Metadata analysis and text reuse

Reassessing public discourse in Finland through newspapers and journals, 1771–1917

Filip Ginter[1], Antti Kanner[2], Leo Lahti[1], Jani Marjanen[2], Eetu Mäkelä[2], Asko Nivala[1], Heli Rantala[1], Hannu Salmi[1], Reetta Sippola[1], Mikko Tolonen[2], Ville Vaara[2], Aleksi Vesanto[1]
[1]University of Turku, [2]University of Helsinki

Introduction

During the period 1771–1917 newspapers developed as a mass media in the Grand Duchy of Finland. This happened in two main languages – Swedish and Finnish. The Computational History and the Transformation of Public Discourse in Finland, 1640–1910 (COMHIS) provides a bird's-eye view of newspaper publishing by tracing the spread of newspapers throughout the country in both languages.

Materials and methods

The analyses are based on Finnish newspapers published by the National Library of Finland, which includes metadata and full text of all newspapers and most periodicals published in Finland between 1771 and 1920.

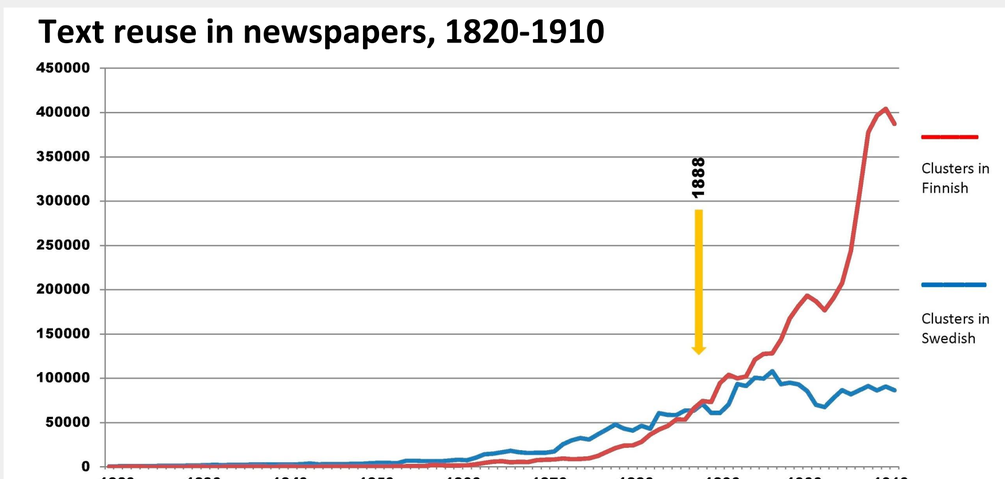
- The statistical analysis builds on metadata that has been harmonized and enriched. For comparisons with books, we rely on the Finnish National Bibliography (Fennica) published by the National Library.
- Text reuse analysis is based on a modified version of the Basic Local Alignment Search Tool (BLAST) algorithm, which detects similar sequences. It was initially developed for fast alignment of biomolecular sequences, such as DNA chains. BLAST is robust to deviations in text content and able to circumvent errors arising from optical character recognition (OCR).

Results

Relating metadata on publication places, language, number of issues, number of words, size of papers, and publishers with the text reuse analysis indicates key moments in the development of journalism. It shows that:

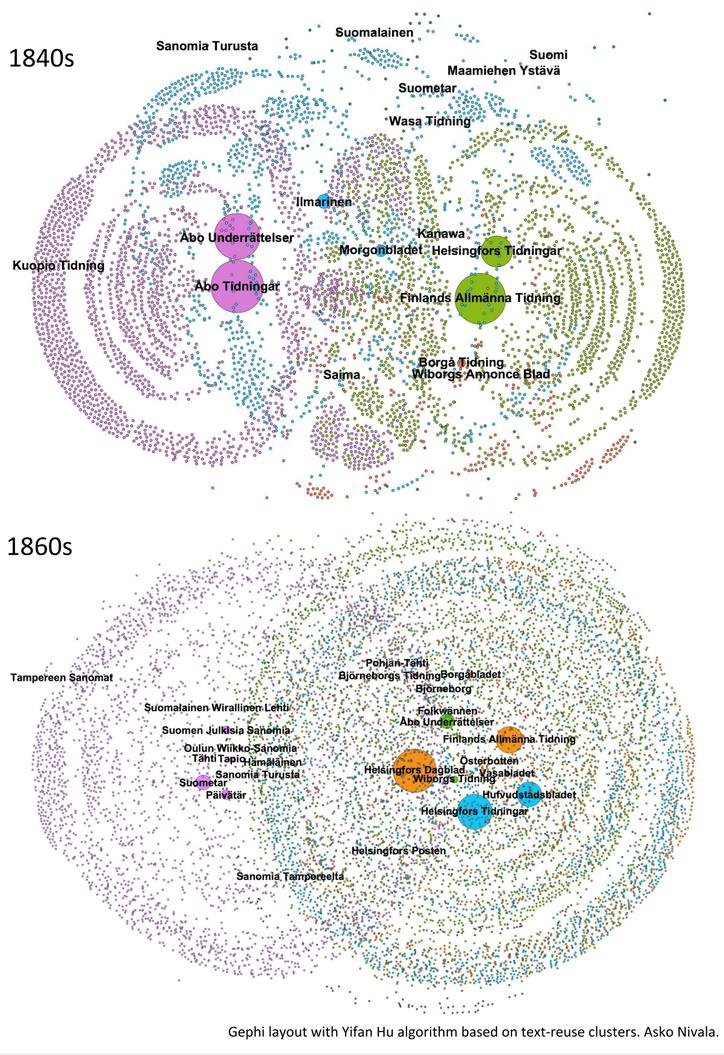
- Discussions in the public were inherently bilingual, but the technological and journalistic developments advanced at different speeds in Swedish and Finnish.
- The increasing number of text reuse clusters indicates a growth in capacity of the papers as well as an increased virality in public discourse.
- Text reuse cases show that Swedish-language papers formed a network that reused paragraphs from the 1840s onwards. Journalistic practises developed slightly later in Finnish-language papers, but catched up especially after the language decree of 1855. In the late 1800s and early 1900s Finnish-language papers multiplied, especially after the relaxing of censorship in 1905.

Text reuse in newspapers, 1820-1910



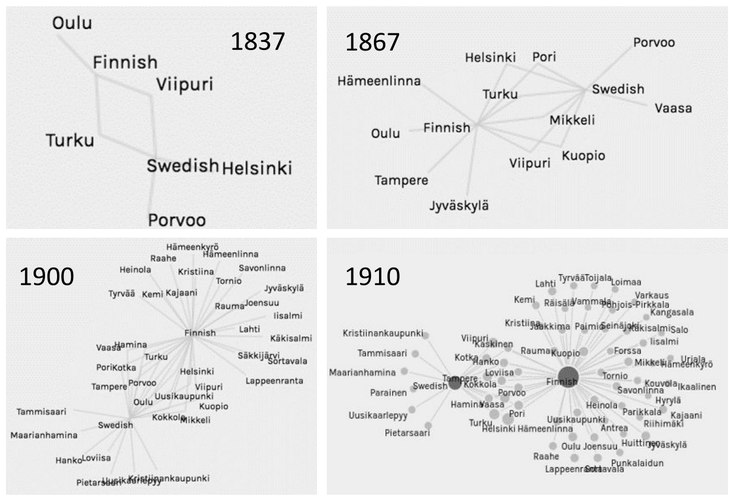
Based on text-reuse clusters. Hannu Salmi.

Network of Finnish press, 1846–50 and 1865



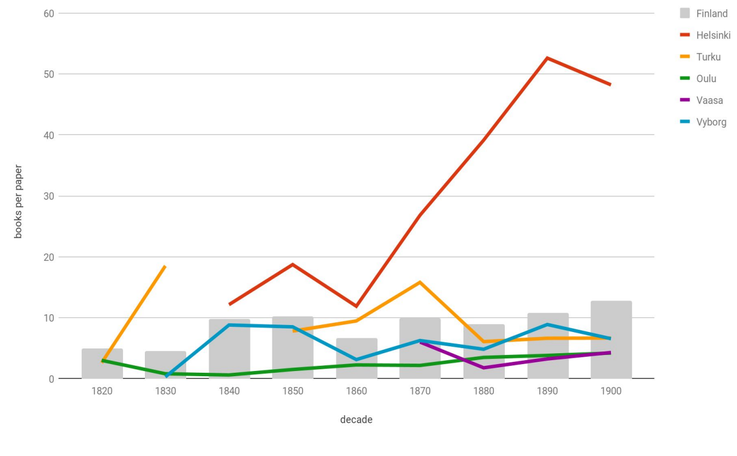
Gephi layout with Yifan Hu algorithm based on text-reuse clusters. Asko Nivala.

Published newspapers in city by language



Palladio network graphs based on newspaper metadata. Mikko Tolonen.

Finnish-language books per newspapers



Based on Fennica and newspaper metadata. Ville Vaara.

Newspapers were crucial to the birth of the nation as an imagined community. Yet, the national public sphere was not without regional asymmetries. The study of text reuse traces "senders" and "receivers" in discourse.

- 1840s: text reuse correlates with location. Helsinki and Turku papers reused local competitors' material.
- 1860s: Finnish-language press was more established and text-reuse clusters group by language.

An analysis of language according to location shows that:

- Newspapers were being established all over the country and becoming forums for local debates.
- Since the emergence of Finnish-language press in 1820, most bigger cities gravitated towards having newspapers in both languages.
- Toward the end of the century smaller mono-lingual municipalities also gained newspapers.

The study further assesses the development of the press in comparison with book production and periodicals which shows that:

- A specialization of newspapers as a medium took place in the period post 1860.
- Book production and publication emerged from university milieus in Helsinki, whereas newspaper production and readership comprised of regional audiences and contributors.

Further information

All code is available at the project's repositories (https://comhis.github.io and https://github.com/avjves/textreuse-blast). The results of the text reuse detection are stored in a database that is available at http://comhis.fi/clusters. It includes an index which combines the rapidity, the amount of newspapers and the geographical expansion in the spread of viral texts.

DIGI hum
1640 KANSALLIS KIRJASTO
HELSINGIN YLIOPISTO HELSINGFORS UNIVERSITET UNIVERSITY OF HELSINKI
University of Turku