The omnipresence of the nation

Jani Marjanen[†], Ruben Ros[§], Simon Hengchen[† 1], Mikko Tolonen[†]

[†] University of Helsinki {firstname.lastname@helsinki.fi}
[§] Utrecht University {r.s.ros@students.uu.nl}

The nineteenth century saw an increased publishing of newspapers in Europe, and Sweden and Finland are no exceptions to that rule. At the same time the nineteenth century is considered the age of nation building. Many influential scholars couple the expansion of print media with the emergence of the nation as a collective imaginary (see particularly Anderson 2006). Through their inherent characteristics, newspapers as a corpus help us study the evolution of a topic (and of the words used to describe said topic) at the macro level as well as at the micro level. We aim to study aspects of nation building through the proliferation of vocabulary with the root "nation". We suggest that terminology relating to nation evolved from the late eighteenth century to the early twentieth century by becoming associated with new domains. Slowly, "nation" and "national" became terms that could be attached to almost anything discussed in public discourse.

To prove our hypothesis, and to further illustrate that this tendency is seen in several countries -- and languages -- within Europe, in other words that the motivation for this semantic change, to use Blank (1997)'s terminology, is a widespread *sociocultural change*, we study newspapers in Swedish (Sweden and Finland), Dutch (the Netherlands), and English (UK). We have used the materials provided the Finnish and Swedish language banks, the Royal Library in the Netherlands and Gale Engage. We query for bigrams[2] of "national" in different languages, and study the productivity and creativity of those bigrams. We expect to capture institutions of cultural nationalism ("national language", "national anthem", etc.), but also rhetorically potent claims like arguments for "national unity". Finally, we also expect to pick up themes that were seen as important, at the time, to be framed as national: "national borders", etc.

The second step of our analysis is to cluster them thematically. Looking at relative frequencies of themes across time allows us to frame the rise and fall of certains themes associated with the adjective "national" and, within those themes, to determine which words were used to describe those themes. Charting this development will paint a picture of the expansion of the vocabulary relating to "national", but also help showcase different patterns in different countries.

We are currently thinking how to best implement more "computationally sophisticated" ways to tackle our research question. One of the ways to track the evolution of a word's sense is to train word embeddings and align them diachronically, as has been done on Swedish

---

[1] Corresponding author
[2] In the case of bigrams containing "meaningless" words such as conjunctions (eg: *nationell och*, *nationaal en*, "national and"), we expand the query until we arrive at the noun modified by the adjective.

newspapers (Tahmasebi 2018), although word embeddings usually only capture the primary sense of a word. Another area of research we are considering is building a dynamic topic model of a word's sense distribution across time, as put forward by Frermann and Lapata (2016). We are at a point in our analysis where we would like feedback on possible methods, as well as on available corpora we have not thought of.

Acknowledgments:

References:
Anderson, B. (2006 [1983]), *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, Verso, London.

Blank, A. (1999). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. *Historical semantics and cognition*, (13):6.

Frermann, L., & Lapata, M. (2016). A Bayesian model of diachronic meaning change. Transactions of the Association for Computational Linguistics, 4, 31-45.

Tahmasebi, N. (2018). A Study on Word2Vec on a Historical Swedish Newspaper Corpus. In *DHN* (pp. 25-37).