische Monatsschrift 2. Bd (1784). Digital Collections of the Bielefeld University Library. http://ds.ub.uni-bielefeld.de/viewer/image/2239816\_004/375/LOG\_0050/.
- 18. On the imagination of the unknown in the nineteenth century, see Latva O (2019) The Giant Squid: Imagining and Encountering the Unknown from the 1760s to the 1890s. Doctoral dissertation. Cultural History, University of Turku.
- Kurunmäki J (2017) Political representation, imperial dependency and political transfer: Finland and Sweden 1809–1819. Journal of Modern European History 15(2): 243–260.