Disappearing Discourses: Avoiding anachronisms and teleology with data-driven methods in studying digital newspaper collections

Elaine Zosa, Simon Hengchen, Jani Marjanen, Lidia Pivovarova, Mikko Tolonen
University of Helsinki — firstname.lastname@helsinki.fi

Newspapers have been a rich source of information for historians for the past hundred years or so. In the past twenty years, digitization of newspapers has made it possible to do simple tasks such as keyword searches or more elaborate text mining analyses. Advancements like this create unprecedented possibilities to the analysis of historical sources. While there is some truth to the promises of the future, the reality is such that the research on digitized newspapers remains underdeveloped with regard to reference corpora and reproducibility of the research. Digitized newspapers are particularly discussed with respect to the development of public discourse, but the idea of entering the realm of past discourse *in toto* through the digitized newspapers may in the end be harmful. In reality, historians are interested in the different layers of newspaper publicity, thus location and temporality always play a crucial role of any historical analysis of public discourse in newspapers.

With these aspects in mind, this paper takes advantage of digitized newspapers and data-driven approaches in identifying disappearing discourses in newspapers. In doing this, we want to revisit one of the key tensions in historiography, that is, the interplay between being relevant for the present and at the same time writing history in a way that is true to the experiences of past actors. History's presentism is sometimes discussed critically from the perspective of anachronism or teleology in history (Koselleck 2010; Skinner 2002), or more appraisingly in terms of genealogies of the present or letting all be the history of the contemporary (Armitage forthcoming). Regardless of the historian's desire for contemporary relevance or for historical antiquarianism, the option to approach history without predefined questions from the present has not been possible. The advent of digitized sources that can be approached in a data-driven way opens up for a possibility of approaching history in a much more open-ended way. Hence, we propose to test the possibility of studying a historical case with as few presupposed categories as possible. To do this we study digitized newspaper collections (specifically, 19th century Finnish newspapers in Finnish and Swedish) through the perspective of discourses that fell out of fashion and disappeared from long-term diachronic newspaper data sets.

We believe there is more potential in the use of digitized newspapers when we are not pinpointing the words and concepts in our approach *a priori*. This may lead us to completely new avenues of research, challenge our take on history as a some sort of progression and, hopefully, show the value of the data-driven approach for the humanities. To understand the boundaries and the development of the public sphere it is useful to identify those discourses

that were important in a particular time and place, but have since disappeared while words and concepts of another discourse have replaced them and started to dominate the ecosystem of print publicity. It is a commonplace to note that religious discourse has lost much of its prominence or that technological advancements have brought with them new topics that have replaced old ones. Still, by turning the question around and asking which discourses disappeared, we get a broader picture. We then turn to the data again and zoom in on localities and languages in order to avoid a totalizing view and move on to looking at where and when discourse changed. Thus, while we produce an analysis of public discourse in Finland, we approach the topic by noting that this is not a unified whole, but composed of different entangled realms of public discourse (Tolonen et al 2019; Marjanen et al 2019a).

Using newspapers and periodicals data in Finnish and Swedish encompassing respectively 5.2B and 3.4B tokens (National Library of Finland 2011a, 2011b), we utilise two different methods: relative word frequencies as proxies for particular discourses enhanced with distributional semantics derived from diachronic word embeddings (Kim et al 2014, Dubossarsky et al 2019), and dynamic topic modeling that captures more general themes.

The former method, i.e. the combination of frequency analysis and vector space similarity allows us to focus on specific themes and track their dynamics along a timeline to detect crucial events related to those themes. This has successfully been carried out by recent previous work on similar data (Martinez-Ortiz et al 2016; Hengchen et al 2019; Marjanen et al 2019b; van Eijnatten and Ros 2019). Training diachronic word embeddings on different time granularities (e.g. months, years, or decades) allows for different views on the evolution of semantic clusters – these themes are then given weight through frequency counts.

The latter method allows us to paint a larger picture of the different dynamics taking place in the data, by harnessing the power of topic models designed to capture trends in time-series data such as Dynamic Topic Models (DTM, Blei and Lafferty 2006). In DTM, the data is divided into discrete time slices and the method infers topics across these time slices to capture topics evolving over time. This method models how a topic changes from one time step to the next. Unlike vanilla LDA topic modelling which does not take into account the evolution of a topic, DTM is more robust to topics that changes vocabulary over time to talk about the same issue. In LDA, topics like these would likely to be separated into separate topics since the words associated with them has changed but in DTM they would be treated as one topic that is developing over time. To address the additional training complexity of this model we subsample the data such that we have the same amount of data for each time slice of our corpus. This would also ensure that the topics inferred are representative of all the time slices in the corpora rather than favoring the latter years which have more articles and newspapers associated with them.

With thematically-labelled temporal representations of newspaper data, it becomes possible to quantify and qualify the evolution of certain themes that have been automatically inferred

from the data — thus removing some bias in topic selection. We further use metadata to zoom in on changes in topics to see which towns, regions or types of newspapers to manually assess the driving locations of change and to produce a typology of disappearing discourses.

Acknowledgements
This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

References

1. Armitage, D. (In Press). In Defense of Presentism. In D. M. McMahon (Ed.), History and Human Flourishing. Oxford: Oxford University Press.
2. Blei, D.M. and Lafferty, J.D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning. ACM, pages 113–120
3. Dubossarsky, H., Hengchen, S., Tahmasebi, N. and Schlechtweg, D. (2019). Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
4. van Eijnatten, J. and Ros, R. (2019). The Eurocentric Fallacy. A Digital Approach to the Rise of Modernity, Civilization and Europe. International Journal for History, Culture and Modernity, 7.
5. Hengchen, S., Ros, R., and Marjanen, J. (2019). A data-driven approach to the changing vocabulary of the 'nation' in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In Proceedings of the Digital Humanities (DH) conference 2019, Utrecht, The Netherlands
6. Kim, Y., Chiu, Y.I., Hanaki, K., Hegde, D. and Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. ACL 2014, p.61.
7. Koselleck, R. (2010). Vom Sinn und Unsinn der Geschichte: Aufsätze und Vorträge aus vier Jahrzehnten von Reinhart Koselleck - Suhrkamp Insel Bücher Buchdetail (C. Dutt, Ed.). Berlin: Suhrkamp.
8. Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., & Tolonen, M. (2019a). 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. https://doi.org/10.21825/jeps.v4i1.10483
9. Marjanen, J., Pivovarova, L., Zosa, E. & Kurunmäki, J. (2019b). Clustering Ideological Terms in Historical Newspaper Data with Diachronic Word Embeddings. in Proceedings of the 5th International Workshop on Computational History. HistoInformatics2019 - the 5th International Workshop on Computational History, 12/09/2019.
10. Martinez-Ortiz, C., Kenter, T., Wevers, M., Huijnen, P., Verheul, J. and Van Eijnatten, J. (2016). Design and implementation of ShiCo: Visualising shifting concepts over time. In HistoInformatics 2016 (Vol. 1632, pp. 11-19).
11. National Library of Finland (2011a). The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016050302.
12. National Library of Finland (2011b). The Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016050301.

13. Skinner, Q. (2002). Visions of politics. Vol. 1, Regarding method. Cambridge University Press.
14. Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2019). A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *52*(1), 57–78. https://doi.org/10.1080/01615440.2018.1526657