The omnipresence of the nation

Marjanen, Jani

2019-03

Marjanen , J , Ros , R , Hengchen , S & Tolonen , M 2019 , ' The omnipresence of the nation ' , Digital Humanities in the Nordic Countries , Copenhagen , Denmark , 05/03/2019 - 08/03/2019 .

http://hdl.handle.net/10138/314884

The omnipresence of the nation

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

[†] University of Helsinki {firstname.lastname@helsinki.fi}

[§] Utrecht University {r.s.ros@students.uu.nl}

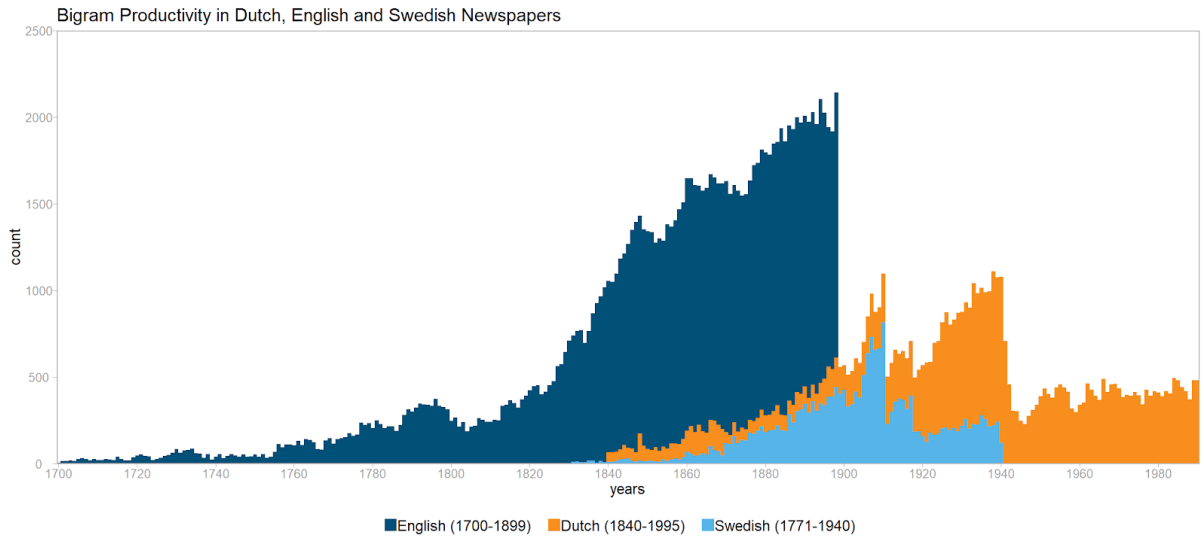
Abstract for Digital Humanities in the Nordic Countries 2019 (DHN2019) conference in Copenhagen, Denmark, 6–8 March 2019

The nineteenth century saw an increased publishing of newspapers in Europe. 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). We suggest turning to newspapers as a source for historical descriptions of the evolution of how historical persons conceptualized the nation. 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". In doing so, we do not claim that the language of nationhood corresponds directly to how people saw the nation, but rather mapping the terminology provides a glimpse of where and for which rhetorical purposes something could be conceptualized as national. The change in the vocabulary around the nation thus reflects historical changes, but were also a way of shaping the national imaginary (cf. Koselleck 1972). We suggest that terminology relating to nation evolved from the late eighteenth century to the early twentieth century by becoming associated with new domains. In general terms, we hypothesised that the terms "nation" and "national" expanded from economy and politics to culture in the nineteenth century and finally the terms could be attached to almost anything discussed in public discourse.

To explore 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 (from 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. The Swedish material consists of Swedish-language newspapers from Finland (1771–1900). The newspapers from Finland cover all newspapers published in the country at this time and are thus a small but representative sample of public discourse. The Dutch data spans almost three centuries (1680-1995) and consists of a relatively representative sample of the Dutch newspaper landscape (in terms of location, political affiliation and circulation). It must be noted that the n-gram data provided by the Dutch Royal Library starts only in 1840. The English data consists of two combined datasets; the Burney Nichols Collection of eighteenth century newspapers, and the British Library Nineteenth Century Newspaper Collection. The latter contains 48 newspapers that span the century between 1800 and 1900. Of the 48

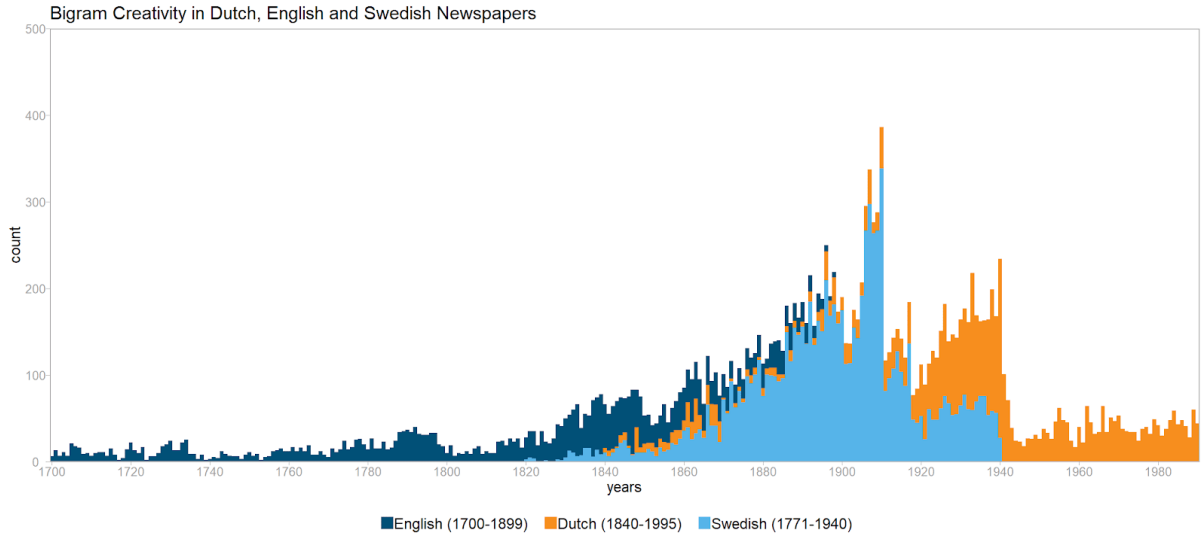
newspapers, 17 are national and 29 are regional. The Burney Nichols Collection consists of 136 volumes of 17th-century newspapers and 1,145 volumes of eighteenth-century newspapers. These newspapers are predominantly London-based, and regional newspapers are scarce.

A first step is to query for bigrams[1] of "national" in different languages, and study the productivity and creativity[2] of those bigrams. We expected to capture institutions of cultural nationalism ("national language", "national anthem", etc.), but also rhetorically potent claims like arguments for "national unity", but studying the totality of bigrams relating to national paints a broader picture. The visualization below shows the bigram 'productivity' and 'creativity'. Productivity maps the growth in vocabularies, whereas creativity is better to show the critical years in which 'national' was particularly productive. Together these measures can be used to investigate into the *extension* of the vocabulary of national, and to visualize the diachronic quantitative change of vocabularies used in connection with "the nation". We identified the French Revolution, the 1848 political ruptures and the Franco-German War (1870) as critical junctures in cross-national vocabulary change. Still, the developments were, as expected, also uneven due to local political circumstances. For instance, in the Swedish-language data (from Finland), the years after 1905 appear as a period of productivity and creativity. This is partly a result of the introduction of freedom of print and universal suffrage in Finland. In the Netherlands, the politically tumultuous 1930s involved a high (and productive) use of 'national' that is contained in the proliferation of vocabulary relating to national socialism (van Sas 1999: 408-411).



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

[2] The definitions of "productivity" and "creativity" are fluid within subfields of linguistics, as already discussed in Lyons (1977: 77). In this paper, we use "productivity" in its corpus linguistics sense, i.e. the proclivity of a linguistic unit to be (re-)used. "Creativity", on the other hand, will characterise this unit's *new* forms: in the case of a bigram, any new bigram with "national + _".



The second step of our analysis is to cluster the bigrams 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 allows to paint a picture of the expansion of the vocabulary relating to "national", but also helps showcase different patterns in different countries.

We are currently using 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 on separate time slices, and to align them diachronically, as has been done on Swedish newspapers (Tahmasebi 2018).[3] Nonetheless, this method has several weaknesses for our purpose: word embeddings usually only capture the primary sense of a word (Iacobacci *et al* 2015), and whilst sense-specific embeddings exist, recent research shows that it is unknown whether this advance helps or hinders the task of word similarity (Dubossarsky *et al* 2018). Additionally, there is no fixed rule for the definition of time periods, which are usually dictated by computational reasons[4] and thus often disregard expert knowledge of the period studied, although we are currently implementing part of the methodology put forward by Hengchen (2017), which pleads for the use of knowledge-created time slices for data-driven methods. Despite the obvious weaknesses of the methods described above, we rely on word embeddings trained on our data and "similarity" distances between top words to semantically cluster top bigrams and, through this process, attempt to achieve a fully-automated and data-driven thematic labelling of bigrans across time.

[3] Embeddings were trained on tokens with a frequency threshold of 300, a CBOW architecture, 100 dimensions, and with a window of 5. Furthermore, mimicking Kim *et al* (2014), the embeddings were trained on different time slices, where embeddings for slice `t+1` are initialised with the embeddings for slice `t`, hence bypassing the need for a "temporal" alignment of the vector space.

[4] Or no reason at all.

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). This method[5] allows, for a specific target word, for the creation of different semantic fields for different time periods, and this for each inferred sense of the target. The vanilla version of the model does not currently incorporate genre information into its inference, a limit than an in-house version [6] of the algorithm corrects.

Finally, we are looking at more statistically robust ways of looking at frequencies of bigrams. To achieve that goal, we rely on a log-likelihood of bigrams, as put forward in the case of verb-object pairs by Cavallin (2012). The higher the ranking, the stronger the connection between the two parts of the bigram. Looking at that ranking across time is a robust data-driven way at looking at the association between two words.

The comparative approach in our study is both revealing and vague in its outcomes. Clustering bigrams by hand and comparing the results suggests that there is some merit to our original hypothesis about the expansion of the the vocabulary relating to "national". In particular it shows the growth of political topics in the vocabulary of national in the nineteenth century. There are some crucial differences in our cases however that seem to be more about the configuration of the data than about the shift in socio-cultural language. The sudden productivity found in Swedish-language newspapers from Finland after 1905 and the rise of national-socialist vocabulary in the Netherlands during the WWII are naturally part of public discourse, but are rather measures of political change than semantic change. At the same time, political and semantic change are in the end inseparable.

Acknowledgments:

This work is partly funded by NewsEye under the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299.

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.

Cavallin, K. (2012). Automatic extraction of potential examples of semantic change using lexical sets. In *KONVENS* (pp. 370-377).

Dubossarsky, H., Grossman, E., & Weinshall, D. (2018) Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research. To be published in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.*

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

Hengchen, S. (2017). When does it mean? Detecting semantic change in historical texts. Ph.D. thesis, Université libre de Bruxelles.

[5] It is not the aim of this abstract to go into details about SCAN. We believe Figures 4 and 5 of the original paper illustrate clearly the strength of this model.

[6] Currently under review.

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Vol. 1, pp. 95-105)

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

Koselleck, R. (1972). "Einleitung", Otto Brunner, Werner Conze & Reinhart Koselleck (eds.), *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*. Band I (Stuttgart: Klett-Cotta), pp. XIII–XXVII.

Lyons, J. (1977). Semantics (vols i & ii). *Cambridge CUP*.

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

van Sas, N.C.F. (eds.) (1999). *Vaderland; een geschiedenis van de vijftiende eeuw tot 1940* (Amsterdam: Amsterdam University Press).