A data-driven approach to studying changing vocabularies in historical newspaper collections

Simon Hengchen
Språkbanken Text, University of Gothenburg, Sweden and iguanodon.ai, Belgium

Ruben Ros
Centre for Contemporary and Digital History (C2DH), University of Luxembourg, Luxembourg

Jani Marjanen and Mikko Tolonen
Helsinki Computational History Group, University of Helsinki, Finland

Abstract

Nation and nationhood are among the most frequently studied concepts in the field of intellectual history. At the same time, the word ‘nation’ and its historical usage are very vague. The aim in this article was to develop a data-driven method using dependency parsing and neural word embeddings to clarify some of the vagueness in the evolution of this concept. To this end, we propose the following two-step method. First, using linguistic processing, we create a large set of words pertaining to the topic of nation. Second, we train diachronic word embeddings and use them to quantify the strength of the semantic similarity between these words and thereby create meaningful clusters, which are then aligned diachronically. To illustrate the robustness of the study across languages, time spans, as well as large datasets, we apply it to the entirety of five historical newspaper archives in Dutch, Swedish, Finnish, and English. To our knowledge, thus far there have been no large-scale comparative studies of this kind that purport to grasp long-term developments in as many as four different languages in a data-driven way. A particular strength of the method we describe in this article is that, by design, it is not limited to the study of nationhood, but rather expands beyond it to other research questions and is reusable in different contexts.

Correspondence: Simon Hengchen, University of Gothenburg, Gothenburg, Sweden.
E-mail:
simon.hengchen@gu.se

1 Introduction

There has been extensive research on the process in which nation-states become pivotal units for international politics and crucial categories of belonging for individuals (Özkırımlı, 2000; Anderson, 2006; Smith, 2013, 2008). Our aim in this article was to use state-of-the-art word embeddings to describe how this process is reflected in the written use of four different languages: Dutch, English, Swedish, and Finnish. Earlier work has focused on words and concepts relating to nationhood, pointing out the general trajectories of word use and the increased levels of vagueness of ‘nation’ as a concept (Kemiläinen, 1964;

doi:10.1093/llc/fqab032

Gschnitzer *et al.*, 1978). Word embeddings are commonly used in natural language processing (NLP), but their application in historical research is still at the experimental stage. There have been many robust attempts at evaluation in state-of-the-art research using embeddings, but in the case of historical data, the evaluation has to be against historical research. It therefore remains difficult to determine what the real object of the modelling is and if the results are transferable to other languages, different corpora, or time spans.

Our large-scale comparative perspective demonstrates changes in the development of nationhood with greater clarity than before in focusing on the term 'national' and using words associated with it to analyse domains that were increasingly being conceptualized as national. One of the benefits of our case is that the words nation and national exist in all four languages as cognates or, in the case of Finnish, as neologisms in the 19th century. The historical translatability makes it ideal for comparative study in that it highlights both similarities as well as differences between the languages in question.

Many studies on semantic change of particular words or concepts over long periods of time tend to focus on changes in words that shift between two distinct senses (Recchia *et al.*, 2017; Tahmasebi *et al.*, 2018). The case of national that we chose for this study is different in that it relates to historical processes that are of interest to historians in particular, but it also provides a challenging case for the use of computational methods as it is not about detecting polysemy, but rather about grasping a vague term and its increasing importance in political discourse over time (on vagueness, see Geeraerts, 1993). This also holds for most of the key terms of interest in understanding political, social, and cultural transformation in the modern period. Words such as state, nation, ideology, culture, gender, and racism have been extensively researched as pivotal terms that have been contested in past debates and whose changing meanings have been indicative of historical transformation, but have also been the cause of change in the past (see e.g. Koselleck, 1972–1997; Ball *et al.* 1989). Although many of these words are polysemous, the aspect that makes them interesting politically and culturally is that they are also vague, at least in one of their senses, that they are used in rather different language domains, and that historical actors seem to have cared a lot about which uses were correct. The vagueness of key terms for navigating society is inherently tied to the complexity of the data required to detect shifts in language use. Historical data have developed in conjunction with societal processes and events (everything from growing wealth to war and censorship practices to changing fashions), and therefore form non-standard data units in terms of computation (Mäkelä *et al.*, 2020). More importantly, developments in the data, in our case, newspapers, are part of the process in which the terminological changes took place. This means that these newspapers cannot be used just to study changes, as the changes in them also need to be factored into the interpretation of the analyses.

The linguistic change relating to 'national' consists of a gradual growth in frequency and an expansion in language domains over time. Setting up a methodology that grasps this development over a long period of time, does so in different languages, is statistically robust and does so in a data-driven way, will pave the way for further historical study that could challenge and complement earlier qualitative accounts of nation-building. We point out that hypotheses developed in earlier studies based on limited source corpora have referred to nations and a shift in focus from the economy to culture and politics (Viroli, 1995; Ihalainen, 2007; Nurmiainen, 2009; Marjanen, 2013). We further propose that a data-driven clustering of the vocabulary relating to national allows for a more fine-grained image of the expansion of the national imaginary. We show signs of change in the language of nationhood that could perhaps be described as processes of culturalization, de-economization, and institutionalization, which should be evaluated more closely in historical research. What this means is that, over the course of the research period, terminology related to culture and political institutions became more commonly labelled as something national (as in national literature or national party), whereas economic terminology became proportionately less dominant in the discourse.

Our method is particularly suited to analysing complex historical keywords that are usually at the heart of studying the history of political and social thought. The development of methods and the concrete plots we devise in this study relate to the vocabulary revolving around the adjective national, but the aim here is not primarily to make a historical argument about nation building. We rather purport to identify ways of using computation to analyse historical trends in past conceptualizations of the world in a more nuanced way than key-word searches, relative frequencies, or topic models have made possible. Ultimately, the methods used to address historically informed questions need some level of tailoring to the data and the type of questions asked. However, given that the bulk of large-scale diachronic text data sets provide possibilities for the study of language in relation to historical processes, there are good possibilities of reuse in other research cases. This goes hand-in-hand with open science and the envisioning of research data as an ecosystem (Lahti *et al.*, 2019).

2 Related Work

2.1 Language and nationalism

Nationalism is a widely studied phenomenon and the role of semantic and lexical change has been noted in literature that provides overviews of the topic (see, in particular, Leersen, 2006, p. 15; Burke, 2013; Gilbert, 2018), but the bulk of the literature on nationalism has still been surprisingly indifferent towards the language of nationhood. This disinterest in the long-term changes in language relating to nationhood means that the analytical distinctions relating to nation-states and their emergence has been prioritized at the expense of enhancing understanding of the historical experience of nationhood.

Studies focusing on language tend to reflect an interest in the differences and similarities between patriotism and nationalism (Cunningham, 1989; Dietz, 1989; Viroli, 1995; Hont, 2005, pp. 447–528; Schierle, 2009). Another strongly related strand concerns the link between nation and fatherland in particular European languages (Kemiläinen, 1964; de Bertier de Sauvigny, 1970; Godechot, 1971; Gschnitzer *et al.*, 1978; Frautschi, 1993; Van Sas, 1999; Brenner, 2013). A few studies have also paid attention to nationalism as an ism in political discourse (Gschnitzer *et al.*, 1978; Bärenbrinker and Jakubowski, 1995; Freeden, 2009, pp. 204–224; Kettunen, 2018; Kurunmäki and Marjanen, 2018).

All of the above-mentioned studies, in one way or another, concern long-term trends in the meanings and uses of the words ‘nation’, ‘national’, and ‘nationalism’. However, apart from in a few isolated cases of resorting to relative frequencies, the use of quantitative methods to trace long-term developments in this vocabulary is almost completely non-existent. The one exception is Van den Bos and Giffard’s study, which focuses on key junctures in Dutch history and the language of nation (van den Bos and Giffard, 2016). The present study takes a step forward and a step backward. On the one hand, it engages in earlier claims about changes in word use being part of the process in which past expectations and experiences about nationhood were articulated (Koselleck, 1972, 2011), which on the other hand leads to claims that an over-arching study of the language of national could, in a general way, describe the process through which the national perspective became dominant in how people saw the world (Anderson, 2006). Although interpretations such as these already exist, they all rely on examples of particular texts rather than any kind of data-driven analysis, which means that they may be detailed in terms of individual examples, but they are not even close to capturing the whole story.

Methods for tracing this kind of change are not a perfect match for the historical questions posed in earlier research (Hengchen *et al.*, 2021). It is clear that earlier abstract claims about the shift in focus in the language of nationhood remain too broad to be captured in a meaningful way by methods for tracing semantic change in that they capture many different and partly conflicting signals from the data. Human interpretation has tended to filter them out, and quantitative methods for assessing shifting vocabulary necessarily have to find a good way of balancing detailed view and a result that is interpretable for human readers. There are good arguments for claiming that modelling may in some cases be less transparent and cannot capture the same things as qualitative interpretation (Biernacki, 2014), but in terms of understanding the evolving language of nationhood, the aspect of modelling and quantification has been completely missing.

2.2 Evolving vocabularies

The traditional focus in conceptual history has been on specific keywords such as ‘democracy’, ‘liberalism’, and ‘nation’, but only to a limited degree has there been any systematic analysis of semantic and lexical fields related to these keywords. The concentration on words has led to extensive discussions about their exact relationship with concepts (Steinmetz, 2012; Bolla *et al.*, 2019; Lähteenmäki and Kaukua, 2019; Bolla *et al.*, 2020). Although we do not assume that we can grasp the conceptual level behind words as such, we take a pragmatic approach and use distributional[1] methods to study changing vocabulary. These methods allow us to broaden the scope from words to groups of words (that are in some way related to concepts) through time. We are still intent on using words as proxies and thus remaining on the level of words and language use, because that will enable us to capture at least some of the personal experiences of historical actors. When they sought to express certain concepts, they chose particular words that reflected their own positions and thus left a trace of their experiences in the data. Moreover, focusing on words allows for the relatively easy tallying of their occurrences.

The challenge in analysing vocabulary is to ‘strike a balance between an adaptive strategy that responds to changes in vocabulary, and a more conservative approach that keeps the vocabulary stable’ (Kenter *et al.*, 2015). The vocabulary must maintain a minimal degree of stability in order for it to be historically relevant and meaningful, but at the same time, it should solve the problem of different words relating to the same concept over time (onomasiology).

Rather than considering a predefined group of words over time, distributional methods allow for a more data-driven approach. Recent scholarship in history has used word embeddings to identify semantically related words and to follow their development over time. This requires an initial set of seed terms that is subsequently expanded by selecting similar words (Kenter *et al.*, 2015; Recchia *et al.*, 2017). Another approach is to identify a vocabulary based on features of single words such as ‘isms’ (Pivovarova *et al.*, 2019, Marjanen *et al.*, 2020), or sequences of words (*n*-grams). The latter approach constructs a vocabulary based on words that are directly preceded (Wevers, 2017; Van Eijnatten and Ros, 2019) or modified (Hill *et al.*, 2018) by a common adjective, and subsequently focuses on the temporal changes. This leads to the quantification of conceptual extension, and gives insights into conceptual and distributional change that would go unnoticed were the focus only on specific keywords. Our method builds on such previous work, and in delegating the choice of ‘seed terms’ to nouns modified by a specific adjective allows for a more data-driven approach, while at the same time retaining some ‘topical control’ and harnessing semantic information from word embeddings.

2.3 Representing meaning in time

As noted above, previous attempts at studying an evolving discourse diachronically made use of computational methods and large corpora. More recent approaches lean on NLP. In this section, we discuss the state-of-the-art and illustrate why studying a specific theme over time is not trivial.

Topic modelling is extensively discussed and is sometimes used in the humanities (Fridlund and Brauer, 2013; Viola and Verheul, 2019). Although the soft clustering method is most commonly used synchronically for exploratory research, there are also dynamic topic models (DTMs) that take time as a variable and allow the extraction of topics across time slices. DTMs (Blei and Lafferty 2006) divide the data into discrete time slices and infer topics across them to capture topics evolving over time. A different approach, Topics over Time (Wang and McCallum 2006), treats time as a continuous variable and the data are not discretized. Although both approaches are promising, their major drawback is that the topic models do not allow for a topic to be defined *a priori*: they allow an exploratory look at the data, but there is no easy way to ensure that a certain topic will be found.

Another field in which meaning is studied computationally across time is that of lexical semantic change, which is particularly suited for conceptual change in that it focuses on words and not general themes (Kutuzov *et al.*, 2018; Tahmasebi *et al.*, 2018; Tang, 2018). To study meaning change, computational methods proceed in two steps: first, they distributionally model meaning in different time bins (subsequent temporal slices of the data at hand). Second, the focus is to detect, for any word *w*, whether the signal between time bins changes in a significant way. In recent years, even laws of semantic change have been proposed (Dubossarsky *et al.*, 2015; Hamilton *et al.*, 2016) and then disproved

(Dubossarsky *et al.*, 2017). Some methods have been under rigorous evaluation (Dubossarsky *et al.*, 2019; Schlechtweg *et al.*, 2019, 2020; Shoemark *et al.*, 2019). At the same time, new methods and paradigms aimed at diachronically modelling semantic information are being developed further: DTMs specifically targeting words (Frermann and Lapata, 2016; Perrone *et al.*, 2019) use bag-of-words to draw sense distributions for certain target words over time, dynamic, and continuous word embeddings (Bamler and Mandt, 2017; Rosenfeld and Erk, 2018; Rudolph and Blei, 2018; Yao et al., 2018; Dubossarsky *et al.*, 2019; Gillani and Levy, 2019) differ from static embeddings in that they use the entirety of the data (i.e. all time bins) to create vector representations, and more recently contextualized word embeddings (which have *token* vectors and not *type* vectors[2]) have been applied to diachronic corpora.

Thus, there have been robust attempts to evaluate and sometimes compare systems, but it remains difficult to determine what is actually being modelled, and whether the performances are transferable to other languages, different corpora, or dissimilar time spans. In short, it is arduous to determine whether NLP systems can be applied as-is to humanities data. Indeed, as McGillivray *et al.* (2019) remark, for example, despite being promising with regard to English, SCAN (Frermann and Lapata, 2016) performs poorly on an Ancient Greek corpus with sparse data and extended time bins, and the performance of an updated model (Perrone *et al.*, 2019) does benefit from additional information such as literary genre.

As anyone who works with historical material is aware, language changes over time. To avoid anachronisms, one has to make sure that texts are understood in their own context, rather than through a contemporary lens. Although historians have been trained to do this, as the above paragraph shows, current NLP methods might not be completely fit for the task. Additionally, NLP usually focuses on relatively straightforward cases,[3] and it is unclear whether or not the signal picked up by computational models is useful for humanities research, given that the changes in meaning being studied may well not be as obvious (Hengchen *et al.*, 2021). Finally, the computational processing of humanities data notoriously poses specific challenges both by its nature (evolving grammar and orthography, uneven size of data across time, for example)[4] and through how researchers can process it electronically (missing, incomplete, or wrong metadata, varying OCR quality, etc.) (Piotrowski, 2012).

3 Methodology

As we point out above, studying the changing vocabulary of a concept is no easy task. Doing so in a way that informs research in the humanities in a data-driven way adds a layer of complexity: if a certain, specific theme is to be studied it has to be defined *a priori*, and operational choices must be made[5] that might bias any quantitative method applied to the resulting subset of the data. Our methodological contribution is an approach that follows the fine line between having a precise research question and making use—in a data-driven way—of all the data available. It is a two-step approach, which we illustrate below in a case study on the changing vocabulary of nationhood in four countries and four languages.

To illustrate that the method is robust enough to tackle different data, languages, and periods, we carry it out on Dutch, Finnish, Swedish, and British newspaper data. The newspapers stem from different sources and countries and are available in different formats. Massive digitized newspaper collections are increasingly used to address historical questions through mining textual data.[6] The material, as well as the pre-processing steps, is laid out below, and the distribution of the data is available in Figs 1 and 2.

The Dutch data are the Delpher open newspaper archive (Royal Dutch Library, 2017) for the period 1618 until 1876 included. This archive is said to contain all newspapers for that period.[7] For the years 1877–1899 included, currently only available through the API, we queried the API for every item of the '*artikel*' type ('article', the Dutch data have article segmentation and further differentiates between advertisements and articles) category containing the determiner *de* ('the') at least once. Although this does not guarantee a 100% recall, *de* is so frequent that we are confident the extreme majority of articles of the necessary length for our tasks are retrieved. For anything pre-1877, we discarded pages that had anything other than exclusively 'nl' or 'NL' as language tags in the metadata. Articles from colonial newspapers were systematically removed. This is motivated

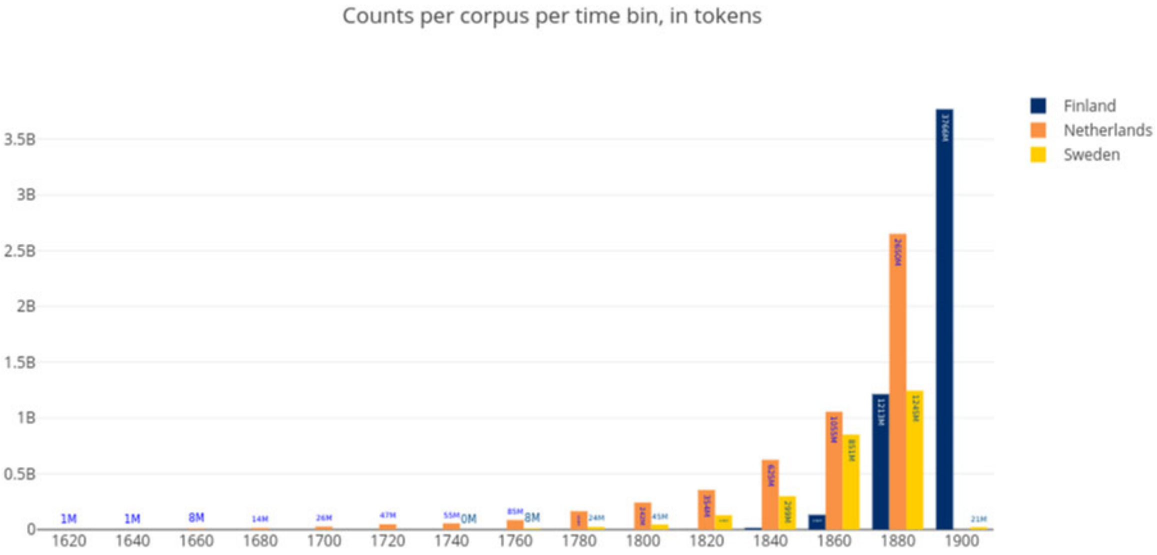

Fig. 1. Distribution of data size over time in token counts for FI, NL, and SV

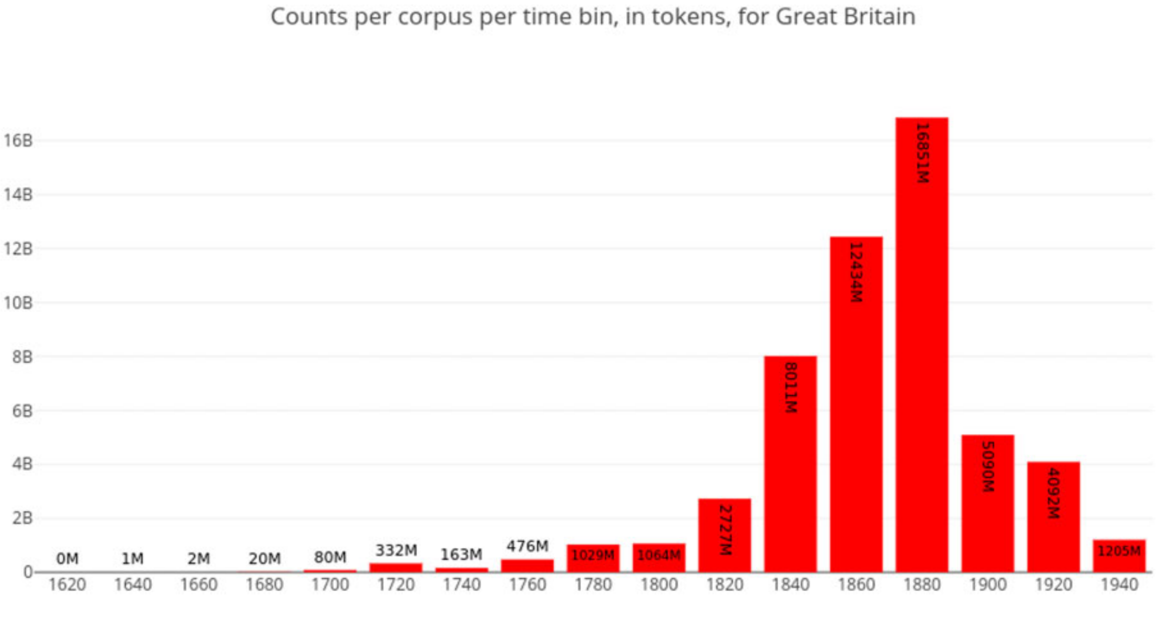

Fig. 2. Distribution of data size over time in token counts for GB

by the fact that only the Dutch dataset has an extensive coverage of colonial newspapers, and including them would have complicated our comparisons with the other countries in our study. The newspapers from Finland comprise two language corpora: we used the entirety of the Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland (National Library of Finland, 2011a) for articles in Finnish, and the corresponding Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland (National Library of Finland, 2011b).[8] Swedish newspapers are available in the Kubhist 2 corpus digitized by the Royal Library of Sweden, processed with the Sparv pipeline (Borin *et al.*, 2016), and made available online[9] by Språkbanken through Korp (Borin et al, 2012). Finally, the British data consist of the British Library Newspapers covering especially the 19th century,[10] the 17th and 18th Century Nichols collection,[11] and the 17th and 18th Century Burney collection.[12]

The changes in corpus size over time bins poses a problem for any computational text-mining task. Our approach creates intermediate data points in separate time bins of 20 years,[13] and it is only the aggregate information that is compared over time. As such, common pitfalls related to aspects such as limited vocabularies or the representativity of the data do not necessarily apply, as we spell out in our evaluation.

A further issue is that our data (historical newspapers) are not only data in which we study changing language: the change in corpus size and the growing importance of newspapers as a medium are parts of the historical process in which the language of nationhood has also changed. Growing amounts of newspapers created a different habitat in which the vocabulary of national could flourish; hence, there is no reasonable way of even trying to achieve a balance with the corpus used for the purpose of computation. Rather understanding changes in the corpus and the development of the public sphere in general is a form of corpus control, which is essential in terms of understanding the changing vocabulary of nationhood (Marjanen *et al.*, 2019; Tolonen *et al.*, 2019).

3.1 Extracting nationhood

First, using dependency parsing,[14] we utilized the method proposed by Hill *et al.* (2018) and extracted all the nouns modified by the adjective at hand, in our case 'national'.[15] With regard to the other languages we extracted nouns modified by *nationaal* and *nationale* in Dutch, *nationella*, *nationell*, and *national* in Swedish, and *kansallinen* in Finnish. Obviously, different languages have different properties. We resorted to splitting the Finnish and Swedish compound nouns starting with *kansallis-* and *national-*, respectively, while making sure they were genuine compounds, removed the 'national', and added the remaining part to our tally. As an example, Swedish *nationalbiblioteket* 'the national library' became *nationell + biblioteket*, but we discarded *nationaliteten* 'the nationality' as it is a noun in its own right,[16] and the Finnish *kansalliskirjasto* 'national library' became *kansallinen + kirjasto*. Only modified nouns are kept. For newspapers in Finnish and in Swedish from Finland, we used linguistic information made available by the language bank of Finland,[17] and similarly for newspapers in Swedish published in Sweden we used information made available by the language bank of Sweden[18]—both sets of data were produced by different versions of the same pipeline, Sparv (Borin *et al.*, 2016). Dutch and English datasets were dependency-parsed using spaCy 2 (Honnibal and Montani, 2017). The large models were chosen for both languages. Unfortunately, no assessment of the quality of the dependency parsing is available for Finnish and Swedish.[19]

The absolute counts of nouns modified by 'national' are displayed in Fig. 3. The relative frequencies show a similar pattern.

Because the meaning of 'national' changes over time, it is possible that other adjectives referred to what we now classify as national. To evaluate the 'centrality' of the adjective in the different time periods, therefore, we aggregated all other adjectives that modified the nouns modified by national. For example, in Dutch, this resulted in adjectives such as 'public', 'Dutch', 'royal', and 'foreign'. The frequencies of these 'competing' adjectives were lower in all decades, however, as well as in the overall time frame. This shows that the adjective 'national' was indeed the most commonly used to modify these nouns. The 'competing' adjectives sometimes perform a supplementary function but, as we will reveal, the discourse of national had a clear role of its own in all languages.

3.2 From words to concepts

Second, to allow the semantic clustering of all nouns relating to the concept of 'nation', we trained diachronic word embeddings on the entirety of the full text. Given that there was no conclusive way of determining what type of embedding was best for our data and that word embeddings are still poorly understood, and since we argue that dynamic and continuous word embeddings models cannot reliably be used here on account of the extremely uneven distribution of the data, we experimented with two fairly old architectures, CBOW and SGNS (Mikolov *et al.*, 2013a,b), which have been studied more thoroughly.[20] For the

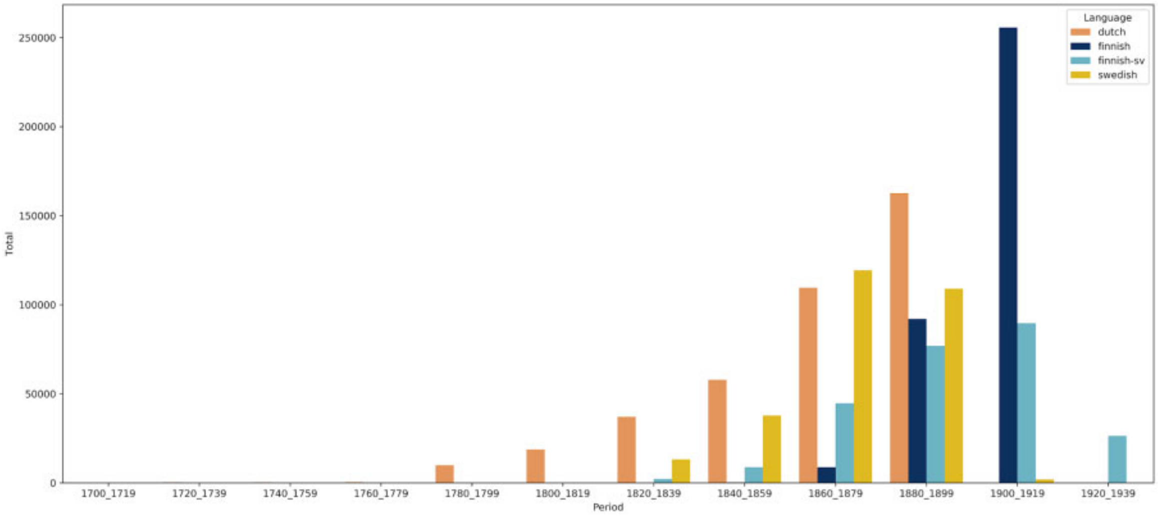

Fig. 3. Absolute counts of nouns modified by 'national' over time in Dutch, Finnish, Swedish from Finland, and Swedish from Sweden. Relative frequencies, not plotted, show a similar pattern

same reason, we created *diachronic* word embeddings using the two most frequently applied methods, post hoc alignment and incremental updating (described in detail below). We chose to train models on double decades for three reasons: first, 20 years roughly corresponds to a 'generation' in historical sociolinguistics (Säily, 2016); second, we needed a certain number of nouns related to the nation for the clustering to make sense, and bins of 20 years allow enough to be gathered, especially in the earlier periods; third, and somewhat echoing the second reason, it allowed us to have relatively stable models for the earlier periods.

For each time bin, we trained two types of word embeddings using gensim (Řehůřek and Sojka, 2010), a Python library for vector space modelling. Because separately trained vector spaces cannot be compared directly, we used two different methods to make the spaces comparable, and thus to ensure a sound diachronic approach. On the one hand, we followed Kim *et al.* (2014) and initialized the vector space for time bin t_1 with the space from t_0,[21] and updated the vectors by continuing the training. This differs slightly from the original approach in setting the learning-rate value of t_1 to that of the end of the previous model (in this case, t_0). The aim was to prevent the models from diverging too rapidly, as successfully reported in previous work based on the same data (Hengchen *et al.*, 2019; Pivovarova *et al.*, 2019; Marjanen *et al.*, 2020). These models are referred to later in this article as *UPDATE*. At the same time, we independently trained word embeddings for all time bins, which we then aligned post hoc as proposed by Kulkarni *et al.* (2015). The spaces were aligned by means of orthogonal Procrustes analysis, as first done by Hamilton *et al.* (2016).[22] We refer to these models later in this article as *ALIGN*. Aside from the frequency threshold, which we raised due to the enormous number of types[23] in our corpora, we used the default (hyper)-parameters.[24] We are releasing the models along with this article.[25]

Once the word embeddings were trained, we built, for each time bin, a similarity matrix between all the nouns extracted above. In other words, we queried the word-embedding models for a degree of 'semantic similarity'[26] between all words at hand and stored those relations in a table.

Semantic clusters can then be created. We used two hard clustering algorithms, which we describe briefly below.

We created the semantic clusters using *k*-means clustering (MacQueen, 1967) and affinity propagation (Frey and Dueck, 2007). The aim in *k*-means is to group similar data points together. Its main limitation, in our case, is that the number of clusters needs to be decided *a priori*. Our second clustering algorithm, affinity propagation, has the advantage of finding the number of clusters automatically: it splits the data into *exemplars* and *instances*, exemplars being representative tokens of their *instances*, the non-exemplar tokens in the same cluster. As Pivovarova *et al.* (2019) point out, ‘Affinity Propagation has been previously used for several NLP tasks, including collocation clustering into semantically related classes (Kutuzov *et al.*, 2017) and unsupervised word sense induction (Alagić *et al.*, 2018)’. Given that, just as in the above-cited article, we lacked a gold standard, we used standard hyperparameters[27] as available in the scikit-learn package (Pedregosa *et al.*, 2011). The main weakness of affinity propagation remains the computational and memory costs: its $O(n^2)$[28] cost is limiting in larger datasets.

As can be inferred from the previous subsections, the main strength of our approach is that it allows researchers to rely on hypotheses stemming from historical research while being data-driven. To a certain extent, we used the entirety of the data available (for English, upwards of 50 billion words) while guiding the process—the only interference, which we admit is crucial and requires domain expertise, was choosing a key adjective on which to focus. The final product, fine-grained on a one-year basis, is refined enough to be analysed in broad strokes as well as to lead to deeper dives into specific periods. Through the use of the entirety of the data, and time-specific meaning representations of words, the method avoids the common trap of teleology. Unfortunately, an inherent weakness to type embeddings is that polysemous words have a single representation in vector space, ‘ironing out’ the polysemy. This is problematic in that some words might have a certain meaning in the context of the topic at hand that is not the main sense of the word, leading to bad clusters.[29]

4 Evaluation

The method proposed in this article can be evaluated from two perspectives. First, intrinsically, we show that choices made in the preparation of the data and in the creation of the intermediate, aggregate data are reliable and produce sound output. Second, through a case study in four languages, we show that the method produces results that are useful for downstream tasks and analyses such as the study of a concept across large time scales.

Word embedding models are commonly evaluated using for example word analogies or word similarities. It should be pointed out that these evaluations are carried out on present-day data for which ground truth exists. To take but one example, Pennington *et al.* (2014) used the analogy task in Mikolov *et al.* (2013a) as well as the word similarities available in WordSim-353 (Finkelstein *et al.*, 2002). However, ground truth is not available for our data. Were we to find enough annotators[30] to create ground truth and evaluate our embeddings, creating such ground truth would entail creating an unreasonable amount,[31] given that we are training different models on different time bins.[32] This is well beyond the scope of this project. Finally, as Chiu *et al.* (2016) point out, there is no guarantee that intrinsic evaluations of word embeddings such as described above indicate better performance in downstream tasks. Instead, we rely on recent conclusions reported by Hill and Hengchen (2019), who point out that, on historical, OCRed, relatively dirty data (i.e. texts with an *F*-score of ~0.75 compared with their corresponding keyed-in ground truth) does not severely impact the performance of vector space models.

Following this, we performed a manual evaluation on certain words to make sure that the models output semantically similar words. The models for all languages except Swedish output words that were deemed correct.[33] As a result, we retrained the Swedish word embeddings after performing some data alteration: we only kept sentences that were at least ten tokens long and for which the Sparv processing pipeline could find at least 50% of lemmas.

Our manual checking confirmed that the word similarities for all languages and models post-1700 (where available) seemed meaningful, and that OCR errors were indeed captured and deemed similar.[34] Similarly, all clusters—either with *k*-means or affinity propagation—were meaningful, as illustrated in the example in Fig. 4. Plotting clusters and their evolution across time—clusters are given weight through a frequency count of their members—showed the expected signal. For example, Fig. 4[35] shows the 1860–1880 situation in Finnish-language Finnish newspapers. The 1863 peak for the legislative cluster

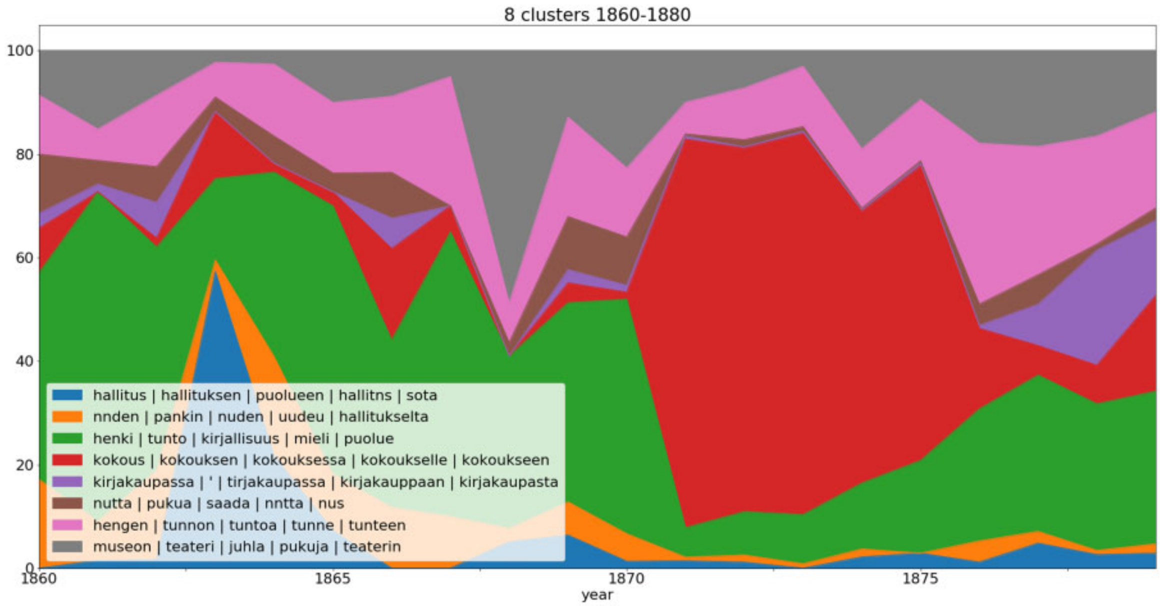

Fig. 4. Finnish-language clusters for 1860–1880

(hallitus, hallituksen, puolueen, hallitns, sota)[36] stems from content about the 1863–1864 session of the Diet of Finland,[37] the legislative assembly of the Grand Duchy of Finland. The red cluster exploding in 1871 (kokous—kokouksen—kokouksessa—kokoukselle—kokoukseen)[38] stems largely from texts relating to the Franco-Prussian War.

As we are proposing a method for which there is no gold standard and to which the notion of 'absolute truth' cannot be applied, the only way to determine whether the approach serves a purpose is to establish its usefulness.[39] The second, extrinsic, part of our evaluation is described in the next section.

5 Findings

Harnessing word embeddings to cluster words is a powerful and useful tool when matched with the right kind of research questions. In the case of the expanding discourse of 'national', for example, our clustering proves the expansion of the vocabulary of nationhood. This does not as such challenge existing historiography, but clusters based on affinity propagation indicate this change in all four languages such that the clusters make sense to a reader with historical knowledge of the period. In English, for instance, we show (in Fig. 5) how affinity propagation produces only one cluster from the time bins from the 18th century, indicating that the language of national was tied to issues related to the military and the economy (debt in particular). Earlier research focusing on the history of economic thought has pointed this out (Hont, 2005), but perhaps because of the focus on the economy, the point has not been widely accepted in the literature. Our analysis on the totality of the material does point to a dominance of economic and military discourse in the period when conceptualizing things as national started to become more common. As expected, we also show that the era of the French revolution heralded a period of gestation in which national themes were associated with political and, to a certain extent, sentiment-related themes. This entailed a clear expansion of the conceptualization of what could be perceived as national. This process continued and, consequently, affinity propagation provides many more distinct clusters for the early 19th century.

The clustering for Swedish, Finnish, and Dutch follows a similar pattern, but there are some differences in the timing and contents. In Finnish, for instance, the word *kansallinen* (as a translation of the

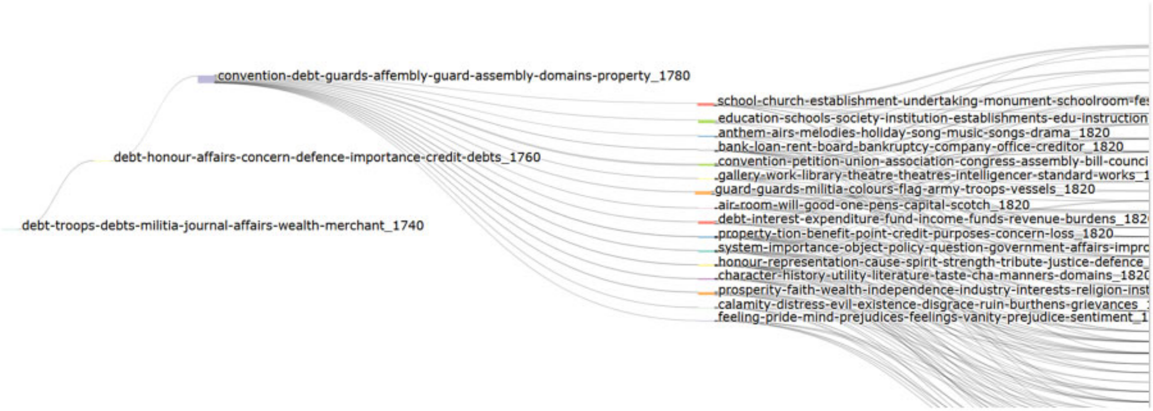

Fig. 5. Sankey chart of British clusters, 1740–1820

national) did not become really frequent before the 1850s (depending on the threshold), so naturally the development is different with a much quicker expansion of the vocabulary as established notions of nationhood readily translated into Finnish from Swedish, German, and English. As such, this suggests that the findings resonate with historical knowledge of the period and could therefore be used to further explore national peculiarities with regard to the vocabulary of nationhood.

One way of looking at the cluster differences in the case of national is to pay attention to the nature of the clusters and not only to the linked individual words. Although focusing on the 19th century, such an approach results in a clear division between sentiment-based (feeling, spirit, pride, prejudice) and object-based (bank, schools, council, government) nouns related to nationhood. When studying nationalism, this is a crucial division as they direct our attention to the growth of discourses relating to identity and affinity on the one hand and state institutions on the other. As such, they channel attention to the hypotheses about culturalization and institutionalization mentioned in Section 1. The aim in this article was not to make a full-fledged historical argument; however, it is enough to observe that, in the study of nationalism, the clusters produced through affinity propagation are perhaps more precise than what a historian reading texts would consider relevant themes, but at the same time, more detailed clusters could be thematically grouped and would seem to capture a greater level of (sometimes conflicting) signals in the data than a human reader could. We may now begin to examine in a data-driven way how different types of attitudes to nationhood emerge over time and in different places and languages.

Our distributional methods based on affinity propagation performed well in tracing general development with regard to nationhood, but also point towards more detailed findings that could be evaluated from the perspective of historical change. As such, we come much further from the use of keyword searches, simple plots of relative frequency, or even topic models in providing methods for diachronic change that relates to theories of long-term historical change. It should also be possible to use the method in analyses of other themes such as (the changing vocabularies) of secularization, modernization, and the process of civilization.

6 Conclusion

The aim in this article was to develop a data-driven method using word embeddings to examine how nation-states became central units for international politics in the 19th century Europe. The study relied on large digitized newspaper datasets in four different languages. To our knowledge, such a large-scale comparative study that grasps long-term development in as many as four languages and is statistically robust has not been attempted before. A major strength of this article is that by design, it is not limited to the study of nationhood but extends beyond it to different research questions and is thus reusable in varying contexts.

Word embeddings, which are also at the core of the method in this article, have recently gained popularity in NLP, but their successful use in historical studies is not so evident. Although there have been robust attempts to evaluate and sometimes compare NLP methods, it remains difficult to determine what is actually modelled in different cases, and whether the performances are transferable to other languages, different corpora, or dissimilar time spans. In semantically clustering, all nouns relating to the word 'national', we trained diachronic word embeddings on the entirety of the full-text historical newspaper corpora at our disposal in Dutch, Swedish, Finnish, and English. We used both *k*-means and affinity propagation clustering, of which the latter seems to provide results that are more intuitive to a domain expert. Given that there is no safe way of determining what type of embedding best suits our purpose, and that no dynamic and continuous word-embedding models could be reliably used due to the extremely uneven distribution of the data, we experimented with two relatively old architectures (CBOW and SGNS). This turned out to be a good, pragmatic choice: our manual evaluation showed that the models output semantically similar words and that the clustering lends itself to historical interpretation. As evaluation in the sense of using a gold standard is not possible, further evaluation of the method is to conduct more case studies that would allow deeper interpretations of changing vocabularies related to historical processes.

Author Contributions

S.H. designed and oversaw the study and experiments, wrote code, and ran experiments. R.R. wrote code and ran experiments, produced data visualizations, and contributed to the historical analysis. J.M. helped to design the experiments and contributed to the historical analysis. M.T. contributed to the historical analysis and acquired the funding. All the authors contributed to the writing of the article and gave final approval for publication.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye), and computational resources were provided by CSC—IT Center for Science Ltd. S.H. is currently funded by the project Towards Computational Lexical Semantic Change Detection supported by the Swedish Research Council (2019–2022; dnr 2018-01184).

Acknowledgements

The authors are grateful to many people for the constructive discussions that led to this article. In particular, and in no particular order: Melvin Wevers, Lidia Pivovarova, Elaine Zosa, Steven Claeyssens, Nina Tahmasebi, Sara Budts, Dominik Schlechtweg, Haim Dubossarsky, Joris van Eijnatten, Eetu Mäkelä, Krister Lindén, Gale Cengage, our reviewers, and the Helsinki Computational History (COMHIS) group.

References

Alagić, D., Šnajder, J., and Padó, S. (2018). Leveraging *l*exical *s*ubstitutes for *u*nsupervised *W*ord *s*ense *i*nduction. *In* Proceedings of the AAAI Conference on Artificial Intelligence (vol. 32, No. 1).

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

Antoniak, M. and Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, **6**: 107–19.

Ball, T., Farr, J., and Hanson, R. L. (1989). *Political innovation and conceptual change*. Cambridge: Cambridge University Press.

Bamler, R. and Mandt, S. (2017). Dynamic word embeddings. In *International Conference on Machine Learning*, (pp. 380–389). PMLR **70**:380–389.

Bärenbrinker, F. and Jakubowski, C. (1995). NATION UND NATIONALISMUS SEIT DEM DEUTSCHEN KAISERREICH: Eine begriffsgeschichtliche Untersuchung anhand von Handbüchern. *Archiv für Begriffsgeschichte*, **38**: 201–22.

Biernacki, R. (2014). Humanist interpretation versus coding text samples. *Qualitative Sociology*, **37:** 173–88.

Blei, D.M. and Lafferty, J.D. (2006). Dynamic topic models. *InP roceedings of the 23rd international conference on Machine Learning*, Pittsburgh, Pennsylvania, pp. 113–120.

Bolla, P. D., Jones, E., Nulty, P., Recchia, G., and Regan, J. (2019). Distributional concept analysis. In *Contributions to the History of Concepts*, Vol. 14, pp. 66–92.

Bolla, P. D., Jones, E., Nulty, P., Recchia, G., and Regan, J. (2020). The idea of liberty, 1600–1800: A distributional concept analysis. *Journal of the History of Ideas*, **81**(3), 381–406. https://doi.org/10.1353/jhi.2020.0023

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp-the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association, pp. 474–478.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., and Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC)*. Umeå, Sweden: Umeå University, pp. 17–18.

Boydens, I. (1999). *Informatique, normes et temps.* Bruxelles, Belgium: Bruylant.

Brandtzæg, S. G., Goring, P., and Watson, C. (2018). *Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century.* Leiden, The Netherlands: Brill Nijhoff.

Brenner, E. (2013). Nationalism; intellectual origins. In Breuilly, J. (ed.), *The Oxford Handbook of the History of Nationalism, Oxford Handbooks.* Oxford: Oxford University Press.

Buntinx, V., Bornet, C., and Kaplan, F. (2017). Studying linguistic changes over 200 years of newspapers through resilient words analysis. *Frontiers in Digital Humanities*, **4**: 2.

Burke, P. (2013). Language and consciousness in early modern Europe. In Breuilly, J. (ed.), *The Oxford Handbook of the History of Nationalism.* Oxford: Oxford University Press.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: how Humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds), *Advances in Neural Information Processing Systems.* Vol. **22**. Red Hook, NY: Curran Associates, Inc., pp. 288–96.

Chiu, B., Korhonen, A., and Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP.* Association for Computational Linguistics, Berlin, Germany, pp. 1–6.

Cordell, R. (2016). What has the digital meant to American periodicals scholarship? *American Periodicals: A Journal of History & Criticism*, **26**: 2–7.

Cunningham, H. (1989). The language of patriotism. In Samuel, R. (ed.), *Patriotism: The Making and Unmaking of British National Identity. Volume I: History and Politics.* London: Routledge, pp. 57–90.

de Bertier de Sauvigny, G. (1970). Liberalism, nationalism and socialism: The birth of three words. *Review of Politics*, **32**: 147–166.

van den Bos, M. and Giffard, H. (2016). Mining public discourse for emerging Dutch nationalism. *Digital Humanities Quarterly*, *10*(3).

Dietz, M. G. (1989). Patriotism. In Ball, T., Farr, J., and Hanson, R. L. (eds), *Political Innovation and Conceptual Change, Ideas in Context.* Cambridge: Cambridge University Press, pp. 177–94.

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-out: temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, pp. 457–70.

Dubossarsky, H., Tsvetkov, Y., Dyer, C., and Grossman, E. (2015). A bottom up approach to category mapping and meaning change. In Proceedings of NetWordS 2015. *Pisa: CEUR-WS.org*, pp. 66–70.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Outta control: laws of semantic change and inherent biases in word representation models. In Proceedings of the 2017 conference on empirical methods in natural language processing. *Association for Computational Linguistics, Copenhagen, Denmark*, pp. 1136–1145.

van Eijnatten, J. V. and Ros, R. (2019). The Eurocentric fallacy. A digital approach to the rise of modernity, civilization and Europe. *International Journal of History and Cultural Studies*, *7*(1): 686–736.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20:** 116–31.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930–1955. *In* Studies in Linguistic Analysis. Oxford: Philological Society, pp. 1–32.

Frautschi, R. L. (1993). The emerging notion of nationalism in French prose fiction of the enlightenment. *History of European Ideas*, **17**: 755–68.

Freeden, M. (2009). Is nationalism a distinct ideology? In Freeden, M. (ed.), *Liberal Languages Ideological Imaginations and Twentieth-Century Progressive Thought.* Princeton: Princeton University Press, pp. 204–24.

Fremann, L. and Lapata, M. (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, **4**: 31–45.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, **315**: 972–76.

Fridlund, M. and Brauer, R. (2010). Historicizing topic models: a distant reading of topic modeling texts within historical studies. In Nikiforova L.V., and Nikiforova N.V. (eds), *Cultural Research in the Context of "Digital Humanities": Proceedings of International Conference 3-5 October 2013*, St Petersburg: Russian State Herzen University, pp. 152–63.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, **4**(3): 223–72.

Gilbert, P. (2018). *The Philosophy Of Nationalism.* Routledge.

Gillani, N. and Levy, R. (2019). Simple dynamic word embeddings for mapping perceptions in the public sphere. ArXiv Prepr. ArXiv190403352.

Godechot, J. (1971). Nation, Patrie, Nationalisme et Patriotisme en France AU XVIII e Siècle. *Annales Historiques de la Révolution Française*, **43**: 481–501.

Gschnitzer, F., Werner, K. F., Schönemann, B., and Koselleck, R. (1978). Volk, nation, nationalismus, masse. In Brunner, O., Conze, W., and Koselleck, R. (eds), Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Stuttgart, Germany: Klett-Cotta.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp. 1489–1501.

Harris, Z. S. (1954). Distributional structure. *Word*, **10**: 146–62.

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 2019 DH Conference.* Utrecht, the Netherlands.

Hill, M. J. and Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, **34**: 825–43.

Hengchen, S. and Tahmasebi, N. (2021). A collection of Swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data, 7*: 2.

Hengchen, S., Tahmasebi, N., Schlechtweg, D., and Dubossarsky, H. (2021). Challenges for computational lexical semantic change. In Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S. (eds) *Computational Approaches to Semantic Change, Language Variation, Chapter 11.* Berlin, Germany: Language Science Press.

Hill, M. J., Kanner, A. O., Marjanen, J. P., et al. (2018). Spheres of "public" in eighteenth-century Britain. In *Book of Abstracts.* Presented at the Digital Humanities in the Nordic Countries (DHN), Helsinki.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hont, I. (2005). *Jealousy of Trade: International Competition and the Nation-State in Historical Perspective.* Cambridge: Harvard University Press,.

Ihalainen, P. (2007). The sanctification and democratisation of "the Nation" and "the People" in late eighteenth--century Northwestern Europe: Proposing a comparative conceptual history. *Contributions to the History of Concepts*, **3**: 125–51.

Kemiläinen, A. (1964). Nationalism; problems concerning the word, the concept, and classification. Studia Historica Jyväskyläensia III, Jyväskylä.

Kenter, T., Wevers, M., Huijnen, P., and de Rijke, M. (2015). Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management—CIKM'15.* Melbourne, Australia: ACM Press, pp. 1191–1200.

Kettunen, P. (2018). The concept of nationalism in discussions on a European society. *Journal of Political Ideology*, **23**: 342–69.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science LACSS 2014.* Association for Computational Linguistics, pp. 61–65.

Koselleck, R. (2011). Introduction and Prefaces to the Geschichtliche Grundbegriffe. *Contributions to the History of Concepts*, **6**: 1–37.

Koselleck, R. (1972). Einleitung. In Koselleck, R., Brunner, O., and Conze, W. (eds), *Geschichtliche Grundbegriffe:*

Historisches Lexikon zur politisch-sozialen Sprache in Deutschland (Vols. 1–8). Stuttgart, Germany: Klett-Cotta.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, pp. 625–35.

Kurunmäki, J. and Marjanen, J. (2018). Isms, ideologies and setting the agenda for public debate. *Journal of Political Ideology*, **23**: 256–82.

Kutuzov, A., Kuzmenko, E., and Pivovarova, L. (2017). Clustering of Russian adjective-noun constructions using word embeddings. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Valencia, Spain, pp. 3–13.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*. ACL, Santa Fe, New Mexico, pp. 1384–97.

Lähteenmäki, V. and Kaukua J. (2019). On the standards of conceptual change. *Journal of the Philosophy of History, 14*(2): 183–204.

Lahti, L., Marjanen, J., Roivainen, H., and Tolonen, M. (2019). Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging and Classification Quarterly*, **57**: 5–23.

Leerssen, J. (2006). *National Thought in Europe: A Cultural History*. Amsterdam University Press, p. 15.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Oakland, CA, pp. 281–97.

Mäkelä, E., Lagus, K., Lahti, L., et al. (2020). Wrangling with non-standard data. In S. Reinsone, I. Skadiņa, A. Baklāne, and J. Daugavietis (eds), *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*. Riga, Latvia, 21–23 October 2020, pp. 81–96 (CEUR Workshop Proceedings; Vol. 2612).

Marjanen, J., Vaara, V., Kanner, A., et al. (2019). A national public sphere? Analyzing the language, location, and form of newspapers in Finland, 1771–1917. *Journal of European Periodical Studies*, **4**: 54–77.

Marjanen, J. (2013). *Den ekonomiska patriotismens uppgång och fall: Finska hushållningssällskapet i europeisk, svensk och finsk kontext 1720–1840*. Helsinki: Helsingin yliopisto.

Marjanen, J., Kurunmäki, J., Pivovarova, L., and Zosa, E. (2020). The expansion of isms, 1820–1917: Data-driven analysis of political language in digitized newspaper collections. *Journal of Data Mining and Digital Humanities* 10.46298/jdmdh.6159.

McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., and Vatri, A. (2019). A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, **34:** 893–907.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., (2013b). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*

Milligan, I. (2013). Illusionary order: Online databases, optical character recognition, and Canadian history, 1997–2010. *Canadian Historical Review*, **94:** 540–569.

Mimno, D. and Thompson, L. (2017). The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. pp. 2873–78.

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.

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.

Nivre, J. (2005). Dependency grammar and dependency parsing. *MSI Reports*, **5133**: 1–32.

Nurmiainen, J. (2009). *Edistys ja yhteinen hyvä vapaudenajan ruotsalaisessa poliittisessa kielessä, Bibliotheca historica*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Özkırımlı, U. (2000). *Theories of nationalism: A Critical Introduction*. New York: St. Martin's Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**: 2825–30.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543.

Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J. Q., and McGillivray, B. (2019). GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Florence, Italy, pp. 56–66.

Piotrowski, M. (2012). Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, **5**: 1–157.

Pivovarova, L., Marjanen, J., and Zosa, E. (2019). Word clustering for historical newspapers analysis. RANLP 2019 3.

Recchia, G., Jones, E., Nulty, P., Regan, J., and de Bolla, P. (2017). Tracing shifting conceptual vocabularies through time. In *Knowledge Engineering and Knowledge Management*. Cham: Springer International Publishing, pp. 19–28 (https://doi.org/10.1007/978-3-319-58694-6_2).

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pp. 45–50.

Rosenfeld, A. and Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers), New Orleans, Louisiana, pp. 474–84.

Royal Dutch Library. (2017). Delpher open krantenarchief (1.0) CC BY 4.0 (text corpus). Royal Dutch Library.

Rudolph, M. R. and Blei, D. M. (2018). Dynamic embeddings for language evolution. In *WWW 2018*. ACM, Lyon, France, pp. 1003–1011.

Säily, T. (2016). Sociolinguistic variation in morphological productivity in eighteenth-century English. *Corpus Linguistics and Linguistic Theory*, **12**: 129–151.

Schierle, I. (2009). Patriotism and Emotions: Love of the Fatherland in Catherinian Russia. *Ab Imperio*, **3**: 65–93.

Schlechtweg, D., Hätty, A., Del Tredici, M., and Schulte im Walde, S. (2019). A wind of change: detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 732–746.

Schlechtweg, D. and Schulte im Walde, S. (2018). Comparing annotation frameworks for lexical semantic change. In *Proceedings of the Workshop on Automatic Detection of Language Change 2018, Gothenburg, Sweden.*

Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of NAACL-HLT*, New Orleans, Louisiana, pp. 169–174.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 Task 1: unsupervised lexical semantic change detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona, Spain: Association for Computational Linguistics.

Shoemark, P., Ferdousi Liza, F., Nguyen, D., Hale, S. A., and McGillivray, B. (2019). Room to glo: a systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics.

Smith, A. (2013). *Nations and Nationalism in a Global Era.* New York: Wiley.

Smith, A. (2008). *The Ethnic Origins of Nations*, **17 edn** [reprint]. Malden: Blackwell.

Steinmetz, W. (2012). Some thoughts on a history of twentieth-century german basic concepts. *Contributions to the History of Concepts*, **7**: 87–100.

van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. In *ICAART (1)*, Valletta, Malta, pp. 484–96.

Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. CoRR abs/1811.06278.

Tang, X. (2018). Survey paper: a state-of-the-art of semantic change computation. *National Language English*, **24**: 649–76.

Tolonen, M., Lahti, L., Roivainen, H., and 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**: 57–78.

Van Sas, N. (1999). *Vaderland: een geschiedenis van de vijftiende eeuw tot 1940 (Reeks Nederlandse begripsgeschiedenis).* Amsterdam, The Netherlands: Amsterdam University Press.

Viola, L. and Verheul, J. (2019). Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. *Digital Scholarship in the Humanities*, **35**(4): 921–43 (https://doi.org/10.1093/llc/fqz068).

Viroli, M. (1995). *For Love of Country: An Essay on Patriotism and Nationalism.* Oxford: Clarendon Press.

Wang, X., and McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, Pennsylvania, pp. 424–433.

Wevers, M. J. H. F. (2017). Consuming America: A Data-Driven Analysis of the United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890–1990. *Utrecht, The Netherlands: Utrecht University.*

Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM'18*. ACM, Marina Del Rey, California, pp. 673–681 (https://doi.org/10.1145/3159652.3159703)

Notes

1 The distributional hypothesis (Harris, 1954), popularised by the sentence 'a word is characterized by the company it keeps' (Firth, 1957) is the idea that words that occur in the same contexts tend to have similar meanings.

2 A type is a class, whereas a token is an instance of that class. In the case of textual corpora, a type is a word, whereas 'tokens' denote all instances of that word. As an example, 'to be or not to be' contains four types ('to', 'be', 'or', and 'not') and six tokens. In the case of embeddings, a type-vector is a vector representation for all instances of a word (and thus conflates polysemy), whereas a token-vector has a distinct vector representation for all occurrences of all tokens in the vocabulary.

3 Some examples from the NLP literature are: 'gay' (cheerful -> homosexual), 'tweet' (bird noise -> twitter post), and 'broadcast' (to sow seeds -> to transmit TV/radio).

4 Nonetheless, initiatives to tackle this very problem at the intersection of NLP and DH are being set up. For example, the International Workshop on Computational Approaches to Historical Language Change 2019 (https://languagechange.org/events/2019-acl-lcworkshop/, last accessed 3 November 2019) was an NLP workshop specifically aiming to be 'an outlet for disseminating cutting-edge research on topics concerning language change', to 'bring together domain experts across disciplines', and to 'raise fundamental theoretical and methodological challenges'. Another example is the SemEval 2020 Task 1 on Unsupervised Lexical Semantic Change Detection (https://competitions.codalab.org/competitions/20948, last accessed 3 November 2019), which uses unbalanced (except for English) historical data.

5 Such choices include: deciding on a (set of) word(s) to follow, 'reading tea leaves' (Chang *et al.*, 2009) in a topic model, etc. We discuss the a priori definition of a topic in the following section.

6 For recent examples and further discussion , see for instance Brandtzæg *et al.*(20182); Buntinx *et al.*(2017); and van den Bos and Giffard (2016). For a discussion on the role of the digitization of newspapers in historical research see Cordell (2016 102) and Milligan (2013).

7 https://www.delpher.nl/nl/platform/pages/helpitems?title=data+in+delpher, last accessed 4 November 2019.

8 We made use of text files created from the original XMLs by Prof. Eetu Mäkelä, whom we thank. The complete list of newspapers is available at https://www.kielipankki.fi/wp-content/uploads/klk-lehdet-fi.pdf for Finnish, and at https://www.kielipankki.fi/wp-content/uploads/klk-lehdet-sv.pdf for Swedish. Both links last accessed 4 November 2019.

9 To the best of our knowledge, there is no official data release for this public data. We retrieved the files starting with 'kubhist2' on https://spraakbanken.gu.se/lb/resurser/meningsmangder/ (last accessed 1 November 2019) and, with the help of a Språkbanken researcher, made sure that the data downloaded were the entirety of the corpus. S.H. was still affiliated with the University of Helsinki at that time.

10 https://www.gale.com/intl/primary-sources/british-library-newspapers (last accessed 1 November 2019).

11 https://www.gale.com/intl/c/17th-and-18th-century-nichols-newspapers-collection (last accessed 1 November 2019).

12 https://www.gale.com/intl/c/17th-and-18th-century-burney-newspapers-collection (last accessed 1 November 2019). Gale's British Library Newspapers collection is large, but it is not comprehensive in the same sense as the National Library of Finland's newspaper coverage, which has certain omissions of individual issues and even newspaper titles but still be considered a full run of newspapers. The nineteenth-century coverage of the British newspaper collection is more local and regional, missing some of the central-London-based newspapers in particular.

13 We clarify the reasons behind the choice of this size in the next section.

14 In NLP, dependency parsing is the task of extracting syntactic relations between words in a sentence. For a gentle introduction to dependency grammar and dependency parsing, we recommend Nivre (2005).

15 For the sake of clarity and unless we refer to a specific instance in another language, we use English words in the remainder of this article.

16 In this precise case, *nationaliteten* is the singular definite form of *nationalitet*.

17 Kielipankki/Språkbanken, https://www.kielipankki.fi/language-bank/ (last accessed 26 October 2019).

18 Språkbanken, https://spraakbanken.gu.se/en (last accessed 18 November 2020).

19 We did not carry out a systematic, large-scale evaluation of the results of dependency parsing on English or Dutch, as this is well beyond the scope of this paper.

A manual evaluation of the results for all time bins indicates relatively high precision (i.e. output is made up of nouns), but we have no information regarding recall. For more on dependency parsing on (historical, OCRed) newspapers, see van Strien *et al.* (2020).

20 See for example: Antoniak and Mimno (2018) and Mimno and Thompson (2017).

21 And thus t_2 with t_1, and t_3 with t_2, etc.

22 We used the code provided by Ryan Heuser, whom we thank. A copy is available at: https://gist.github.com/faustusdotbe/5a87007aaccc1342608c049af83fc5d2. As the code effectively deletes vectors that are not in all time bins, we made sure our nation-related nouns were not deleted.

23 As an extreme example: the 1880–1900 time bin of British newspapers contains 16,853,339,700 tokens for 664,897,852 (lower-cased) types. Dropping tokens with a frequency below 100 for the two decades leaves us with 2,043,844 types (0.3% of the original), while still retaining 15,661,451,873 tokens (92.9% of the original). To drive the point home: the enormous number of types is attributable to OCR errors.

24 For both strategies and all languages, tokens were lower cased and the (hyper)parameters were as follows: continuous bag-of-words (CBOW) for UPDATE and skip-gram with negative sampling (SGNS) for ALIGN, epochs = 5, window = 5, min_count = 100, negative = 5 (for SGNS), alpha = 0.025 (for ALIGN and the first UPDATE model) and sample = 0.001.

25 https://zenodo.org/record/3585027 (last accessed 18 November 2020).

26 For words w_1 and w_2, the similarity score is $(w1 \cdot w2)/(\|w1\|\ \|w2\|)$, where w_1 and w_2 are, respectively, the L2-normalised vectors for w_1 and w_2 and $\| \cdot \|$ denotes the Euclidean norm.

27 Although the number of clusters cannot be set, the *preference* hyperparameter (which defines the 'will' of an item to be an exemplar) can be tuned.

28 The 'Big O notation' is used to describe how calculation time or space requirements grow as the input of a certain algorithm grows. In the case of $O(n^2)$, this means that the growth is quadratic: with an input of 10 the requirement is $10^2 = 100$, with an input of 100 the requirement is $100^2 = 10{,}000$, etc.

29 We did not find that to be the case in our case study, and do not expect it to be a large problem.

30 In general, between three and five annotators are needed to be able to calculate satisfactory inter-annotator agreement. See Schlechtweg *et al.* (2018, 2020) and Schlechtweg and im Walde (2018) for a discussion on creating annotated ground-truth in diachronic corpora, and particularly on how domain experts (in that case, historical linguists) have better agreement scores—reinforcing our intuition that domain experts are needed.

31 Additionally, we echo similar work (Hengchen and Tahmasebi, 2021) in stating that diachronic word embedding models finetuned on one task are not necessarily perfect for other tasks.

32 Even if the OCR tool used is the same across all time bins, there is still variation in the input data (fonts, columns, etc.), forcing us to be thorough, and to evaluate all models.

33 A more thorough analysis should be conducted, but it seemed that the model could not abstract from the structure of definites and indefinites: for example, *brev* 'letter' and *brevet* 'the letter' were deemed less similar than *brevet* and *gardet* 'the guard'—two tokens that share nothing apart from their suffix. Note: this observation was made on the first version of the corpus (Kubhist), the final embeddings used and released were trained on the larger, more recent version (Kubhist 2).

34 The reason we do not delve into the difference between clusters created with ALIGN and UPDATE models is that we could not find a meaningful difference.

35 The clusters were created using *k*-means with eight clusters. Different cluster sizes for the same time bin show the same behaviour.

36 Literally: 'government' in the nominative, 'government' in the genitive, 'political party' in the genitive, 'government' in the nominative with an OCR error, 'war' in the nominative.

37 Note that even if the original noun (*waltiopäivät, valtiopäivät*, 'state diet') is not present as a keyword because it was not modified by 'national', the event is still present.

38 Literally 'meeting/assembly' in different cases.

39 For a more extensive discussion on the notion of 'fitness for use', see Boydens (1999).