Erik Edoff, Dept. of Culture and Media Studies, Umeå University, Sweden

A newspaper atlas: Named entity recognition and geographic horizons of 19th century Swedish newspapers

What was the outside world for 19th century newspaper readers? That is the overarching problem investigated in this paper. One way of facing this issue is to investigate what geographical places that was mentioned in the newspaper, and how frequently. For sure, newspapers were not the only medium that contributed to 19th century readers' notion of the outside world. Public meetings, novels, sermons, edicts, travelers, photography, and chapbooks are other forms of media that people encountered with a growing regularity during the century; however, newspapers often covered the sermons, printed lists of travelers and attracted readers with serial novels. This means, at least to some extent, that these are covered in the newspapers columns. And after all, the newspapers were easier to collect and archive than a public meeting, and thus makes it an accessible source for the historian.

Two newspapers, digitized by the National Library of Sweden, are analyzed: *Tidning för Wenersborgs stad och län* (TW) and *Aftonbladet* (AB). They are chosen based on their publishing places' different geographical and demographical conditions as well as the papers' size and circulation. TW was founded in 1848 in the town of Vänersborg, located on the western shore of lake Vänern, which was connected with the west coast port, Göteborg, by the Trollhätte channel, established in 1800. The newspaper was published in about 500 copies once a week (twice a week from 1858) and addressed a local and regional readership. AB was a daily paper founded in Stockholm in 1830 and was soon to become the leading liberal paper of the Swedish capital, with a great impact on national political discourse. For its time, it was widely circulated (between 5,000 and 10,000 copies) in both Stockholm and the country as a whole. Stockholm was an important seaport on the eastern coast. These geographic distinctions probably mean interesting differences in the papers' respective outlook. The steamboats revolutionized travelling during the first half of the century, but its glory days had passed around 1870, and was replaced by railways as the most prominent way of transporting people.

This paper is focusing on comparing the geographies of the two newspapers by analyzing the places mentioned in the periods 1848–1859 and 1890–1898. The main railroads of Sweden were constructed during the 1860s, and the selected years therefore cover newspaper geographies before and after railroads.

The main questions of paper addresses relate to media history and history of media infrastructure. During the second half of the 19th century several infrastructure technologies were introduced and developed (electric telegraph, postal system, newsletter corporations, railways, telephony, among others). The hypothesis is that these technologies had an impact on the newspapers' geographies. The media technologies enabled information to travel great distances in short timespans, which could have homogenizing effects on newspaper content, which is suggested by a lot of traditional research (Terdiman 1999). On the other hand, digital historical research has shown that the development of railroads changed the geography of Houston newspapers, increasing the importance of the near region rather than concentrating geographic information to national centers (Blevins 2014).

The goal of the study is in other words to investigate what these the infrastructural novelties introduced during the course of the 19th century as well as the different geographic and demographic conditions meant for the view of the outside world or the imagined geographies provided by newspapers. The primary goal with this paper is to investigate a historical-geographical problem relating to newspaper coverage and infrastructural change. The secondary aim is to tryout the use of Named Entity Recognition on Swedish historical newspaper data.

Named Entity Recognition (NER) is a software that is designed to locate and tag entities, such as persons, locations, and organizations. This paper uses SweNER to mine the data for locations mentioned in the text (Kokkinakis et al. 2014). Earlier research has emphasized the problems with bad OCR-scanning of historical newspapers. A picture of a newspaper page is read by an OCR-reading software and converted into a text file. The result contains a lot of misinterpretations and therefore considerable amount of noise (Jarlbrink & Snickars 2017). This is a big obstacle when working with digital tools on historical newspapers. Some earlier research has used and evaluated the performance of different NER-tools on digitized historical newspapers, also underlining the OCR-errors as the main problem with using NER on such data (Kettunen et al. 2017). SweNER has also been evaluated in tagging named entities in historical Swedish novels, where the OCR problems are negligible (Borin et al 2007). This paper, however, does not evaluate the software's result in a systematic way, even though some important biases have been identified by going through the tagging of some newspaper copies manually. Some important geographic entities are not tagged by SweNER at all (e.g. Paris, Wien [Vienna], Borås and Norge [Norway]). SweNER is able to pick up some OCR-reading mistakes, although many recurring ones (e.g. Lübeck read as Liibeck, Liibcck, Ltjbeck, Ltlbeck) are not tagged by SweNER. These problems can be handled, at least to some degree, by using "leftovers" from the data (wrongly spelled words) that was not matched in a comparison corpus. I have manually scanned the 50,000 most frequently mentioned words that was not matched in the comparative corpus, looking for wrongly spelled names of places. I ended up with a list of around 1,000 places and some 2,000 spelling variations (e.g. over 100 ways of spelling Stockholm). This manually constructed list could be used as a gazetteer, complementing the NER-result, giving a more accurate result of the 19th century newspaper geographies.

There is a big difference in size between the corpuses of two newspapers. As mentioned above, *AB* was published daily, which makes that corpus several times larger than the *TW* one. For the sample years I have analyzed (1850 and 1890), the corpus for *TW* is about 0,4 million words for 1850 and about 1,7 million for 1890. The preliminary result based on a sample from *TW* is shown in the table below. Regarding the precision of the NER location detection, some hints are given in the relation between manually tagged locations and locations tagged by NER, even though no formal precision and recall analysis has been performed yet.

Year	1850	1890
Corpus size (circa)	400,000 words	1,700,000 words
Locations tagged by NER	1,963	19,914
Locations tagged manually	1,359	8,129
Locations (total)	3,322	28,043
Unique locations	305	1,206
Locations/100,000 words	830	1,171

REFERENCES

Blevins, C. (2014), ”Space, nation, and the triumph of region: A view on the world from Houston”, *Journal of American History*, Vol. 101, no 1, pp. 122–147.

Borin, L., Kokkinakis, D., and Olsson, L-G. (2007), “Naming the past: Named entity and animacy recognition in 19th century Swedish literature”, *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pp. 1–8, available at: http://spraakdata.gu.se/svelb/pblctns/W07-0901.pdf (accessed October 31 2017).

Jarlbrink, J. and Snickars, P. (2017), “Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive”, *Journal of Documentation*, Vol. 73, no 6, pp. 1228–1243.

Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., and Löfberg, L. (2017), ”Old content and modern tools: Searching named entities in a Finnish OCRed historical newspaper collection 1771–1910”, *Digital Humanities Quarterly*, (preview) Vol. 11, no 3.

Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., and Borin, L., (2014), ”HFST-SweNER – A new NER resource for Swedish”, *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik 26–31 May 2014., pp. 2537-2543

Terdiman, R. (1999) “Afterword: Reading the news”, *Making the news: Modernity & the mass press in nineteenth-century France*, Dean de la Motte & Jeannene M. Przyblyski (eds.), Amherst: University of Massachusetts Press.