Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910

Aleksi Vesanto
Turku NLP Group
Department of FT
University of Turku
aleksi.vesanto@utu.fi

Asko Nivala
Cultural History and
Turku Institute for Advanced Studies
University of Turku
asko.nivala@utu.fi

Heli Rantala
Cultural History
University of Turku
heli.rantala@utu.fi

Tapio Salakoski
Turku NLP Group
Department of FT
University of Turku
tapio.salakoski@utu.fi

Hannu Salmi
Cultural History
University of Turku
hannu.salmi@utu.fi

Filip Ginter
Turku NLP Group
Department of FT
University of Turku
filip.ginter@utu.fi

Abstract

We present the results of text reuse detection, based on the corpus of scanned and OCR-recognized Finnish newspapers and journals from 1771 to 1910. Our study draws on BLAST, a software created for comparing and aligning biological sequences. We show different types of text reuse in this corpus, and also present a comparison to the software Passim, developed at the Northeastern University in Boston, for text reuse detection.

1 Introduction

The dataset of the National Library of Finland (NLF) contains 1.95 million pages of digitized historical newspapers and journals from 1771 to 1910. Approximately half of the content is in Swedish, the other half in Finnish, although there are also a few German and Russian papers included (Pääkkönen et al., 2016). We aim to trace the influential texts that were copied and recirculated in Finnish newspapers and journals in this time period. This is done by clustering the 1771–1910 NLF corpus with a text reuse detection algorithm. Our approach enables us to study the dissemination of news and other information and to reconstruct the development of the newspaper network as a part of Finnish public discourse: What kinds of texts were widely shared? How fast did they spread and what were the most important nodes in the Finnish media network?

Our research project builds on a similar study of nineteenth-century US newspapers by Ryan Cordell, David A. Smith and their research group (Cordell, 2015; Smith et al., 2015). However, in contrast to the US press, the nineteenth- and early twentieth-century Finnish newspapers were typically printed in the *Fraktur* typeface, which (together with other possible sources of noise) poses unusual difficulties for Optical Character Recognition (Kettunen, 2016). To solve this problem, we have developed a novel text reuse detection solution based on BLAST (Vesanto et al., 2017) that is accurate and resistant to OCR mistakes and other noise, making the text circulation and virality of newspaper publicity in Finland a feasible research question.

2 Detecting Text Reuse

In the nineteenth century, contemporaries saw newspapers as reflections of modern culture. Many phenomena were amplified by the increasing power of the press, including urbanization, consumerism, and business life. The changes in transport technology led to more efficient distribution of information. Before 1880s, there was no copyright agreement to regulate the free copying of texts, which became a distinctive feature of the press. To understand this process, it is essential to analyze how texts were copied and reprinted.

In Finland, newspaper publishing started slowly, the first paper being *Tidningar Utgifne af et Sällskap i Åbo* in 1771. According to the NLF metadata, in 1850 there were only ten papers and six journals. A rapid upheaval occurred at the

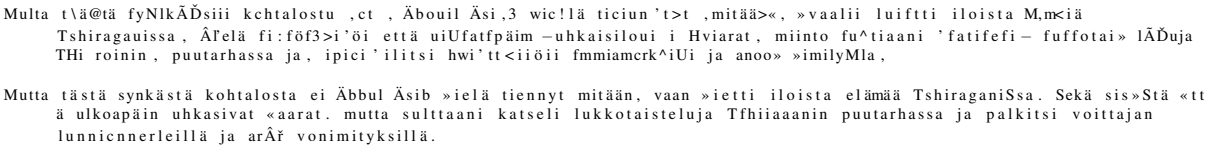

Figure 1: Example of how low the OCR quality can be. Both passages are identical in the original issues.

end of the century, resulting in 89 papers and 203 journals in 1900. The volume of the press was thus very limited during the first half of the century, which also means that text reuse was small-scale. Towards the end of the period the situation changed dramatically, offering more volume for viral chains of reprints. These chains had different origins: they were internal chains within the press, translations from abroad, stories from books, telegrams, or official announcements. Therefore, text reuse detection can shed essential light on how information flowed between centers within the country and how, in the end, Finnish press participated in the global circulation of information.

The primary obstacle in detecting text reuse in the NLF dataset is the poor OCR recognition rate, as illustrated in Figure 1. This makes any approach which assumes exact seed overlaps of several words in length infeasible, and calls for a fuzzy matching method highly tolerant to noise. To this end, we have applied BLAST (Altschul et al., 1990), a sequence alignment software developed for fast matching of biological sequences against very large sequence databases. The main features of BLAST are speed and the ability to retrieve also distantly related sequences – which in our case translates to the ability to withstand the OCR noise present in the data. We index each page of the NLF data as a sequence in BLAST, translating the 23 most common lowercase letters into the amino-acid sequences which BLAST is hard-coded to handle, and subsequently matching the pages in an all-against-all scenario, and post-processing the results to recover the repeated text segments. We choose not to describe the technical details of the process in this paper, and rather focus on the results obtained. The implementation will be made available as open-source software and in the following, we focus on presenting the main results in context of processing historical texts.

System	% of text
BLAST	0.177
Passim (default)	0.057
Passim (optimized)	0.080

Table 1: Text reuse recall comparison of the BLAST-based method relative to Passim with its settings left at their default values, as well as optimized to maximize recall.

3 Text Reuse Clusters – Quantitative Analysis

In total, we found around 8 million clusters of repeated texts that have a total of 49 million occurrences (hits) longer than 300 characters. Note, however, that some clusters refer to the same, larger repeated news piece, in different lengths. This is due to the fact that at times the OCR quality is too low, allowing only for a shorter hit to be identified in some of the repetitions of an otherwise larger text. Since the surrounding text of a shorter hit is too dissimilar (which, after all, is the very reason why only a shorter segment was found), it is difficult to establish whether these clusters can be safely merged. Therefore, the number of found hits does not necessarily fully correspond to the number of unique text reuse.

3.1 BLAST evaluation

As there is not a feasible manner in which to directly estimate the recall of the system on this data, we compare our system to *Passim*, a popular tool for text reuse detection (Smith et al., 2014) used in many similar studies previously, so as to establish a relative comparison to the state-of-the-art. We form a dataset of 2,000 randomly selected documents from the NLF corpus, apply both systems to it, and for every document, we calculate the fraction of its text that was identified as text reuse, with a text length minimum set to 100. The results are summarized in Table 1, and demonstrate a substantial recall gain of the BLAST-based method.

We can see that the BLAST-based method

vastly outperforms Passim in terms of recall. In order to establish that this gain in recall is not at the expense of precision, we sample clusters both randomly and at the very bottom of BLAST similarity scores still acceptable for inclusion in the results and manually verify the proportion of those that are true positives. The proportions are shown in Table 2. The results naturally depend on the length of the texts in the cluster, with shorter texts less likely to be correct hits than the longer ones, given a constant alignment score.

Range	BLAST Precision random	BLAST Precision low	BLAST Coverage	Passim Coverage
300 - 350	1.00	1.00	0.108	0.076
250 - 299	1.00	0.94	0.120	0.078
200 - 249	0.94	0.94	0.133	0.079
150 - 199	0.92	0.86	0.154	0.080
100 - 149	0.86	0.70	0.177	0.080

Table 2: The precision and coverage of the BLAST method on 50 clusters of varying text hit lengths, sampled randomly and at the lowest alignment scores acceptable.

To understand to what extent the hits identified as text re-use are dissimilar, we randomly selected 1000 clusters which contain only two hits of at least 300 characters in length. We then calculate the pairwise character alignment between these two hits and measure the proportion of matching characters, i.e. not gaps nor misalignments. As shown in Figure 2, the alignment values range from around 99% down to as low as 40%, with the bulk of the data in the 70–90% range. For the most part, the repeated texts thus differ in 10–30% of positions, but the difference can be as much as 60%. Partly, these are cases of e.g. advertisements which differ only in numerical values, but partly these are in fact fully identical texts with a massive OCR error rate.

The gain in recall comes at the expense of compute time, with BLAST being about three orders of magnitude slower than Passim. Applying BLAST to the entire NLF dataset required around 150,000 CPU-core hours. This is certainly out of reach for a single computer, but well within modern cluster computing resources, especially since the historical text collection is static and the run only needs to be carried out once.

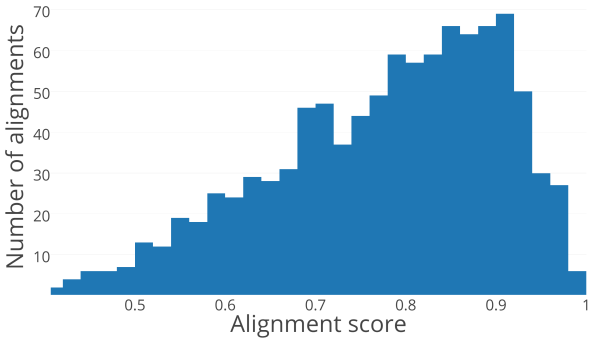

Figure 2: The distribution of alignment scores (horizontal axis) and the number of clusters out of 1000 with the given alignment score (vertical axis). Minimum text reuse length is 300.

4 Text Reuse Clusters – Qualitative Analysis

Copying and reprinting texts from other newspapers took three different forms, which is why we need to differentiate between text reuse, long-term reuse and virality. First, the majority of reprinted clusters consists of advertisements and notices, official announcements and ecclesiastical material. Second, many clusters include old news items, anecdotes, stories and poems that are suddenly reprinted many decades – sometimes even a hundred years – later. This second group is an example of longitudinal text reuse. Finally, the third group is viral news proper. The amount of viral news increases towards the end of the nineteenth century and these texts are often reprinted very rapidly within a short time frame.

4.1 Advertisements and announcements

The first group of clusters, advertisements and announcements, might be interesting sources in their own right for specific research questions. But above all, the changes in their amount tell us a lot about the scope of the public communication network and its historical development year by year even if we completely overlook the content of shared texts. For instance, by importing all text clusters as nodes and edges to a network analysis software, one is able to produce visualizations of the development of relationships between the newspapers and show what were the most dominating nodes in the network.

4.2 Longitudinal text reuse

Because of the wide time frame of the NLF dataset, long-term text reuse opens up an important perspective on historical memory. For instance, the Fennomans were an influential group in the nineteenth-century publicity of the Grand Duchy of Finland. The papers published around the turn of the nineteenth century often reprinted old articles and shorter quotations that supported their cause. To give an example of this, the Table 3 shows the reprints of a patriotic student song "Ännu på tidens mörka vågor" that was probably written by Gustaf Idestam (1802–1851) and first printed in *Åbo Morgonblad* in 3 March 1821 – a newspaper edited by the polemical Romanticist author Adolf Ivar Ardwidsson (1791–1858). The lyrics of the song were then reprinted three times in 1891, crediting the Swedish humorous magazine *Söndags-Nisse* as their source in addition to Arwidsson's paper. The song is also described in *Nya Pressen* as the favourite anthem of the Turku students before the introduction of "Vårt land", the later national anthem. We have found many other similar examples that show the way in which much earlier historical texts were reused for political purposes, opening up an important research question for the strategies of Finnish nationalism.

One example explicitly connected with the state of the Finnish press is the closing down of Arwidsson's newspaper *Åbo Morgonblad* by the officials in October 1821. The reasons for this act were political since Arwidsson had been calling for the wide freedom of the press in his paper. In the last issue of *Åbo Morgonblad* Arwidsson published the document on the official decision for the act. In 1891, 70 years later, this text was reprinted by five different newspapers. The first reprint was in *Åbo Tidningar* (30 September 1891), which used the censorship case of 1821 to discuss the state of the press freedom in 1891 – the censorship law was tightened in Autumn 1891. Other newspapers continued this discussion by reprinting the document from 1821 and also commenting the state of the censorship in 1891. This way the reuse of old news item offered a way to discuss and criticize the situation of the press freedom in 1891.

4.3 Viral news

The third group of text reuse are the actual viral news. According to our preliminary survey of the clustered NLF newspaper corpus, their amount increases rapidly after the The Crimean War (1853–1856). For instance, a bank robbery in Helsinki broke the news in 20 newspapers in 1906. This item was disseminated very rapidly in the Finnish-language press, as is shown in the Table 4. Only in six days it traveled from the urban communication hubs like Helsinki, Turku, Viipuri and Tampere to smaller towns in Ostrobothnia, Savonia, Karelia and Lapland. The viral chain served the need to rapidly tell about a current incident, although this happened without any particular plan, through the existing network of newspapers.

Cluster	Date	Title
639828	1821-03-03	Åbo Morgonblad
639828	1891-02-20	Nya Pressen
639828	1891-02-20	Folkwännen
639828	1891-02-21	Åbo Tidning

Table 3: Reprints of a patriotic song.

The three categories of text reuse could also overlap. Longitudinal chains, for example, might later on transform into viral texts. Old stories or anecdotes could be reactivated after several decades and reused in an infectious manner. BLAST is effective in revealing these different temporal rhythms of text reuse.

Place	Date	Title
Helsinki	1906-11-07	Uusmaalainen
Helsinki	1906-11-07	Helsingin Sanomat
Turku	1906-11-08	Uusi Aura
Helsinki	1906-11-08	Elämä
Tampere	1906-11-08	Tampereen Sanomat
Turku	1906-11-08	Sosialisti
Helsinki	1906-11-08	Uusi Suometar
Jyväskylä	1906-11-09	Suomalainen
Oulu	1906-11-09	Kaleva
Kuopio	1906-11-09	Pohjois-Savo
Tampere	1906-11-09	Kansan Lehti
Viipuri	1906-11-09	Karjala
Sortavala	1906-11-10	Laatokka
Heinola	1906-11-10	Heinolan Sanomat
Savonlinna	1906-11-10	Keski-Savo
Joensuu	1906-11-10	Karjalatar
Lahti	1906-11-11	Lahden Lehti
Kemi	1906-11-12	Pohjois-Suomi
Kristiina	1906-11-12	Etelä-Pohjanmaa
Lahti	1906-11-13	Lahti

Table 4: Reprints of a bank robbery news.

5 Conclusion

We have presented the use of the BLAST method to analyze text reuse in a massive corpus of historical newpapers of poor OCR quality. We have shown that, given sufficient computational power, the method is capable of identifying reprinted text passages that, due to OCR noise, may differ in up to 60% characters when aligned. Analysis of the clusters discovered by the method provides us with new insights into the magnitude and different types of text reuse, and reveals a number of individual examples of historical interest. As a future work, we will strive to develop a text classifier of the different types and topics of text reuse to be able to provide their quantitative analysis. The software developed to carry out the study will be made publicly available as open-source.

Acknowledgments

The work was supported by the research consortium *Computational History and the Transformation of Public Discourse in Finland, 1640-1910*, funded by the Academy of Finland. Computational resources were provided by CSC — IT Centre for Science, Espoo, Finland.

References

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct.

Ryan Cordell. 2015. Reprinting, Circulation, and the Network Author in Antebellum Newspapers. *American Literary History*, 27(3):417–445.

Kimmo Kettunen. 2016. Keep, change or delete? setting up a low resource ocr post-correction framework for a digitized old finnish newspaper collection. In D. Calvanese, D. De Nart, and C. Tasso, editors, *Digital Libraries on the Move. IRCDL 2015. Communications in Computer and Information Science*, volume 612. Springer, Cham.

Tuula Pääkkönen, Jukka Kervinen, Asko Nivala, Kimmo Kettunen, and Eetu Mäkelä. 2016. Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. *D-Lib Magazine*, 22(7).

David A. Smith, Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. Detecting and modeling local text reuse. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 183–192, Piscataway, NJ, USA. IEEE Press.

David A. Smith, Ryan Cordell, and Abby Mullen. 2015. Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, 27(3):E1–E15.

Aleksi Vesanto, Asko Nivala, Tapio Salakoski, Hannu Salmi, and Ginter Filip. 2017. A system for identifying and exploring text repetition in large historical document corpora. In *Proceedings of NoDaLiDa 2017*.