Metadata Analysis and Text Reuse Detection: Reassessing public discourse in Finland through newspapers and journals 1771–1917

Poster at DHN 2018 conference

Presenters: Ginter, Filip (1); Kanner, Antti (2); Lahti, Leo (1); Marjanen, Jani (2); Mäkelä, Eetu (2); Nivala, Asko (1); Rantala, Heli (1); Salmi, Hannu (1); Sippola, Reetta (1); Tolonen, Mikko (2); Vaara, Ville (2); Vesanto, Aleksi (2)

Organisation(s): 1: University of Turku; 2: University of Helsinki

During the period 1771–1917 newspapers developed as a mass medium in the Grand Duchy of Finland. This happened within two different imperial configurations (Sweden until 1809 and Russia 1809–1917) and in two main languages – Swedish and Finnish. The *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (COMHIS) project studies the transformation of public discourse in Europe and in Finland via an innovative combination of original data, state-of-the-art quantitative methods that have not been previously applied in this context, and an open source collaboration model.

In this study the project combines the statistical analysis of newspaper metadata and the analysis of text reuse within the papers to trace the expansion of and exchange in Finnish newspapers published in the long nineteenth century. The analysis is based on the metadata and content of digitized Finnish newspapers published by the National library of Finland. The dataset includes full text of all newspapers and most periodicals published in Finland between 1771 and 1920. The analysis of metadata builds on data harmonization and enrichment by extracting information on columns, type sets, publications frequencies and circulation records from the full-text files or outside sources. Our analysis of text reuse is based on a modified version of the Basic Local Alignment Search Tool (BLAST) algorithm, which can detect similar sequences and was initially developed for fast alignment of biomolecular sequences, such as DNA chains. We have further modified the algorithm in order to identify text reuse patterns. BLAST is robust to deviations in the text content, and as such able to effectively circumvent errors or differences arising from optical character recognition (OCR).

By relating metadata on publication places, language, number of issues, number of words, size of papers, and publishers and comparing that to the existing scholarship on newspaper history and censorship, the study provides a more accurate bird's-eye view of newspaper publishing in Finland after 1771. By pinpointing key moments in the development of journalism the study suggest that while the discussions in the public were inherently bilingual, the technological and journalistic developments advanced at different speeds in Swedish and Finnish language forums. It further assesses the development of the press in comparison with book production and periodicals, pointing towards a specialization of newspapers as a medium in the period post 1860. Of special interest is that the growth and specialization of the newspaper medium was much indebted to the newspapers being established all over the country and thus becoming forums for local debates.

The existence of a medium encompassing the whole country was crucial to the birth of a national imaginary. Yet, the national public sphere was not without regional intellectual asymmetries. This study traces these asymmetries by analysing text reuse in the whole newspaper corpus. It shows which papers and which cities functioned as "senders" and "receivers" in the public discourse in this period. It is furthermore essential that newspapers and periodicals had several functions throughout the period, and the role of the public sphere cannot be taken for granted. The analysis of text reuse further paints a picture of virality in newspaper publishing that was indicative of modern journalistic practices but also reveals the rapidly expanding capacity of the press. These can be further contrasted to other items commonly associated with the birth of modern journalism such as publication frequency, page sizes and typesetting of the papers.

All algorithms and software will be made openly available online, and can be located through the project's repositories (https://comhis.github.io/ and https://github.com/avjves/textreuse-blast). The results of the text reuse detection carried out in BLAST are stored in a database that has already been opened at http://comhis.fi.