DHQ: Digital Humanities Quarterly
2022
Volume 16 Number 2

Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914

Jaap Verheul <j_dot_verheul_at_uu_dot_nl>, Utrecht University
Hannu Salmi <hansalmi_at_utu_dot_fi>, University of Turku
Martin Riedl <riedl_dot_ma_at_gmail_dot_com>, University of Stuttgart
Asko Nivala <aeniva_at_utu_dot_fi>, University of Turku
Lorella Viola <lorella_dot_viola_at_uni_dot_lu>, University of Luxembourg
Jana Keck <keck_at_ghi-dc_dot_org>, German Historical Institute Washington
Emily Bell <E_dot_J_dot_L_dot_Bell_at_leeds_dot_ac_dot_uk>, University of Leeds

Abstract

Linking large digitized newspaper corpora in different languages that have become available in national and state libraries opens up new possibilities for the computational analysis of patterns of information flow across national and linguistic boundaries. The significant contribution this article presents is to demonstrate how word vector models can be used to explore the way concepts have shifted in meaning over time, as they migrated across space, by comparing newspapers from different countries published between 1840 and 1914. We define a concept, rather pragmatically, as a key term or core idea that has been used in historical discourse: an abstraction or mental representation that has served as a building block for thoughts and beliefs. We use historical newspapers in English, Finnish, German and Swedish from collections in the UK, US, Germany, and Finland, as well as the Europeana collection. As use cases, we analyze how the different conceptual constructs of "nation" and "illness" emerged and changed between 1840 and 1920. Conceptual change over time is simulated by creating a series of overlapping word vector models, each spanning ten years. Historical vocabularies are retrieved on the basis of vector space proximity. Conceptual change across space is simulated by comparing the historical change of vocabularies in newspaper collections from different nations in several languages. This computational approach to conceptual history opens up new ways to identify patterns in public discourse over longer periods of time and across borders.

1. Introduction

Big data opens a window onto the global dimensions of our cultural heritage and history [Eijnatten et al. 2014]. The availability of large datasets of digitized newspapers offers unprecedented opportunities to explore the transnational and intercultural connections of the western world. This article explores the ways in which word vector models can be used to analyse how the exchange of knowledge, ideas, and concepts across borders and languages is reflected in newspapers. It focuses on the period of globalization between the middle of the nineteenth century and the return to isolationism and nationalism in the wake of the First World War. 1

This publication is based on a strand of the international research project "Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914" (OcEx). OcEx has brought together an international and interdisciplinary research consortium to examine patterns of information flow across national and linguistic boundaries in nineteenth-century newspapers by linking digitized newspaper corpora currently siloed in national collections. Funded by the Trans-Atlantic Platform for the Social Science and Humanities' "Digging Into Data Challenge," this effort builds on the recent rediscovery of global and world history to overcome the national perspectives of the past.[1] 2

The larger academic research question this article addresses is how we can trace the migration of concepts over time and across space. Specifically, this article discusses two sets of very different concepts which changed during the second half of the nineteenth century and the early twentieth century. The first is the set of concepts that denotes the most public aspects of human existence: the collective identity that is framed in terms of "nation," "state," "people," and "national identity." The analysis of this category of concepts also allows comparison with traditional approaches to conceptual history, as discussed below. The second, in contrast, deals with the ultimate personal experience reflected in discourses about "illness" and "health." Although belonging to a different domain than national identity, illness and health likewise reflect fundamental experiences in human life which can reach collective dimensions, as in the case of epidemics and pandemics. More importantly, the concepts that people in different nations and cultures have employed to understand and respond to these circumstances of well-being or illness are conditioned, or mediated, by shared ideas about the human condition which have changed over time. Both national identity and health, then, are cultural constructions which should be understood in their historical and geographical context. These conceptual sets are interesting use cases because there is already a rich body of academic literature that suggests important changes in the understanding of both "nation" and "illness" as labels for abstract ideas and knowledge domains took place during this period [Hobsbawm 2012] [Gellner 2007] [Anderson 2006]. 3

How can we use computational word vector models to identify the changing vocabulary in which concepts are expressed? While many definitions of concepts circulate in various academic disciplines such as linguistics, philosophy, and psychology, no consensus has emerged [Margolis and Laurence 2005] [Margolis and Laurence 1999]. Within this article, we define a concept rather pragmatically as a key term or core idea that has been used in historical discourse, and as an abstraction or mental representation that has served as a building block for thoughts and beliefs. We approach historical concepts not as analytical research tools constructed by historians to understand the past, but rather as the mental constructions of actual historical actors as they are expressed in formal and informal public discourse, political tracts, fiction, or life writing. We focus on historical newspapers as a serial and coherently structured source to demonstrate how such concepts are articulated [Broersma and Harbers 2018] [Douglas 1999] [Ginneken 1998] [Kunczik 1997] [Moran 1978]. 4

The significant contribution this article presents is to demonstrate how word vector models can be used on different newspaper collections in different languages to explore the way concepts have shifted in meaning, as they migrated, between 1840 and 1914. Rather than attempting a comprehensive overview of conceptual history in this period, the aim of this article is to demonstrate the viability of this methodology and the significance of newspapers as representatives of the public sphere [Habermas 1991, 181–95]. This computational approach to conceptual history opens up new ways to identify patterns in public discourse over longer periods of time and across borders. We also contribute to the discussion about the relationship between concepts and natural language in the larger process of knowledge discovery [Jackson and Moulinier 2007] [Harras 2000] [Weitz 1988]. 5

2. Conceptual change and word vector models

Historical concepts are essential to our understanding of the past [Weitz 1988]. Concepts such as citizenship, democracy, migration, liberty, security, trust, and health constitute the continuous foundation of changing historical debates and have been compared to the "unit-ideas" that form the building blocks of human discourse, similar to the elements that form chemical compounds [Lovejoy 1933, 4]. Tracing the change, continuity, and replacement of these concepts is vital for historians and other humanities scholars, since concepts are the lenses through which people in the past understood the world. 6

Our approach builds on the tradition of conceptual history (*Begriffsgeschichte*), as established by the Bielefeld school of Reinhart Koselleck and the Cambridge School associated with J. G. A. Pocock and Quentin Skinner. Koselleck argues key terms such as nation, citizenship, and family reflected fundamental changes in the social and political structures of European societies which took place from the mid-eighteenth century onwards. The magnum opus of eight volumes, which he edited together with Otto Brunner and Werner Conze, traces the shifting meaning of one hundred and thirty of these leading concepts (*Geschichtliche Grundbegriffe*). These concepts reflect modernity, and social and political change [Brunner et al. 1972]. The methodological question this article addresses is whether computational methods can identify the fundamental changes in vocabulary that demonstrate the kind of conceptual change Koselleck's group describes [Koselleck 2002] [Koselleck 2004]. 7

The Bielefeld group is often contrasted with the conceptual history tradition established in the United Kingdom. The Cambridge School of Intellectual History established by Pocock and Skinner is interested in the inherent historical contextuality of conceptual change. Their work contributed to a shift from a history of ideas to a history of concepts, and pushes against any sense of keyword "definitions" [Skinner 1969]. Skinner's work on concepts such as the "state" and "liberty" has been key to the development of conceptual history as a field. Pocock's study of the key ideas of "Machiavellian thought," which moved from Renaissance Italy to Civil War Britain and then crossed the Atlantic to inform the American Revolution, suggests an interconnectedness of concepts across linguistic and geographical boundaries [Pocock 2016] [Skinner 2012] [Skinner 1978]. 8

Building on these German and British research initiatives, and related projects in Scandinavia, the Netherlands and elsewhere, conceptual history became an established – if undertheorized – field 9

in the 1970s, producing a number of monumental studies that have been described as a true “pyramid of the mind” [Steinmetz 2016] [Müller and Schmieder 2016]. Conceptual drift – the historical shift in the meaning of concepts and the words that articulate them – has mainly been analyzed chronologically in the context of a single language [Betti and van den Berg 2014] [Kuukanen 2008]. De Bolla demonstrates how keyword search methods can be used to trace how the concept of human rights developed during the era of the American Revolution [Gavin 2015] [De Bolla 2013].

10 Little has been done to study the way in which concepts change if they are translated into different languages and cultural contexts. Studying “what happens when concepts move between different kinds of modernities and their associated temporalities” [Müller 2014, 88] remains one of the unsolved challenges in the history of concepts, and is thus key to the methodology of this article. A better understanding of conceptual migration is essential if we want to know how news, knowledge, ideas and ideologies circulated in the globalizing world that emerged at the end of the nineteenth century, how nations emerged as imagined communities around common newspaper readership, or how people in various parts of the world learned about and gave meaning to pandemics diseases and disasters.

11 In recent years, word vector models have been used to represent concepts in text corpora. Rather than relating words to objects or events in the real world, word vectors represent a multidimensional network of relations between words in a large textual corpus. Word vectors are representations based on the distributional hypothesis, which assumes that words tend to be similar if they occur in a similar context. The meaning of words is represented by the relative positions of their vectors in that semantic space. The effectiveness of this associative approach to identifying concepts has been demonstrated [Landauer et al. 1997] [Lund and Burgess 1996] and applied to the scientific discourse of the seventeenth century; Pumfrey, Rayson and Mariani (2012) have conducted significant work that compares a manual approach to a digital humanities methodology using corpus linguistics tools.

12 Related approaches have also been proposed in the computational linguistic community. However, most approaches target a change of word senses, i.e. the different meanings a word can have, rather than concepts. In distributional semantic models (based on the distributional hypothesis by Harris) a word is defined by the context in which it appears [Harris 1954]. This notion became popularized as the idea that “You shall know a word by the company it keeps!” [Firth 1957, 11]. Such representations are used to show that word senses change over time. Distributional thesauri have been used, for instance, to compute word similarities for different time points in Google Books data [Mitra et al. 2015]. Computing different embedding representations for several time spans can show how word sense changes over time [Hamilton et al. 2016]. This unsupervised computational methodology demonstrates how new senses are born, disappear, or remain consistent. By aligning the embeddings of different time spans in the same vector space, it has been possible to demonstrate how terms change position over time in the vector space.

13 The use of word vectors models received a boost by the word2vec algorithm developed by Google. Using a large corpus of text as input, this algorithm offers an efficient and reliable method to produce a multi-dimensional vector space, in which words that share common contexts in the corpus are located close to each other [Mikolov et al. 2013a] [Mikolov et al. 2013b]. As Word2vec can place words from a large text corpus such as books, web pages or newspaper articles in a vector space that represents semantic similarity it can indicate, or “predict,” semantic relations between words [Baroni et al. 2014]. Yet most corpus-based distributional semantic methods use a bag-of-words approach that lumps words together without taking their temporal order or historical origin into account, assuming that the meaning of words remains stable over time. As such, the challenge remains finding a computational approach that accounts for both the historicity of changing word meanings and also conceptual change.

3. Computational methodology

14 In this article, concepts are hypothesized as semantic spaces within a vector space in which vocabularies can be identified that express core ideas, abstractions or mental representations. We operationalize this by following conceptual change in different regions of the western world, tested computationally by comparing the changing vocabularies in word vector spaces of a number of concepts with a global presence, and then interpreting them with specific domain knowledge.

15 A methodology has recently been developed to use word vector models, created with the word2vec algorithm, to interrogate historical changes in vocabularies linked to concepts. The digital history team at Utrecht University, in collaboration with the Netherlands eScience Center, has developed the tool ShiCo (Mining Shifting Concepts through Time) which enables researchers to test this methodology on collections of digitized newspapers [Martinez-Ortiz 2016] [Kenter et al. 2015] [Kenter 2013]. This tool creates word vector models of ten years each, with an overlap of two years, e.g. 1840–49, 1842–51, 1844–53. When users enter one or more search terms (or “seed words”), the algorithm will return the vocabularies found in the surrounding vector space for each model. This offers an overview of gradual changes over time in vocabularies associated with the concepts [Wevers and Koolen 2020] [Viola and Verheul 2020].

16 We use ShiCo as a backend to access the embeddings of all corpora used in this study. [2] In “adaptive mode” ShiCo is able to use the vocabularies that are found in a vector model as seed words for the next model, as a way to trace gradual concept shift. We used the “non-adaptive” mode to trace the changing vocabularies associated with the same seed words over time, in order to create a more stable baseline from which to compare vocabularies from different language corpora in the same time period. For the generation of the word embeddings, we deploy gensim and compute CBOW models, with 100 dimensions, a window size of five, a minimum word count of five, and with five negative samples. [3]

17 In our study, we compute similarities for a term for each window of time as set in the interface, using the same seed words. Since the windows of time overlap, most of the semantic relations between words remain stable, resulting in gradual changes over the years. We argue that this method offers a way to understand gradually-changing words that are used to articulate the same topic, concept, or idea. One can debate to what extent these semantic spaces in a word vector model represent historical concepts [Recchia et al. 2017]. This article tests this hypothesis by employing ShiCo on several different digitized newspaper collections.

18 Due to OCR issues with historical newspapers, particularly issues arising from the similarity of some characters, our corpora have large amounts of words with erroneous variations. Furthermore, the fraktur font is used in historical German-language newspapers, and the OCR models deployed to convert German-language newspapers are not trained on such a font. For example, the term *Krankheit* (illness) is also similar to the writing variations *Krankbeit*, *Krankhcit*, *Kraiikheit*, *Kraukheit.* To solve this issue, we merge variations with manual correction lists (e.g. *Krankheit*: *Krankbeit*, *Krankhcit*, *Kaiikheit*, *Kaukheit*) and use the mean of all similarity scores as the similarity score.

4. A multilingual dataset of digital newspaper corpora

19 We deploy parallel instances of the ShiCo software to create word vector models for the newspaper corpora we use as datasets (https://oceanicexchanges.org/news/). We use eight different corpora including four different languages, namely English, German, Finnish and Swedish (Table 1). These corpora were selected because they represent large national newspapers collections which can be compared to gain an understanding of the transnational and cross-cultural circulation of knowledge and ideas, as was the starting point of the “Oceanic Exchanges” project, and on the basis of their availability in digitized form with sufficient OCR reliability [Beals and Bell 2020]. The date range 1840–1914 was selected to capture a the period of rapidly expanding cross-Atlantic trade, traffic, migration, and cultural exchange which can be understood as the first wave of globalization. This period was also the heyday of newspapers as the first big data for a mass audience [O’Rourke 1999] [Osterhammel 2014] [Nolan 2012].

20 *The Times Digital Archive* (TDA) was the first online digitized newspaper collection of British newspapers. Currently, it contains material up to 2010 comprising over 1.6 million pages from 70,000 issues of *The Times of London*, sub-divided or zoned into 11.8 million articles, catalogued by category, including advertising, editorial and commentary, news, business, news, people and photojournalism. The data for the digitized newspapers comes in two forms: a scanned image of each newspaper page at 300 DPI, zoned and sub-divided at article level, and an XML file containing the text (OCR) and metadata for each article. The machine-readable text appears within the XML file, surrounded by metadata that describes various features about the article, including the title, issue, date, section, and page number. The collection is available in many state and institutional libraries throughout the world through a commercial licencing arrangement with Gale. The underlying text and metadata can be accessed by request, with a cost recovery fee. [4]

21 The Finnish newspaper corpus has been provided by the National Library of Finland and is downloadable as a data dump via the Language Bank of Finland. The collection includes all published issues from the birth of newspaper publishing in the country in 1771 up to 1920. Since the Finnish press has mainly been published in two languages, Swedish (SE-NLF) and Finnish (FI-NLF), we compute two models on newspapers from 1840 to 1914. Within this timeframe 369 different titles were published, totalling 3.6 million pages: 26% in Swedish, 74% in Finnish. In the corpus, the amount of data in Finnish is especially thin in the 1840s and 1850s as the Finnish-language press only expanded towards the end of the century. By 1890, 47% of all published newspaper pages were still in Swedish. The Finnish-language press was furthermore printed mostly with Gothic typeset, which results in a substantial amount of OCR noise.

22 To investigate the shift of concepts in German-language newspapers, we compute embeddings for three different corpora: the German parts of the *Europeana* (DE-EU) corpus, the German-language newspapers from *Chronicling America* (DE-CA) and the *Berlin State Library* (DE-SBB) corpus.

23 The *Europeana* corpus[5] is a collection of 50 million digitized items such as books, newspapers, music and artworks that have been published in Europe. From these items, there are 876,724 newspapers available which have been digitized and OCRed. The majority of pages written in German are from Austria (1,184,091), followed by Germany (822,085), Italy (683,062), Estonia (39,540) and Lithuania (27,030). This corpus comprises 129 different newspapers from 1840 to 1913, 6,272.3 million tokens, and around 494.8 million words. The number of tokens indicates the

words in a corpus regardless of how often they are repeated, while the word count reflects the number of distinct word types.

Chronicling America is a web-based platform that gives access to newspapers published in the United States from 1789 to 1963, with descriptive information about the newspapers and digitization of historic pages.[6] The majority of newspapers fall within the range 1850–1922. In addition to American titles published in English, it also includes newspapers of twenty ethnicities such as, for instance, Native American, Czech, Swedish, Icelandic, Danish, Finnish, French, German, or Italian. We use newspapers that were published in German (DE-CA) between 1840 and 1913. This collection consists of 57 newspaper magazines, resulting in a corpus of 929.8 million tokens and 49.5 million words. 24

The newspaper corpus from the Berlin State Library (DE-SBB) is a collection of historical newspapers published in the German states. We selected newspapers from 1872 to 1913, which is a corpus of 3,111.6 million tokens and 119.8 million words. Compared to the previous two corpora, this corpus is very specific according to its locations, as it contains only articles from three newspaper publishing houses in Berlin. 25

Name	Repository	Language	Origin	Timespan	No. of words (million)	No. of tokens (million)
TDA	Times Digital Archive (Gale Cengage)	English	UK	1840-1920	221.5	3,544.5
FI-NLF	National Library of Finland	Finnish	Finland	1840-1914	225.8	2,966.4
DE-CA	Chronicling America	German	USA	1840-1910	49.5	929.8
DE-EU	Europeana	German	Austria, Germany, Estonia, Lithuania [26] [27]	1840-1912	494.8	6,272.3
DE-SBB	Berlin State Library	German	Germany	1872-1912	119.8	3,111.6
SE-NLF	National Library of Finland	Swedish	Finland	1840-1914	80.9	2,321.0

Table 1.

Table 1: Newspaper datasets 28

5. Case study I: Nations and national identity

The nineteenth century was, in many respects, the age of nationalism and national identity. Although new nation states produced long genealogies of invented history that suggested perennial antecedents in tradition, common language, and ethnic affiliation, those were relatively new constructs cobbled together out of the regional and tribal identities of the *ancien régime*. Modern nation states were very much the product of modernization and industrialization, processes which required unification of time, language, education, and collective behaviour, political emancipation and mobilization of the middle classes, and the integration of the new urban masses. Rather than emerging from natural, perennial, or "primordial" identities, nations were deliberately based on "constructed" ideas of national identity [Gellner 1997] [Gellner 2007] [Hobsbawm and Ranger 2010] [Hobsbawm 2012]. 29

Several authors have indicated the importance of newspapers and other mass media in the formation of these new national identities [Andrews 2014] [Rosie et al. 2004] [Billig 1995]. If the nineteenth century was the age of national identity, it was also the age of newspapers, magazines, journals, and the book industry. As the many local newspapers fostered urban and regional identities, the new national newspapers connected and informed the rapidly emerging readership of wealthy middle-class citizens. Anderson demonstrates that modern mass media played a vital role in the formation of these "imagined communities," as they created collective illusions of shared experience, group solidarity, and common fate [Anderson 2006]. National identity and the emergence of capitalist media appear, therefore, to be interconnected. At the same time, the growing global information networks of the nineteenth century increasingly allowed national newspapers to inform their readers about developments in the world. These newspapers not only provided factual news about government decisions, political upheavals, military expeditions, trade opportunities, or new inventions, but also constructed political ideologies and movements. 30

Conceptual history has produced impressive studies to trace the emergence of the new concepts of nation, nationalism, and the people. A large section of *Geschichtliche Grundbegriffe* (*GGB*, 1972–97) is dedicated to the emergence of the related concepts "Volk, Nation, Nationalismus, Masse" in German discourse [Brunner et al. 1972]. Although this *GGB* chapter traces the concept from antiquity to the end of the Cold War, the main conceptual turning point can be traced to the period from the second half of the nineteenth century to the early twentieth century, when the nation became the focus of political mobilization. Conservative, liberal, Catholic, and socialist movements developed their own political and social lexicons to express their ideas about national identity. The concept of "nation" thus became a synonym for social integration. As Koselleck hypothesizes, the German word *Volk* (people) remained a concept (*Begriff*) with a mainly state-centred and political meaning, whereas the loanword *Nation* (nation) was largely apolitical. With the rise of the working-class movement during the second half of the century, the concept of the nation became gradually connected with the democratic participation of the masses reflected in references to terms such as proletariat, masses, and mob (*Proletariat*, *Masse*, and *Menge*) [Koselleck 1992a] [Koselleck 1992b]. 31

This case study aims to trace such conceptual changes in large datasets of digitized newspapers repositories, to explore how the development of national discourse unfolded in different European countries and across the Atlantic. The objective is to address the question of whether the emergence of national discourse can be seen as a universal and uniform phenomenon in the western world, or rather whether national traditions can be discerned as they are reflected in the vocabulary used to express the concept of the nation. In other words, by using word vector models, the analysis will shed light on the way in which newspapers "invented" national communities and created shared national experiences. What conceptual changes emerge in the public debates in the different countries around this discursive topic? 32

In this experiment, we compare the changing vocabulary around the concept of nation by implementing word vector modelling on national newspaper corpora in different languages in the period 1840–1914. Word vector models were used for newspapers published in English, Swedish, Finnish, and German, as published in the UK, Finland, Germany, and the US. 33

As entry points into the semantic vector space (seed words), we use different synonyms for the term "nation" derived from the *GGB* and the secondary literature, to which we add the terms that resulted from the word vector modelling.[7] To compare the changing vocabulary and improve readability, we use the (modern) English translations of these terms as anchor points. 34

A. The transnational migration of national identity

The term "nation" derives from the Latin word *nationem* for birth, origin, or tribe. The geographical concept of the nation is thus firmly established, and is still evident in the 1840–1914. Ironically, the conceptual stability of nation in public discourse seems more stable than the geopolitical realities of the time. Within the word vector space of TDA, the search term "England," for example, produced a stable set of names of other nations; "England" shares a semantic space with "Scotland," "Ireland," "France," and "Belgium," and also with "America." Perhaps more remarkably, the term "Europe" is also part of that stable presence. "Germany" and "Spain" are mentioned in 1848, as the political revolutions in both countries inject them into public discourse, and ensure they persist after (see Table 2 below). 35

decade	france	ireland	scotland	america	germany	europe	england	principality	spain	italy	belgium	wales
1840s	0.68	0.71	0.73	0.70	0.25	0.52	0.52	0.4	0.26	0.00	0.13	0.00
1850s	0.69	0.68	0.64	0.69	0.65	0.66	0.39	0.00	0.38	0.00	0.00	0.00
1860s	0.66	0.71	0.70	0.69	0.64	0.65	0.53	0.27	0.00	0.00	0.13	0.00
1870s	0.73	0.52	0.66	0.69	0.73	0.70	0.00	0.00	0.66	0.66	0.66	0.00
1880s	0.65	0.63	0.72	0.65	0.65	0.26	0.36	0.68	0.49	0.00	0.00	0.49
1890s	0.62	0.64	0.67	0.48	0.63	0.59	0.65	0.70	0.00	0.35	0.00	0.12
1900s	0.62	0.67	0.72	0.46	0.23	0.11	0.64	0.68	0.00	0.12	0.11	0.47
1910s	0.65	0.68	0.69	0.61	0.00	0.56	0.00	0.62	0.00	0.56	0.60	0.00

Table 2. Similarity scores for the most frequent word vectors in TDA related to “England,” grouped by decade.

In the German newspaper corpus (DE-SBB), the conceptual neighbours of “Deutschland” are also its geographical neighbours France, Austria-Hungary, England, Russia, Belgium and Italy, more or less in that order, along with Europe. 36

decade	deutschland	frankreich	england	europa	rußland	oesterreich-ungarn	oesterreich	belgien	spanien	italien
1870s	0.82	0.85	0.81	0.84	0.78	0.80	0.39	0.39	0.40	0.19
1880s	0.82	0.85	0.84	0.83	0.81	0.48	0.16	0.00	0.00	0.00
1890s	0.83	0.86	0.86	0.82	0.32	0.00	0.32	0.00	0.00	0.00
1900s	0.81	0.85	0.85	0.39	0.00	0.39	0.81	0.00	0.00	0.00

Table 3. Similarity scores for the most frequent word vectors in DE-SBB related to “Deutschland,” grouped by decade.

The Finnish dataset FI-NLF shows a sense of Nordic insularity, as all the proper names of nation states are from Nordic countries: Finland, Sweden, Norway, Denmark (*Suomi*, *Ruotsi*, *Norja*, *Tanska*). In Finnish there is a distinction between *Suomi* (Finland) and *suomi* (Finnish language), but it is lost here because the word embeddings are not case sensitive. Therefore, many word vectors are also related to the Finnish language (like *suomenkieli*, *opetuskielenä*, *kieli*, *suomenkielinen*). *Historia* (history) and *Pohjanmaa* (Ostrobothnia, a region in Finland) are also included, but references to other European nations or to the concept of Europe are remarkably absent in the Finnish word vector space of Nordic countries. 37

decade	suomi	ruotsi	suomenmaa	norja	suomenkieli	opetuskielenä	kieli	historia	tanska	pohjanmaa	finlands	suomenkielinen
1840s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00
1850s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.28	0.00	0.00	0.58	0.0
1860s	0.06	0.35	0.19	0.06	0.21	0.25	0.40	0.29	0.00	0.00	0.07	0.33
1870s	0.59	0.70	0.65	0.31	0.62	0.50	0.29	0.00	0.05	0.00	0.00	0.11
1880s	0.61	0.66	0.60	0.61	0.46	0.05	0.00	0.00	0.50	0.00	0.00	0.00
1890s	0.66	0.62	0.47	0.47	0.11	0.00	0.00	0.00	0.00	0.52	0.00	0.00
1900s	0.77	0.65	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4. Similarity scores for the most frequent word vectors in FI FNL related to “suomi,” grouped by decade.

This semantic stability may be seen as an illustration of the fact that the nation states are defined in relation to each other, as Koselleck observes [Koselleck 1992a, 145–6]. The relative stability in the word vector models stands in sharp contrast to the geopolitical upheavals of this period. The First World War, for instance, is virtually invisible in the long-term representation of the word vector models. 38

The shifting vocabularies in which the concept of the nation is expressed in public discourse, have changed over time considerably. Although references to the nation appear in all European countries during the nineteenth and early twentieth century, the semantic associations reflect the different contexts in which this concept functions. In TDA, the semantic connection between the term “nation” and “monarchy,” which is initially very strong, disappears after 1884. The nation is consistently associated with “people” over the full period under review, reflecting the broad meaning of the word in which national identity is based on the notion of a people sharing similar language and ethnic origins. Unsurprisingly, nation is also associated with the term “empire” in England, although this connection weakens after 1908 (see Table 5). 39

decade	people	empire	patriotism	independence	community	mankind	monarchy	democracy	nations	fatherland	nationality	rulers
1840s	0.76	0.79	0.74	0.75	0.00	0.75	0.6	0.14	0.15	0.00	0.74	0.76
1850s	0.79	0.76	0.77	0.79	0.00	0.76	0.76	0.00	0.45	0.00	0.76	0.79
1860s	0.75	0.77	0.77	0.80	0.60	0.75	0.76	0.29	0.00	0.15	0.45	0.47
1870s	0.59	0.77	0.45	0.79	0.75	0.00	0.75	0.75	0.73	0.60	0.15	0.00
1880s	0.74	0.75	0.59	0.75	0.74	0.29	0.30	0.76	0.00	0.30	0.44	0.00
1890s	0.73	0.78	0.74	0.29	0.73	0.44	0.00	0.29	0.58	0.76	0.00	0.14
1900s	0.74	0.59	0.72	0.43	0.74	0.44	0.00	0.79	0.74	0.74	0.00	0.00
1910s	0.77	0.00	0.71	0.00	0.77	0.77	0.00	0.83	0.75	0.73	0.00	0.00

Table 5. Similarity scores for the most frequent word vectors in TDA related to “nation,” grouped by decade.

A narrower political interpretation of national identity appears only in the 1860s in TDA, when nationalism became associated with conservatism, but also with the competing political affiliations of liberalism, democracy, and Toryism. The growing political awareness and polarization in the 1870s is reflected by the connection between “nationalism” and the term “radicalism” within the same word vector space. That nationalism is also placed in a religious context is suggested by the emergence of religious terms such as “protestantism,” “clericalism,” and “ultramontane” (the latter reflecting the political ambitions of the pope). The role of the state is difficult to assess on the basis of the word vector models because the word “state” possesses several meanings in English. Yet the term “state” seems associated with the more administrative terms “to govern,” “government,” and “administration” from the 1840s on, and the more negative term “disorganisation” from the mid-1850s on, hinting at a tendency perhaps, to describe the state, rather than the nation, in utilitarian terms. 40

decade	conservatism	liberalism	democracy	radicalism	toryism	republicanism	imperialism	protestantism	separatist	clericalism	puritanism	exclusiveness
1840s	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00
1850s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1860s	0.56	0.56	0.26	0.00	0.55	0.15	0.00	0.43	0.00	0.00	0.54	0.41
1870s	0.75	0.59	0.73	0.29	0.44	0.44	0.29	0.00	0.59	0.00	0.00	0.44
1880s	0.77	0.78	0.76	0.76	0.75	0.00	0.00	0.46	0.29	0.64	0.45	0.00
1890s	0.81	0.82	0.80	0.82	0.31	0.64	0.48	0.64	0.32	0.83	0.00	0.00
1900s	0.77	0.81	0.77	0.83	0.15	0.78	0.80	0.00	0.15	0.00	0.00	0.15
1910s	0.00	0.82	0.78	0.83	0.76	0.81	0.80	0.00	0.00	0.00	0.00	0.00

Table 6. Similarity scores for the most frequent word vectors in TDA related to “nationalism,” grouped by decade.

After German unification in 1871, the newspapers reflected the strong cultural origins of the German sense of national identity. The term “nation” is strongly associated with “christianity” (*Christenheit*) in the 1870s, the period marked by Bismarck’s cultural war (*Kulturkampf*) against Catholicism. Although the Iron Chancellor faced defeat in the political arena, the emergence of the secular alternative “civilization” (*Zivilization*) in the 1880s suggests that he won in public discourse. This seems to confirm Koselleck’s conclusion that the German sense of nation is primarily a cultural construct (*Kulturbegriff*). Nevertheless, the word vector models also show the connection with power. The emergence of geopolitical and military terms such as “sea power,” “great power,” and “world power” (*Seemacht*, *Großmacht*, *Weltmacht*) in the vocabulary around the term “nation” reflects Germany’s global geopolitical ambitions, which were translated into colonial expansion and maritime muscle-flexing at that time. A similar German sense of superiority and the civilizing mission is reflected in the close relation emerging between the term “civilization” (*Civilisation*) and geopolitical markers such as “sea power” (*Seemacht*) and “world politics” (*Weltpolitik*) at the end of the nineteenth century. 41

decade	dynastie	nation	civilisation	einheit	republik	zivilisation	diplomatie	demokratie	volksvertretung	seemacht	politik	colonialpolitik	weltmacht
1870s	0.57	0.87	0.77	0.79	0.77	0.76	0.19	0.81	0.76	0.00	0.00	0.00	0.00
1880s	0.78	0.86	0.76	0.78	0.77	0.15	0.60	0.61	0.61	0.15	0.15	0.30	0.00
1890s	0.76	0.78	0.59	0.44	0.44	0.29	0.29	0.00	0.00	0.74	0.43	0.29	0.43
1900s	0.75	0.00	0.37	0.00	0.00	0.73	0.74	0.00	0.00	0.76	0.73	0.36	0.74

Table 7. Similarity scores for the most frequent word vectors in DE-SBB related to “nation,” grouped by decade.

decade	nation	dynastie	monarchie	volksvertretung	demokratie	diplomatie	einheit	nationalität	großmacht	unabhängigkeit	rationalität	aristokratie	völker
1840s	0.82	0.86	0.79	0.00	0.00	0.30	0.00	0.46	0.81	0.15	0.63	0.62	0.00
1850s	0.82	0.87	0.62	0.00	0.16	0.79	0.15	0.46	0.78	0.60	0.31	0.62	0.31
1860s	0.84	0.86	0.80	0.15	0.80	0.81	0.00	0.15	0.31	0.00	0.00	0.00	0.47
1870s	0.85	0.85	0.80	0.63	0.47	0.32	0.31	0.15	0.00	0.00	0.15	0.00	0.00
1880s	0.87	0.84	0.81	0.80	0.47	0.15	0.64	0.63	0.00	0.00	0.47	0.00	0.00
1890s	0.82	0.85	0.63	0.79	0.32	0.00	0.79	0.15	0.00	0.79	0.00	0.00	0.00
1900s	0.00	0.84	0.00	0.80	0.83	0.00	0.79	0.00	0.00	0.80	0.00	0.00	0.00

Table 8. Similarity scores for the most frequent word vectors in DE-EU related to “nation,” grouped by decade.

The political interpretation of national identity within Germany is expressed by the term “nationalism,” which can be understood as a developing term first associated with foreign influences from the West and East (*Bonapartismus*, *Polenklubs*) and the indigenous national spirit (*Nationalgeistes*) propagated to counter France’s national ambitions. Interestingly, similarly to the UK, the German term “nationalism” becomes contested at the end of the century, as is evinced by its semantic association with pejorative terms with the suffix -ismus such as “clericalism,” “imperialism,” “absolutism,” “radicalism,” “chauvinism,” and “fanaticism” (*Klerikalismus*, *Imperialismus*, *Absolutismus*, *Radikalismus*, *Chauvinismus*, *Fanatismus*). In German discourse, the term “state” (*Staat*) does not make inroads, as the associated vocabulary suggests competition between governmental intervention and private initiatives, such as “compulsory insurance” (*Versicherungszwang*), “private firm” (*Privatbetrieb*), “artisans” (*Handwerkerstand*), and, at the end of the century, “taxation” (*Steuerbezahler*, *Fiscus*). The Hegelian conception of national identity, lastly, is reflected in the frequent use of the term “spirit of the nation” (*Nationalgeist*) in connection with the German “nation.” In the 1870s, this term is associated with terms such as “idealism,” “spirit,” “patriotism” and “national identity” (*Idealismus*, *Geist*, *Patriotismus*), which may reflect an essentialist, or Hegelian, interpretation of national identity. In the 1880s, terms such as “idealism,” “altruism,” “fatalism,” “naturalism,” and “national character” (*Idealismus*, *Altruismus*, *Fatalismus*, *Volkscharacter*) dominate, perhaps suggesting a conceptual change towards a more moral and personal interpretation of national identity. 42

decade	despotismus	radikalismus	klerikalismus	absolutismus	imperialismus	individualismus	socialismus	ultramontanismus	parlamentarismus	chauvinismus	sozialismus
1870s	0.00	0.00	0.19	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00
1880s	0.33	0.16	0.00	0.33	0.16	0.52	0.00	0.00	0.00	0.00	0.50
1890s	0.67	0.84	0.69	0.50	0.34	0.52	0.68	0.69	0.67	0.68	0.17
1900s	0.86	0.88	0.9	0.87	0.87	0.00	0.84	0.85	0.85	0.86	0.00

Table 9. The similarity scores for the most frequent word vectors in DE-SBB related to “nationalismus,” grouped by decade.

decade	sozialismus	radicalismus	rationalismus	radikalismus	klerikalismus	ultramontanismus	chauvinismus	sozialismus	liberalismus	antisemitismus	konservatismus
1840s	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.41
1850s	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1860s	0.49	0.64	0.81	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.16
1870s	0.15	0.45	0.61	0.00	0.16	0.15	0.00	0.16	0.00	0.00	0.43
1880s	0.67	0.32	0.00	0.33	0.33	0.49	0.33	0.17	0.49	0.17	0.17
1890s	0.90	0.71	0.00	0.91	0.90	0.53	0.88	0.89	0.89	0.90	0.52
1900s	0.90	0.00	0.00	0.92	0.91	0.89	0.91	0.44	0.00	0.88	0.88

Table 10. The similarity scores for the most frequent word vectors in DE-EU related to “nationalismus,” grouped by decade.

In the German-language newspapers that were published in the United States (DE-CA), the concept of the nation is used in a radically different political context. Across the Atlantic, the nation became immediately connected with “justice,” “politics”, “party,” “government,” and “principles” (*Nation*, *Gerechtigkeit*, *Politik*, *Partei*, *Regierung*, *Grundsätze*), which illustrates how readily German migrants absorbed the constitutional context of their adopted nation. That adaptation may also explain the emergence of the term “republic” (*Republik*) in the 1850s, and “race” (*Rasse*) during the last years of the nineteenth century, in the discourse of nation. Similarly, the term “democracy” (*Demokratie*) became associated with heated party politics in the post-Jackson years of the 1840s, resulting in terms such as “slave owner” and “slavery question” (*Sklavenhalter*, *Sklavenfrage*) in the late 1840s, and “Republican” in the 1850s. The occurrence of references to silver and prohibition (*Silberfrage*, *Silberleute*, *Prohibition* and *Prohibitionisten*) in the vocabulary around democracy during the last decades of the century illustrates that German immigrants absorbed the key concerns of political populism, such as alcohol prohibition and resentment against the silver standard, which divided the rest of the United States [Wells 2015] [Kazin 2007]. For German immigrants in the United States, the term “folk” (*Volk*) is associated with “germanness” (*Deutschthum*) from the late 1850s and “fatherland” (*Vaterland*) from the 1880s. During the 1860s, references to “civil rights” and 43

“citizenship” (*Bürgerrecht*, *Bürgerthum*) are associated with people (*Volk*), and with “workers” (*Proletariat*) from the 1890s, which may be a reflection of the emerging radicalism in the United States of socialist and anarchist groups, inspired by European movements [Kazin 2012] [Foner 2014].

decade	partei	republik	politik	monarchie	einheit	demokratie	regierungsform	unabhängigkeit	dynastie	regierung	administration	civilisation	parle	menschheit	ar
1840s	0.62	0.00	0.81	0.00	0.00	0.00	0.00	0.45	0.00	0.45	0.30	0.00	0.00	0.00	0.
1850s	0.15	0.64	0.60	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.77	0.
1860s	0.79	0.87	0.30	0.77	0.61	0.30	0.15	0.15	0.15	0.15	0.46	0.59	0.00	0.00	0.
1870s	0.80	0.86	0.75	0.78	0.30	0.75	0.60	0.00	0.29	0.45	0.31	0.00	0.29	0.00	0.
1880s	0.79	0.83	0.75	0.79	0.76	0.30	0.30	0.60	0.59	0.00	0.00	0.15	0.30	0.00	0.
1890s	0.75	0.78	0.14	0.58	0.14	0.58	0.57	0.14	0.14	0.00	0.00	0.15	0.29	0.00	0.
1900s	0.72	0.71	0.00	0.00	0.67	0.73	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.

Table 11. Similarity scores for the most frequent word vectors in DE-CA related to “nation,” grouped by decade.

decade	bourgeoisie	partei	sozialdemokratie	Orthodoxie	socialdemokratie	demagogie	reaktion	wählerschaft	liberalismus	nation	fortschrittspartei	gewerkschaftsbewegu
1840s	0.85	0.83	0.62	0.40	0.41	0.61	0.40	0.00	0.00	0.81	0.20	0.00
1850s	0.84	0.84	0.88	0.81	0.82	0.83	0.16	0.65	0.48	0.16	0.66	0.00
1860s	0.87	0.83	0.88	0.83	0.85	0.49	0.33	0.32	0.49	0.00	0.00	0.48
1870s	0.82	0.40	0.86	0.82	0.84	0.00	0.82	0.00	0.00	0.00	0.00	0.82
1880s	0.85	0.83	0.62	0.40	0.41	0.61	0.40	0.00	0.00	0.81	0.02	0.00
1890s	0.84	0.84	0.88	0.81	0.82	0.83	0.16	0.65	0.48	0.16	0.66	0.00
1900s	0.87	0.83	0.88	0.83	0.85	0.49	0.33	0.32	0.49	0.00	0.00	0.48

Table 12. Similarity scores for the most frequent word vectors in DE-SBB related to “demokratie,” grouped by decade.

The case of Finland provides an interesting point of comparison with English and German datasets. Finland was originally a region of the Swedish Kingdom, annexed by the Russian Empire during the Napoleonic Wars. The Grand Duchy of Finland (1809–1917) was an autonomous part of Russia and can be considered as a predecessor of the Republic of Finland, founded in 1917. The fact that Finland is a bilingual country had consequences for Finnish nationalism. In the beginning of the nineteenth century, Finnish intellectuals and civil servants understood only Swedish, whereas the common people tended to use Finnish in their communication. This led to a situation where the early promoters of Finnish nationalism typically published in Swedish, and could not understand Finnish. However, towards the end of the nineteenth century, the Finnish language slowly replaced Swedish as the main language of the press, and during the peak of Finnish nationalism (Fennomania) , many Swedish-speaking families changed their first language to Finnish.[8] 44

Academic scholarship has emphasized the role of the German example in the early development of Finnish nationalism. In the beginning of the nineteenth century, folk (*folk* in Swedish or *kansa* in Finnish) was a term used in the idealized context of national Romanticism. Johann Gottfried von Herder and German Romanticism were important models for early Finnish nationalists like Adolf Ivar Arwidsson (1791–1858), active in the 1810s and 1820s. From the beginning, the press was the most important forum for Arwidsson’s nationalist activity. He founded the radical journal *Åbo Morgonblad*, which was suppressed by the Russian Tsar in 1821. After the short period of Finnish Romanticism in Turku, the intellectual center moved to Helsinki where the Hegelian philosophy became popular. When writing about the conceptual history of folk (*kansa*), Liikanen (2003) emphasizes the importance of the Fennomans and their Hegelian understanding of national spirit (*kansallishenki*, *nationalanda* from German *Volksgeist*). Inspired by Hegelian philosophy, the Finnish philosopher Johann Vilhelm Snellman (1806–81) introduced the concept of “spirit” (*Geist* in German, *anda* in Swedish, and *henki* in Finnish) into the Finnish nationalist discourse. Although our word vector models start from the 1840s, the Hegelian tradition is reflected in these German and Swedish word vector models. Both German and Swedish word vector models associated the Hegelian concept of “spirit” with “patriotism.” *Fosterlandskärlek* (“love for country”) is also strongly linked with the concept of “spirit” in Swedish results. Finally, the concept of “national spirit” (*kansallishenki*) is also associated with a mixture of nationalist terms and Hegelian technical terminology in the Finnish results (see Table 12). 45

decade	kansallistunto	kansallistunne	siwistys	kansallisuus	isänmaanrakkaus	itsetietoisuus	kansallinen	kulttuuri	yhteishenki	itsetunto	itsetajunta	ruotsalaisuus
1840s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1850s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1860s	0.23	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07
1870s	0.75	0.68	0.71	0.77	0.14	0.00	0.50	0.00	0.00	0.07	0.07	0.43
1880s	0.76	0.80	0.29	0.59	0.43	0.50	0.43	0.35	0.51	0.07	0.62	0.00
1890s	0.80	0.69	0.22	0.37	0.77	0.75	0.07	0.46	0.46	0.39	0.00	0.00
1900s	0.88	0.86	0.58	0.43	0.76	0.58	0.58	0.76	0.00	0.77	0.00	0.00

Table 13. Similarity scores for the most frequent word vectors in FI-NLF related to “kansallishenki” (national spirit), grouped by decade.

B. The ethnic roots of national identity

It has been customary to make a distinction between cultural and ethnic nationalism [Leerssen 2018] [Smith 2008] [Alter 1994]. For example, the late eighteenth-century philosopher Johann Gottfried von Herder is considered an important early theoretician of nationalism, although he in fact objected to the biological race theories of his time. When Herder emphasized the importance of local national cultures, his intention was to defend small states against the imperialism of the multinational empires of the time [Nisbet 1999]Nisbet 1999). However, after the development of positivism, scientism, and Darwinism, late nineteenth-century nationalism became more based on ideas of ethnicity, race, and the shared biological descent of national populations. This is also reflected in our results. In the Fenno-Swedish corpus, “nationality” (*nationalitet*) is associated with descent and origins. It seems that “nationality” and “ancestry” (*härkomst*) are also associated with “religious confession” (*trosbekännelse*). In contrast to these Swedish results, German word vector models for the concept of nationality (*Nationalität*) indicate a clear association between “race” and “ancestry.” Finally, the Finnish model is very different to the Fenno-Swedish and German results. The concepts in the word vector space linked with “nationality” (*kansallisuus*) refer to *siwistys* (cultivation, cf., *Bildung*), *kulttuuri* (culture), *kirjallisuus* (literature), *sanomakirjallisuus* (press), *kieli* (language), and *aate* (idea), which are related more to cultural nationalism than ethnic nationalism (see Table 13). 46

decade	siwistys	kansallinen	kansallishenki	kansallistunne	kulttuuri	kansallistunto	aate	kirjallisuus	kieli	edistyminen	sanomakirjallisuus	isänmaallisuus
1840s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1850s	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1860s	0.32	0.41	0.00	0.00	0.00	0.00	0.39	0.15	0.32	0.64	0.00	0.00
1870s	0.78	0.77	0.77	0.22	0.00	0.00	0.76	0.36	0.6	0.15	0.14	0.00
1880s	0.72	0.75	0.59	0.72	0.56	0.21	0.00	0.50	0.00	0.00	0.43	0.00
1890s	0.73	0.14	0.44	0.5	0.86	0.74	0.00	0.00	0.00	0.00	0.00	0.29
1900s	0.73	0.42	0.72	0.75	0.81	0.75	0.00	0.00	0.00	0.00	0.00	0.58

Table 14. Similarity scores for the most frequent word vectors in FI-NLF related to “kansallisuus” (nationality), grouped by decade.

In German newspapers (DE-SBB) of the same period, the ethnic roots of national identity are reflected by the close association between the term “folk” (*Volk*), “christianity” (*Christentum*) and “army” (*Heer*). In similar fashion, the term “nationality” (*Nationalität*) is used in the context of words that strongly suggest heritage, such as “descent” (*Abkunft*), “heredity” (*Abstammung*), “religion” (*Konfession*, Religion, Christian, Mission), and “language,” both “spoken” and “written,” “local” and “national” (*Schriftsprache*, *Landessprache*, *Muttersprache*, *Volkssprache*). The term “nation” is also closely related to “civilization” (*Civilisation*) and “mores” (*Gesittung*). 47

C. Citizens, peoples, and masses

As discussed, Koselleck emphasizes that the politicization of nation and folk in Germany towards the end of the nineteenth century was connected with the development of the labour movement and concepts related to anonymous multitudes of people, like mass or mob [Koselleck 1992b, 389]. In other words, the politicization of “folk” did not only happen against other nations, but also in relation to the inner power structures of each country. This temporal process, where the connotations of “folk” change substantially, is reflected very clearly in all of the national newspaper corpora analysed here. For example, in FI-NLF the term “the poor” (*köyhälistö*) becomes associated with “folk” (*kansa*) only from the 1890s onwards. In SE-NLF, “social class” (*samhällsklass*) appears in association with “folk” from the 1870s (Figure 1). In DE-SBB, references to “popular representation” (*Volksvertretung*) appear in the word vector space around “nation” in the 1870s and 1880s, and “workers” (*Arbeiter*) appears mainly in the 1880s. Similarly, terms such as “mob” (*Janhagel*) and “rabble” (*Pöbel*) appear in the vocabulary around “nation” with connotations such as “riot,” “fright,” “alarm,” and “rebellion” (*Aufruhr*, *Schrecken*, *Allarm*, *Aufstand*). The term “folk” (*Volk*) becomes semantically linked to “workers” (*Proletariat*) in the 1890s (Figure 1). 48

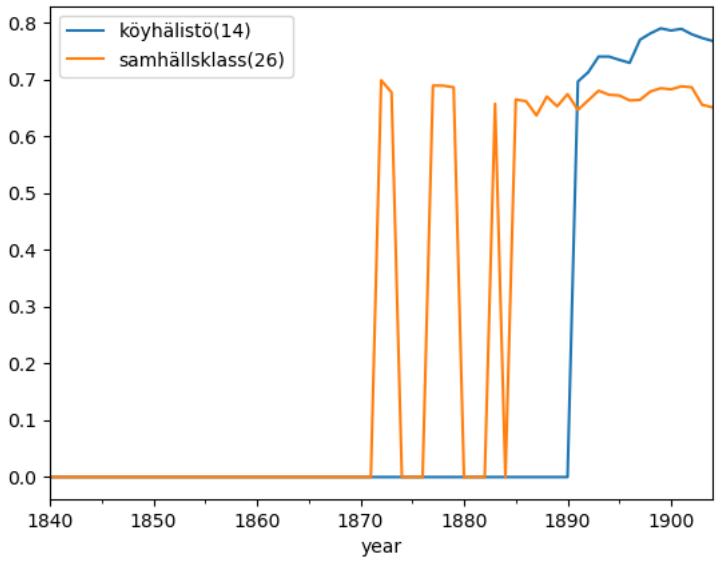

Figure 1. The rise of terms “köyhälistö” (the poor) and “samhällsklass” (social class) in FI-NLF in association with “kansa” (nation).

Whereas the labour movement sympathized with the poor and the proletariat, the political conservatives often made a distinction between “folk” – the ideal common people – and the disobedient “rabble.” In SE-NLF it seems that when “common people” are idealized and well-behaving, they are *folk*, but when they rebel, they become an anonymous rabble. In the DE-CA corpus, for instance, the concept of *Volk* (folk) is associated with *Schicksal* (destiny) in the beginning of the nineteenth century, whereas the association with Proletariat rises towards the end of the century (Figure 2). However, the etymological connections between Germanic languages can be misleading in this case. For example, English “people” is relatively neutral, whereas its German and Swedish homonyms *Pöbel* and *pöbeln* refer to a pejorative political concept meaning “mob” or “rabble,” i.e. a disorderly crowd of people [Koselleck 1992b, 143]. The division between idealized folk and dangerous rabble is reflected in Fenno-Swedish word vector models as well: the concept of “rabble” (*pöbeln*) is associated with “rebels” (*rebellerna*). In Finnish, *rahvas* (or *rahwas*) was first used as a neutral term to refer to “common people,” but acquired the pejorative meaning of “rabble” (*roskajoukko*) towards the end of the nineteenth century. 49

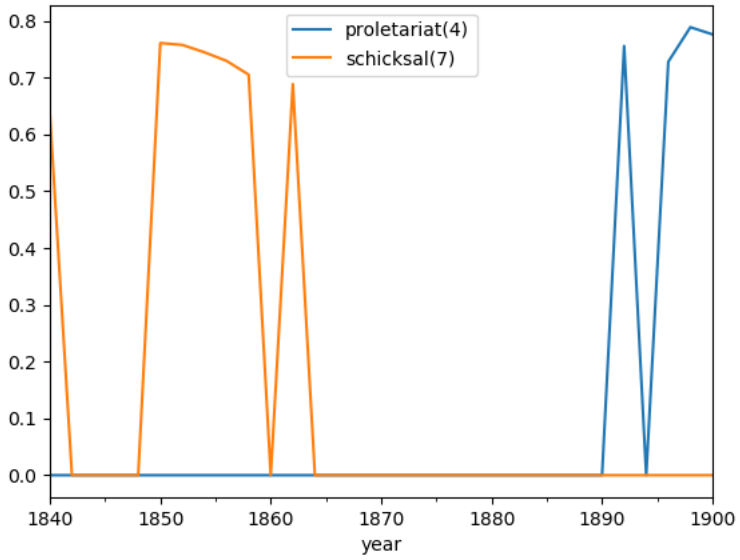

Figure 2. Frequency of "proletariat" and "schicksal" in DE-CA. The concept of *Volk* (folk) is associated with *Schicksal* (destiny) in the beginning of the nineteenth century, whereas the association with *Proletariat* rises towards the end of the century.

In TDA, too, the term "nation" becomes associated with the concept of "democracy" from the 1860s on, and the concept similarity score gradually increases until 1900 A contested aspect of national unity within the United Kingdom is suggested by the terms which appear in the context of the "people," where a consistently strong association with "countrymen" and "peasantry," as well as with "Englishmen" and "Irishmen," reflects the growing tensions around industrialization, urbanization, and home rule. References to Ireland typically evoke subaltern terms such as "oppressed," "colonists," and "loyalists" in the word vector space. The proximity between "people" and the term "agitators" in the models of the last decades of the century may reflect the rise of trade unions in the UK. 50

An interesting transatlantic context is presented by the words surrounding the concept of citizenship within the UK, as the term "inalienable" remains dominant until the first decade of the twentieth century. This seems a clear reference to the "inalienable rights" that were promised in the American Declaration of Independence, which may have informed the British debate about citizenship and national identity. The proximity of terms such as "privileges," "birthright," and "heritage" also illustrates the British dissension about the concept of nation and the right to be a citizen. If we superimpose the timeseries of "independence" and "citizenship" from the word vector model based on the query of "nationality" on the same plot, we can see that, during the beginning of the nineteenth century, "independence" is more frequently used in association with "nationality," whereas "citizenship" dominates at the end of the nineteenth century (Figure 3). 51

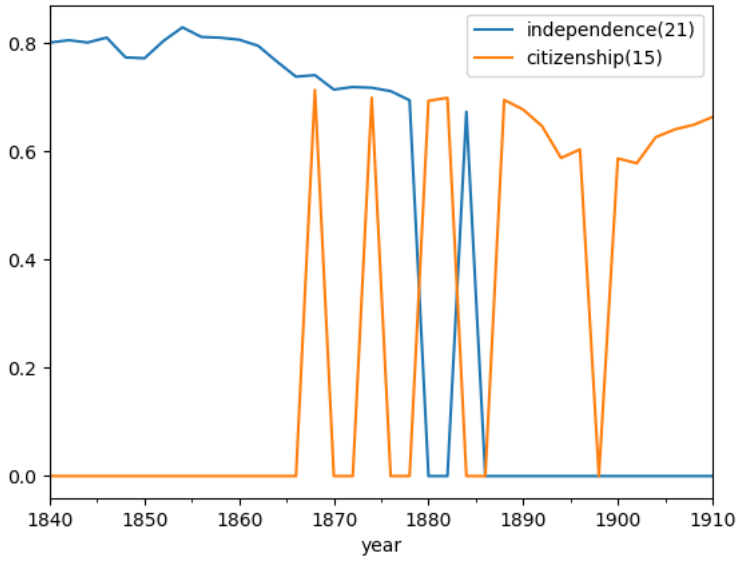

Figure 3. The time series of "independence" and "citizenship," plotted from the "nationality" vector space model, built on the dataset from TDA.

D. Comparing nations, linking vocabularies

The word vector models illustrate that the concept of "nation," in spite of its common origin, is expressed in divergent semantic contexts in the European countries examined here. The Finnish connotations of "nation" lean towards a cultural interpretation, suggested by terms like spirit and idea, culture, literature and language. German discourse points at a shared heritage that also suggests a cultural interpretation, as is visible in words such as Christianity and civilization. Although the English-language word vector models of TDA were not as clear, the British concept of "nation" seems to emphasize historical continuity as the basis of national identity, apparent in terms such as destiny and fatherland. British and German word vectors demonstrate a conceptual shift from a public to a more personal understanding of "nation," in which state and government are seen as intruders of the private sphere, suggested by references to disorganization, taxation and private enterprises. Germany and Finland share a tendency to understand the nation in essentialist and spiritual terms, where the term "folk" embodies the body politic. At the same time, all language sets show that the "people" evoked negative connotations by the vicinity of alarming terms such as "rabble," "mob," and "revolt". The transatlantic connections, lastly, were visible in the 52

British corpus (especially in the references to independence, citizenship and inalienable rights), but virtually absent in the German and Finnish newspapers. Although the word vector models of the German newspapers printed in the United States were not as conclusive, they suggested an interesting avenue of transnational comparison for future research, particularly with respect to the way the nation is understood in different geographical and historical contexts.

If, on the one hand, word vectors lack the precision of a thorough conceptual analysis such as the one conducted by *GGB*, on the other, they offer the advantage of extending the analysis on much larger datasets that, it can be argued, reflect the public discourse of a larger readership. The *longue durée* has shown, for instance, the relative instability of the connection between nation and monarchy in the UK. Due to the limitations of having to work with small datasets, these are subtle changes that the *GGB*, which is based on a limited set of key texts, could not have noticed. 53

More importantly, the parallel presentation of word vector models from newspapers in different languages opens up a multilingual comparison of public debates around common concepts. Although our methodology has had to rely on modern translations of seed words, detailed comparisons of domain knowledge still allow for a thorough interpretation of the research results. This offers many interesting heuristic possibilities to discover changes in newspaper discourse that may result from conceptual change. The current methodology does not allow us to track cross-border migration of concepts, but it does demonstrate commonalities and contrasts that can direct future research. 54

6. Case study II: Illness and disease

Compared to the concepts of “nation” and “nationhood,” “illness” and “disease” are different in nature: in all languages studied in this experiment, “illness” is not a migratory concept produced by modernization, but has profound local cultural roots that are based on older migratory ties. For example, the Finnish word *sairaus* (illness) derives from *sairas* (ill), which originates from the Germanic word *sairaz* which means sore or painful. The German word for ill, *krank*, also has proto-Germanic roots; it derives from *krangaz* or *krankaz*, which means crooked or weak [Kluge 1891]. Despite the regional etymological rootedness of the concept of illness, our experiment shows that, increasingly, new conceptual linkages and influences emerged during the nineteenth century. 55

In general, the word embeddings in Finnish, Swedish, German, and English refer to different horizons of meaning from three frames of reference. The vocabularies list words that 1. refer to symptoms or signs of being ill such as, for instance, fever or weakness; 2. identify particular diseases that raged in the nineteenth century, from isolated cases of the common cold to influenza and cholera pandemics; and 3. Show the consequences of becoming ill, like passing away, or becoming disabled through injury. Thus, the models retrieve vocabularies that conceptualize “illness” from distinct perspectives, ranging from symptoms and specific diseases to their consequences for human beings. Our experiment shows that illness cannot be traced to one stable ontological category, because these aspects are affected by shifting social and cultural contexts that result from specific economic, political, and social histories. 56

A. Towards the transnational semantics of “illness”

In the first experiment, we search for the 15 most related terms for the concept term “illness” in the word embeddings of all corpora for the different time spans. For all different corpora, we encounter spelling variations that have resulted from OCR problems. For example, for the German word *Krankheit* (illness) we also retrieve terms like *Kraukheit* or *Krankhett*, and the Swedish word *sjukdom* (illness) was retrieved as *fjukdom* and *fjuldom*. To enable a better view of the results, we merge these spelling variations for all terms and compute their average similarity score. 57

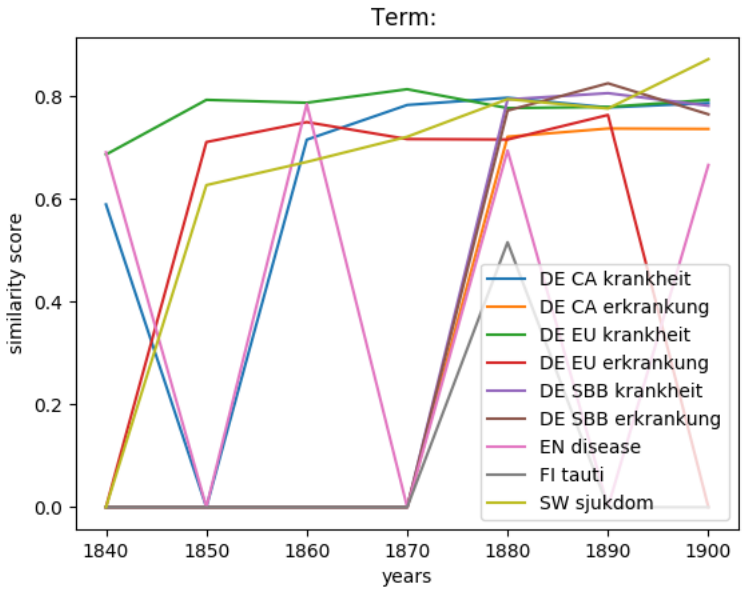

Figure 4. Combined similarity scores for synonyms of “illness/disease” in all datasets.

	DE-CA	DE-CA	DE-EU	DE-EU	DE-SBB	DE-SBB	TDA	FI-NLF	SW-FNL	SW-NLF
decade	krankheit	erkrankung	krankheit	erkrankung	krankheit	erkrankung	disease	tauti	sjukdom	ohälsa
1840s	0.59	0.00	0.69	0.00	-	-	0.69	0.00	0.00	0.00
1850s	0.00	0.00	0.80	0.71	-	-	-	0.00	0.62	0.00
1860s	0.71	0.00	0.78	0.75	-	-	0.78	0.00	0.67	0.00
1870s	0.78	0.00	0.81	0.72	-	-	-	0.00	0.72	0.00
1880s	0.80	0.72	0.78	0.72	0.79	0.77	0.69	0.5154	0.79	0.00
1890s	0.78	0.73	0.78	0.76	0.81	0.82	-	0.00	0.78	0.00
1900s	0.7863	0.74	0.79	0.00	0.78	0.76	0.67	0.00	0.87	0.70

Table 15. Similarity scores for the term illness/disease for all datasets. A similarity score of 0.00 indicates that the term is not similar or cannot be found in the corpus. A “-” indicates that the corpus does not cover these years, and thus a model could not be computed.

In addition, we retrieve synonyms of the term illness such as sickness and disease, which validate the reliability of the semantic models. We focus on the most similar terms that appear in at least four of the different models with a continuous time span, confirming their significance in the corpora. This enables us to identify discrepancies across the different languages and corpora. 58

B. The conceptualization of pandemics

The modern era has been characterized by pandemics [Caduff 2015]. The word pandemic itself is a Greek loanword that refers to diseases that affect all (*pan*) people (*demos*). In the nineteenth century, new forms of transport and mobility enabled the rapid spread of diseases. Cholera and influenza in particular raged on a global scale [Hamlin 2009] [Honigsbaum 2014]. This effect of globalization is visible in our results. Since the corpora cover slightly different timespans, they do not offer any comprehensive picture of how pandemics emerged and developed conceptually. 59

However, the results refer to different regional perceptions of pandemics. In themselves, “epidemic” and “pandemic” are ancient words that have been used in many European languages since at least the seventeenth century [OED 2019]. In SE-NLF, the Swedish word “epidemi” appears eleven times from 1858–68 onwards. At the same time, it also shows up in DE-CA, first as “Seuche,” which was retrieved as similar to “Krankheit,” 16 times. The loanword “Epidemie” appears in the results slightly later, in 1864–74. In DE-CA, the term “cholera” is retrieved in 1858–68, obviously as a result of the so-called third cholera pandemic of 1846–60. In Europe, cholera raged particularly violently during the Crimean War of the 1850s, but the pandemic soon reached the Americas in the late 1850s, which accounts for discussions in the German immigrant press that label cholera *the* disease. While the closeness of real pandemics to their conceptual reflections is obvious, the word embeddings indicate that pandemics also caused conceptual changes.

In addition to cholera, influenza constitutes a further interesting example that gained weight in the nineteenth century. The so-called “Russian flu” pandemic was particularly lethal in 1889–90. In the Swedish corpus, too, “influenza” appears in 1888–98. In FI-NLF “influenza” appears for the first time as “influentsa” in 1884–94, not as a synonym for “sairaus” (see above) but for “tauti,” another Finnish term for illness. As a concept, “influenza” went almost as viral as the disease itself. There were, of course, references to influenza before the 1880s, but the international loanword (and its variations) was not used, according to our empirical data. Semantically, influenza hides behind different words, such as flu, or breast or lung disease. 60

C. Nervous diseases and the social boundaries of illness

At the beginning of the nineteenth century, the understanding of the mechanisms of nervous diseases was mostly pathological. However, in the course of the century, modern psychology developed a more internal view of the human mind [SChultz and Schultz 2012]. In the newspaper corpora, nervous diseases appear increasingly towards the end of the nineteenth century. At the turn of the twentieth century there were particularly vivid discussions about neurasthenia. This term for weakness of the nerves was coined in the 1820s, but became popular from the 1860s onwards, especially after the publication of George Miller Baird’s article “Neurasthenia, or nervous exhaustion,” in *The Boston Medical and Surgical Journal* in 1869 [Beard 1869, 217–21]. Neurasthenia was often seen as a disease caused by the accelerating rhythms of modern culture [Uimonen 2000] [Salmi 2013]. Nervous diseases start to appear in our models in the 1860s, and remain an essential conceptual aspect of illness going forward. 61

	Mental illness	neuropathy			
	DE-CA	DE-SBB	SW	DE-SBB	SW
decade	geistesstörung	geisteskrankheit	sinnessjukdom	nervenkrankheit	nervsjukdom
1840-1850	0.00	-	0	-	0.00
1842-1852	0.00	-	0	-	0.00
1844-1854	0.00	-	0	-	0.00
1846-1856	0.00	-	0	-	0.00
1848-1858	0.00	-	0	-	0.00
1850-1860	0.00	-	0	-	0.00
1852-1862	0.00	-	0	-	0.00
1854-1864	0.00	-	0	-	0.00
1856-1866	0.00	-	0	-	0.00
1858-1868	0.00	-	0	-	0.00
1860-1870	0.00	-	0	-	0.00
1862-1872	0.00	-	0.58	-	0.00
1864-1874	0.67	-	0.62	-	0.00
1866-1876	0.00	-	0.00	-	0.00
1868-1878	0.00	-	0.61	-	0.00
1870-1880	0.00	-	0.00	-	0.00
1872-1882	0.00	0.72	0.61	0.00	0.00
1874-1884	0.00	0.72	0.67	0.00	0.00
1876-1886	0.00	0.72	0.66	0.00	0.00
1878-1888	0.00	0.73	0.69	0.00	0.00
1880-1890	0.00	0.73	0.67	0.00	0.00
1882-1892	0.00	0.75	0.66	0.00	0.00
1884-1894	0.00	0.72	0.67	0.00	0.00
1886-1896	0.00	0.73	0.68	0.00	0.00
1888-1898	0.00	0.00	0.67	0.74	0.00
1890-1900	0.00	0.00	0.70	0.72	0.00
1892-1902	0.00	0.00	0.72	0.80	0.00
1894-1904	0.00	0.00	0.68	0.81	0.00
1896-1906	0.00	0.73	0.68	0.77	0.00
1898-1908	0.00	0.00	0.68	0.76	0.66
1900-1910	0.00	0.71	0.66	0.76	0.68
1902-1912	-	0.00	0.66	0.72	0.00
1904-1914	-	-	0.70	-	0.00
1908-1918	-	-	-	-	-

Table 16. Similarity scores between “illness” and the terms “mental disease” and “nervous disease.” While we observe “mental disease” in DE-EU, DE-SBB and SE-NLF, we retrieve “nerve disease” only in DE-SBB and SE-NLF.

The same observation applies to mental diseases. *Nervenkrankheit* (nervous disease) is retrieved in DE-SBB first in 1882–892, and four times thereafter. In the DE-EU, it comes up in 1888–98, in total eight times during the whole timespan, and in the SE-NLF in 1898–1908 three times in total as *nervsjukdom.* Mental health issues also appeared through a plethora of similar words: in Swedish as *sinnessjukdom* (mental disease), as many as forty times after 1862–72; in DE-EU as *Geisteskrankheit* ten times after 1872–82; and in the DE-SBB four times during the same period. This is not to argue that nervous or mental problems had not been diagnosed before, but conceptually they became more eminent as defining principles for illness. 62

There are clear differences between the corpora, however. The examples from the Swedish and German material refer to different timescales in the discourses and articulations on nervous and mental issues. It also appears that *Nervenkrankheit* is not being retrieved in DE-CA or in FI-NLF. Although the models do not make the reasons for this absence clear, these results may be explained by the different audiences of the Finnish-language and German immigrant press. The former was addressing a rural population, whereas neurasthenia was mainly discussed in a middle-class context. This is also supported by the fact that the Finnish-language models yielded many words that refer to animal and plant diseases, which were important for the rural readership of the Finnish-language press. These findings support the conclusion that “illness” fuelled conceptual migrations across borders, while at the same time reflecting social boundaries for the conceptualization of diseases. However, the word vector models do not allow for further exploration of the context of these concepts, which means that the semantic interaction between human and natural domains in a period when Darwinian evolution theory was debated worldwide would require more research [Hawkins1997] [Bowler 2003, 224–324] 63

D. Debating causes, symptoms, and consequences

The etiological standpoint, coined by Robert Koch in the 1870s, refers to the assumption that diseases are best controlled and understood by means of their causes. Since this theory has dominated medical discourse for the past two centuries, much of contemporary medical practice concentrates on identifying specific causes of disease. A disease, however, is regularly defined by specifying its nature. To study the representation of illness at community-level in the nineteenth century, our results show that this identity can be expressed by more than causes. Our word embeddings reflect this by showing vocabularies which conceptualize illness from different perspectives such as symptoms, their consequences, synonymous terms, and treatment. By breaking up illness into categories (e.g. treatment, consequences, and symptoms) we can demonstrate its multimodal nature as a concept without fixed entities. Moreover, the word “illness” is used in different senses or meanings. The aim of this use case is to investigate whether the dominant sense of illness changed over time. The findings support the claim that “there is no single way of defining, interpreting, experiencing or managing disease” [Jackson 2017, 6]. Our results show that, in contrast to medical discourse, the nineteenth-century press did not concentrate on causes and aetiology of diseases, but rather focused on their human consequences. 64

To investigate the change in word sense, we first translate all similar terms for “illness” into nineteenth-century English. Then, two annotators assign categories to each term. After a close reading of the vocabulary, we define the subcategories listed in Table 16: 65

Category	Abbreviation	Example
treatment	T	surgery
consequences	C	death
synonymy	M	disease
symptoms	S	pain
specific disease	D	cholera
other	?	Words of other categories and words of other part-of-speech

Table 17. Subdivided categories of “illness,” their abbreviation, and textual examples.

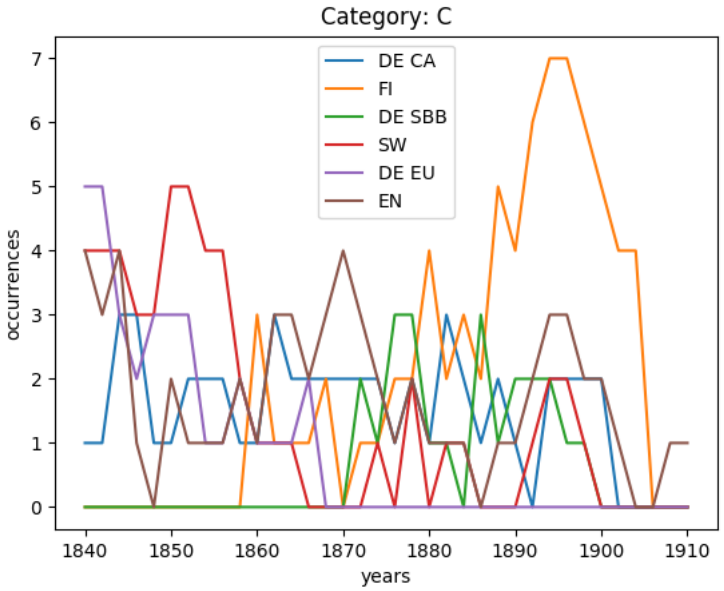

Figure 5. The consequences of diseases. The shifting temporal and corpus-specific patterns in all datasets of “disease,” in terms of its retrieved vocabulary referring to consequences of diseases. This is based on our categorization of “illness,” category C (see Table 16).

We assigned words such as “death,” “crisis,” or “incurable” to the consequences category, “sore throat,” “tiredness,” or “vomiting” to the symptoms category, “ill,” or “sickness” to the synonymy category, “cholera,” “influenza,” or “mental disease” to a category for specific diseases, “surgery,” or “inoculation” to treatment, and words that were not context-related to the other category. 66

The annotation and comparison of the different subcategories highlights that the majority of generated words are synonyms and consequences of diseases. These categories are dominant in all six corpora. As Figure 5 illustrates, analysis of word embeddings in diverse newspaper corpora suggests the discussion of consequences of diseases rarely respects local, regional, or national borders. This leads to the conclusion that the press constructed illness as a concept with an emphasis on its consequences. This emphasis is supported by mainly negative characteristics such as, for instance, “epidemic,” “poverty,” “died”, and “suffering.” Thus, diseases are represented as agents of suffering, misery, and death, which have a strong negative impact on human lives. 67

However, shifting temporal and corpus-specific patterns can be observed, as illustrated by Figure 5. Whereas DE-SBB only textualizes consequences from 1870 onwards, no words categorized as consequences seem to occur in DE-EU after 1878. Moreover, words that refer to consequences are widely discussed in FI-NLF and SE-NLF throughout the entire time span. 68

Analysing the different categories in which the concept illness is discussed in newspapers allows us to trace the fluctuating meanings of disease and illness as a superordinate category. The overview of subcategories also helps us to consider the role of the press in defining diseases. The fact that there was a strong emphasis on the consequences of illness may be explained by the dual nature of newspaper discourse. On the one hand, the press constructed fear of diseases by describing the disastrous consequences of illness, becoming a platform for emotional expression. On the other, the press expressed social observations and shared them with a wider audience. Pandemics such as influenza and cholera were strongly present in the nineteenth-century public sphere, and papers informed their readers of their impact on the life of an individual as well as on society. 69

The combination of synchronic and diachronic analysis of word vector models allows us to trace the development of an abstract concept such as illness over time in newspaper corpora in different languages. We can discover how specific diseases emerged in public discourse in Europe and the United States at different times in history, such as the epidemics of cholera and influenza that affected these parts of the world. More interestingly, word vectors are an effective instrument to trace the vocabularies with which national publics discussed – and constructed – diseases, and their supposed causes and consequences. Our study shows the influence of public discourse, newspaper publishing in particular, on definitions of illness and the multidimensional network of relations between symptoms, specific diseases, and their consequences to construct the concept of illness. Although this case study was limited to newspaper collections from a small number of nations, the differences and silences in discourse, and the changing vocabularies that emerge from the word vector models, offer interesting heuristic starting points for a close reading of these corpora within their proper historical context. 70

7. Discussion

Although the two use cases demonstrate the ways in which word embeddings can be used for the transnational comparison of concepts and vocabularies over time, this experiment also illustrates the following methodological and technological challenges in using multilingual historical newspaper datasets for the construction of word vector models. 71

OCR issues: One of the major issues we face with historic newspaper data is the OCR quality. Some newspaper pages simply cannot be converted due to low scan quality, and further issues arise with spelling variations. For example, the German word *Volk* appears also as “Bolk” and “Volt.” To solve this issue, we merged variations and used their average similarity in our study.

Different language properties: For the computation of the models, we perform the same pre-processing by using standard tokenization methods. However, because we are using tokenization that mostly splits words by white spaces, multiword expressions composed of two words separated by white space are also split. While this is not an issue for the Germanic languages, where multiword expressions are often close compounds, we lose these terms for the English corpus. This was particularly evident for the illness case study, where specific diseases are often represented as multiword units, leading to entirely different results. In this article, however, we cannot address the task of computing paraphrase embeddings or detecting multiword expressions, and will leave this to future work.

Corpus differences: We use dense vector space models to compare the shift of concepts across different languages and corpora. However, besides language differences, the corpora we use to compute the models are very diverse in size and composition, and represented different time periods. Comparisons between concepts across corpora is only possible for overlapping time periods. While we have the entire time span of newspapers in TDA, we have no, or only few, newspapers from 1840 to 1855 for most other corpora. Models for semantic similarity tend to be more stable if the size of the corpora used for the computations is large [Riedl and Bieman 2013] [Altszyler et al. 2017]. We observe the general trend that the number of newspapers increased with time. For the first decades of the nineteenth century, we often have only a few texts, making the models less reliable. In addition, the individual newspaper datasets differ from one another. While TDA only contains newspaper issues from a single source, we use German-language newspapers of several European countries in the German Europeana corpus. In addition, newspapers are always targeted to a specific audience. The Finnish corpus is mostly comprised of newspapers from rural regions, and thus represents specific concerns (e.g. about farming) and mostly does not cover urban trends.

Parameters of word embeddings: The embeddings depend on the parameters used for the computation, as well as the number of similar terms that are extracted. For the computation of the embeddings, we relied on standard parameters as provided by the ShiCo tool (see footnote 2).

Conceptual dissimilarities: This article studied conceptual change in two different sets of concepts, those of collective nationhood, and those of the personal circumstances of health and illness. As we have argued, both are cultural constructions that change over time, and both are embedded in their specific cultural context. Koselleck et al also draw attention to the frequent use of metaphors of illness, health, and the body to express concerns about the state of the nation as body politic [Koselleck et al. 2006, 163–4, 205]. However, the word vector models shows no overlap between the two semantic domains within the corpora that were used. The connection between the two conceptual sets can only be established on the basis of historical domain knowledge which takes the specific historical context into account.

8. Conclusion

Historical newspapers allow us to study how everyday concepts have been used in public discourse. This article discusses how word vector models can help us to trace how such concepts were articulated in shifting vocabularies as they moved over time and space. This addresses the methodological question: how can computational methods be applied to broaden the scope of conceptual history? In order to test the usability of word vector models, we address two urgent academic questions within the field of conceptual history. The first is to what extent concepts are stable entities that are expressed in changing vocabularies over large periods of time, as Lovejoy (1933) famously suggested, or whether changing vocabularies should be understood as an indication of conceptual change, as other practitioners of conceptual history argue. We test this by concentrating on a historical time frame, the period between 1840 and 1914, in which the western world experienced the rapid changes of modernization and globalization and became interconnected through the new mass media of newspapers. The second, perhaps more complex, question is how we can use this computational methodology to trace how concepts change as they migrate over geographical and linguistic borders. For our dataset, we use digitized historical newspaper corpora in five different languages. As use cases, we select two radically different concepts that have a global presence, the collective identity of the “nation” and the deeply personal experience of “illness.” 72

This international and interdisciplinary research project, in which researchers from four academic institutions in Europe collaborate, illustrates the possibilities and challenges of using computational methods to analyse conceptual change over longer periods of time in different cultural and national contexts. We can formulate a number of promising findings: 73

Newspaper corpora. The computational tool ShiCo, which has been developed to trace shifting concepts over time by constructing series of overlapping word vector models, can effectively be applied to historical newspaper repositories of different provenance, metadata structure, and OCR quality, and, most interestingly of all, written in different languages. The resulting vocabularies were consistent enough to allow meaningful scholarly interpretation. Even if newspapers are commercial and professional media enterprises with their own ideological agendas, they can only survive if they reflect ongoing debates in society. The outcomes of this research project confirm that these combined newspapers collections are rich and promising sources for a computational approach to conceptual history, and may offer us a novel entry point into the historical public sphere.

Digital conceptual history. Our computational approach to conceptual history validates the application of a quantitative and big data method to this discipline. In comparison to traditional approaches to conceptual history, which tend to use limited textual corpora mostly produced by scholarly communities, this computational approach enables us to draw on the big data repositories of historical newspapers which reflect a much broader public discourse.

Time. Word vector models prove to be a convincing and promising tool to show how different vocabularies share the same semantic space over time. Although we define the term “concept” rather pragmatically, the word models show a consistent configuration that changed over time, sometimes gradually, sometimes rapidly, introducing new words or terms as a result of historical changes. The domain knowledge of the authors enables us to offer meaningful interpretations of these semantic changes. This offers a new way to trace how collective concepts such a nation, national identity, the people, and the more personal discussions of causes, symptoms and results of illness were articulated over a longer period of time, even spanning eight decades.

Space. The comparison of the word models created by applying the ShiCo software to separate newspaper repositories in different languages allows researchers to test the promise of comparative and transcultural conceptual history. Even if this method relies on translation of key terms and vocabularies in English, and on manual comparison based on historical domain knowledge, the results are significant and allow us to situate concepts such as nation and illness in its historical contexts.

Distant reading. As Moretti quipped, “ambition is now directly proportional to the distance from the text: the more ambitious the project, the greater must the distance be” [Moretti 2013, 48]. Indeed, word vector models seem to allow for an extreme form of distant reading of textual corpora which takes researchers far from the contextual content of the newspapers. But even as the text itself disappears out of sight, larger patterns emerge that carry heuristic value. By identifying trends and discontinuities in the way concepts are articulated in shifting vocabularies as they move over time and space, this form of distant reading can help us to understand how concepts are discussed in newspapers in various languages. Comparing word vector models based on newspapers from different national collections offers us a glimpse at the circulation of knowledge in the western world.

The outcomes of this article also suggest a number of avenues for future research: 74

1. Although word vectors models offer quantitative data about the words that appear in the semantic space around search queries or seed words, the comparison between different language corpora still relies on qualitative interpretation. A more robust method may be found in cross-mapping word vector models to represent the translation of concepts and terms [Jansen 2017] [Luong et al. 2015] [Mikolov et al. 2013b]. The structure of the data and models currently does not allow for such computational application of translation vectors.
2. It would be valuable to divide the national newspaper corpora into different sub-collections of newspapers according to their political, religious or regional affiliation. Such segmentation would allow for comparisons of, for instance, Catholic or socialist newspapers from different linguistic and national backgrounds, or could offer indications of the audiences that are receptive to international influences or ideological affiliation. Segmentation would require server space and computer power that was not available for this project.
3. Another approach would be to develop a robust computational method to determine in which specific discursive contexts certain vocabularies and concepts were used (e.g. in which sections of the newspapers can references to national identity be found in newspapers from a particular period, and where were issues of health and illness discussed? Can the discursive context of these references be established?). This would necessitate combining the creation of word vectors of sufficient volume with automated segmentation of newspapers in departments, genres, or other sections that semantically belong to each other.
4. A more ambitious way to trace cross-national migration of ideas would be to implement algorithms that identify text reuse. Although this has been successfully attempted in monolingual corpora [Smith et al. 2015] [Cordell 2015], text reuse in multilingual corpora is still in development. If combined with the heuristic potential of word vector models, text reuse could bring us closer to the scalable, zoomable and explorative readings that are the meeting ground between the big-data ambitions of distant reading and the semantic precision of close reading.

Acknowledgements

We would like to thank prof. Marc Priewe of the University of Stuttgart for hosting the project conference from which this publication originated and for his feedback on the concept version 75

Notes

[1] “Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840–1914.” DOI 10.17605/OSF.IO/WA94S. Available at: osf.io/wa94s.

[2] For querying similar terms, we use the following ShiCo parameters: non-adaptive search, maximal terms: 15, related terms: 20 and minimum similarity: 0.1. The other parameters were left on their default settings: word boost = 1, boost method = sum similarity, words per year = 5, weighing function = gaussian, etc. Documentation of these parameters and the ShiCo software is available at https://github.com/NLeSC/ShiCo.

[3] Software and scripts developed for this project are available on https://github.com/OceanicExchanges/wp5.

[4] For the TDA corpus models were created for the period 1840–1920 for reasons of repurposability

[5] https://www.europeana.eu/portal/en

[6] https://chroniclingamerica.loc.gov/

[7] The search terms were: citizenship, civilisation, civilization, country, democracy, destiny, empire (realm), fatherland, masses, mob, nation, nationalism, nationality, people(s), rabble, spirit of nation, state.

[8] During the 1860s language strife in Finland, advocates of the Finnish language were called the Fennomans, and they founded the Finnish Party. A significant number of the Fennomans originated in the mostly Swedish-speaking upper classes who Finnicized their Swedish family names and promoted Finnish culture and language.

Works Cited

Alter 1994 Alter, Peter. 1994. *Nationalism*. 2nd ed. New York: Edward Arnold.

Altszyler et al. 2017 Altszyler, Edgar, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. “The Interpretation of Dream Meaning: Resolving Ambiguity Using Latent Semantic Analysis in a Small Corpus of Text.” *Consciousness and Cognition* 56 (November): 178–87. https://doi.org/10.1016/j.concog.2017.09.004.

Anderson 2006 Anderson, Benedict R. O’G. 2006. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Rev. ed. London: Verso.

Andrews 2014 Andrews, Ann. 2014. *Newspapers and Newsmakers: The Dublin Nationalist Press in the Mid-Nineteenth Century*. Liverpool: Liverpool University Press.

Baroni et al. 2014 Baroni, Marco, Dinu Georginana, and Germán Kruszewski. 2014. “Don’t Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors.” In: Proceedings of ACL:171–81. East Stroudsburg PA. http://anthology.aclweb.org/P/P14/P14-1023.xhtml.

Beals and Bell 2020 Beals, Melodee, and Emily Bell. 2020. “The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges,” Figshare, 19 january 2020. https://doi.org/10.6084/M9.FIGSHARE.11560059.

Beard 1869 Beard, George Miller. 1869. “Neurasthenia, or Nervous Exhaustion.” *Boston Medical and Surgical Journal* (April): 217–21.

Betti and van den Berg 2014 Betti, Arianna, and Hein van den Berg. 2014. “Modelling the History of Ideas.” *British Journal for the History of Philosophy* 22 (4): 812–35. https://doi.org/10.1080/09608788.2014.949217.

Billig 1995 Billig, Michael. 1995. *Banal Nationalism*. Thousand Oaks, CA: Sage.

Bowler 2003 Bowler, Peter J. 2003. *Evolution: The History of an Idea*. 3rd ed. Berkeley: University of California Press.

Broersma and Harbers 2018 Broersma, Marcel, and Frank Harbers. 2018. “Exploring Machine Learning to Study the Long-Term Transformation of News.” *Digital Journalism* 6 (9): 1150–64. https://doi.org/10.1080/21670811.2018.1513337.

Brunner et al. 1972 Brunner, Otto, Werner Conze, and Reinhart Koselleck, eds. 1972. *Geschichtliche Grundbegriffe: Historisches Lexikon Zur Politisch-Sozialen Sprache in Deutschland*. 8 vols. Stuttgart: E. Klett.

Caduff 2015 Caduff, Carlo. 2015. *The Pandemic Perhaps: Dramatic Events in a Public Culture of Danger*. Oakland, CA: University of California Press.

Cordell 2015 Cordell, Ryan. (2015) “Reprinting, Circulation, and the Network Author in Antebellum Newspapers.” *American Literary History*, 27(3), pp. 417-445.

De Bolla 2013 De Bolla, Peter. 2013. *The Architecture of Concepts: The Historical Formation of Human Rights*. New York: Fordham University Press.

Douglas 1999 Douglas, George H. 1999. *The Golden Age of the Newspaper*. Westport, CT.: Greenwood Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=497022.

Eijnatten et al. 2014 Eijnatten, Joris van, Toine Pieters, and Jaap Verheul. 2014. “Big Data for Global History: The Transformative Promise of Digital Humanities.” *Low Countries Historical Review* 128 (4): 55–77.

Firth 1957 Firth, John Rupert, ed. 1957. *Studies in Linguistic Analysis*. Oxford: Blackwell.

Foner 2014 Foner, Eric. 2014. *Give Me Liberty!: An American History*. Fourth edition. New York: W.W. Norton & Company.

Gavin 2015 Gavin, Michael. 2015. “The Arithmetic of Concepts: A Response to Peter de Bolla,” in: *Modelling Literary History* (blog). modelingliteraryhistory.org/2015/09/18/the-arithmetic-of-concepts-a-response-to-peter-de-bolla.

Gellner 1997 Gellner, Ernest. 1997. *Nationalism*. New York University Press.

Gellner 2007 — — — . 2007. *Nations and Nationalism*. 2. ed., Malden, MA: Blackwell Publ.

Ginneken 1998 Ginneken, Jaap van. 1998. *Understanding Global News: A Critical Introduction*. Thousand Oaks, CA: Sage.

Habermas 1991 Habermas, Jürgen. 1991. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Cambridge, MA: MIT Press.

Hamilton et al. 2016 Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 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)*, 1489–1501. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141.

Hamlin 2009 Hamlin, Christopher. 2009. *Cholera: The Biography*. Oxford: Oxford University Press.

Harras 2000 Harras, Gisela. 2000. “Concepts in Linguistics – Concepts in Natural Language.” In *Conceptual Structures: Logical, Linguistic, and Computational Issues*, edited by Bernhard Ganter and Guy W. Mineau, 1867:13–26. Berlin: Springer. https://doi.org/10.1007/10722280_2.

Harris 1954 Harris, Zellig S. 1954. “Distributional Structure.” *Word* 10 (2–3): 146–62. https://doi.org/10.1080/00437956.1954.11659520.

Hawkins1997 Hawkins, Mike. 1997. *Social Darwinism in European and American Thought, 1860-1945: Nature as Model and Nature as Threat*. Cambridge: Cambridge University Press.

Hobsbawm 2012 Hobsbawm, Eric J. 2012. *Nations and Nationalism since 1780: Programme, Myth, Reality*. Second edition. Cambridge: Cambridge University Press.

Hobsbawm and Ranger 2010 Hobsbawm, Eric J., and Terence O. Ranger, eds. 2010. *The Invention of Tradition*. Cambridge: Cambridge University Press.

Honigsbaum 2014 Honigsbaum, Mark. 2014. *A History of the Great Influenza Pandemics: Death, Panic and Hysteria, 1830–1920*. New York: I.B. Tauris.

Jackson 2017 Jackson, Mark, ed. 2017. *The Routledge History of Disease*. London: Routledge.

Jackson and Moulinier 2007 Jackson, Peter, and Isabelle Moulinier Moulinier. 2007. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. nd revised ed. Amsterdam: John Benjamins.

Jansen 2017 Jansen, Stefan. 2017. “Word and Phrase Translation with Word2vec.” *ArXiv:1705.03127 [Cs]*, May. http://arxiv.org/abs/1705.03127.

Kazin 2007 Kazin, Michael. 2007. *A Godly Hero: The Life of William Jennings Bryan*. New York: Anchor Books.

Kazin 2012 — — — . 2012. *American Dreamers: How the Left Changed a Nation*. New York: Vintage Books.

Kenter 2013 Kenter, Tom. 2013. “Filtering Documents over Time for Evolving Topics.” *Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013)*.

Kenter et al. 2015 Kenter, Tom, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 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, 1191–1200. Melbourne, Australia: ACM Press. https://doi.org/10.1145/2806416.2806474.

Kluge 1891 Kluge, Friedrich. 1891. *Etymological Dictionary of the German Language*. London: George Bell & Sons.

Koselleck 1992a Koselleck, Reinhart. 1992a. “Einleitung: Volk, Nation, Nationalismus, Masse.” In: *Geschichtliche Grundbegriffe: Historisches Lexikon Zur Politisch-Sozialen Sprache in Deutschland*, edited by Otto Brunner, Werner Conze, and Reinhart Koselleck, 7:141–51. Stuttgart: E. Klett.

Koselleck 1992b — — — . 1992b. “Lexikalischer Rückblick.” In *Geschichtliche Grundbegriffe: Historisches Lexikon zur Politisch-Sozialen Sprache in Deutschland*, edited by Otto Brunner, Werner Conze, and Reinhart Koselleck, 7:380–89. Stuttgart: Klett-Cotta.

Koselleck 2002 — — — . 2002. *The Practice of Conceptual History: Timing History, Spacing Concepts*. Translated by Todd Samuel Presner. Stanford, CA: Stanford University Press.

Koselleck 2004 — — — . 2004. *Futures Past: On the Semantics of Historical Time*. New York: Columbia University Press.

Koselleck et al. 2006 Koselleck, Reinhart, Ulrike Spree, Willibald Steinmetz, and Carsten Dutt. 2006. *Begriffsgeschichten: Studien zur Semantik und Pragmatik der politischen und sozialen Sprache*. Frankfurt am Main: Suhrkamp Verlag.

Kunczik 1997 Kunczik, Michael. 1997. *Images of Nations and International Public Relations*. LEA’s Communication Series. Mahwah, N.J: Erlbaum.

Kuukkanen 2008 Kuukkanen, Jouni-Matti. 2008. “Making Sense of Conceptual Change.” *History and Theory* 47 (3): 351–72. https://doi.org/10.1111/j.1468-2303.2008.00459.x.

Landauer et al. 1997 Landauer, Thomas K., and Susan T. Dumais. 1997. “A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge.” *Psychological Review* 104 (2): 211–40. https://doi.org/10.1037/0033-295X.104.2.211.

Leerssen 2018 Leerssen, Joep. 2018. *National Thought in Europe: A Cultural History*. Amsterdam: Amsterdam University Press.

Lenci 2018 Lenci, Alessandro. 2018. “Distributional Models of Word Meaning.” *Annual Review of Linguistics* 4 (1): 151–71. https://doi.org/10.1146/annurev-linguistics-030514-125254.

Lovejoy 1933 Lovejoy, Arthur O. 1933. *The Great Chain of Being: A Study of the History of an Idea*. Cambridge, MA: Harvard University Press. http://site.ebrary.com/id/10314249.

Lund and Burgess 1996 Lund, Kevin, and Curt Burgess. 1996. “Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence.” *Behavior Research Methods, Instruments, & Computers* 28 (2): 203–8. https://doi.org/10.3758/BF03204766.

Luong et al. 2015 Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. 2015. “Bilingual Word Representations with Monolingual Quality in Mind.” *NAACL Workshop on Vector Space Modeling for NLP*, 151–59.

Margolis and Laurence 1999 Margolis, Eric, and Stephen Laurence, eds. 1999. *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Margolis and Laurence 2005 — — — . 2005. “Concepts.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford University. https://plato.stanford.edu/archives/spr2014/entries/concepts/#Aca.

Martinez-Ortiz 2016 Martinez-Ortiz, Carlos, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul, and Joris van Eijnatten. 2016. “Design and Implementation of ShiCo: Visualising Shifting Concepts over Time.” Edited by Marten Duering, Adam Jatowt, Antal van den Bosch, and Johannes Preiser-Kappeller. *Proceedings of the 3th Histoinformatics Conference, Krakow, July 11 2016*.

Mikolov et al. 2013a Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *ArXiv:1301.3781 [Cs]*, January. http://arxiv.org/abs/1301.3781.

Mikolov et al. 2013b Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013. “Exploiting Similarities among Languages for Machine Translation.” *ArXiv:1309.4168 [Cs]*, September. http://arxiv.org/abs/1309.4168.

Mikolov et al. 2013c Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” *ArXiv:1310.4546 [Cs, Stat]*, October. http://arxiv.org/abs/1310.4546.

Mitra et al. 2015 Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. “An Automatic Approach to Identify Word Sense Changes in Text Media across Timescales.” *Natural Language Engineering* 21 (5): 773–98. https://doi.org/10.1017/S135132491500011X.

Moran 1978 Moran, James. 1978. *Printing Presses: History and Development from the Fifteenth Century to Modern Times*. Berkeley: University of California Press.

Moretti 2013 Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Müller 2014 Müller, Jan-Werner. 2014. “On Conceptual History.” In: *Rethinking Modern European Intellectual History*, edited by Darrin M. McMahon and Samuel Moyn, 74–93. Oxford: Oxford University Press.

Müller and Schmieder 2016 Müller, Ernst, and Falko Schmieder. 2016. *Begriffsgeschichte Und Historische Semantik: Ein Kritisches Kompendium*. Berlin: Suhrkamp.

Nisbet 1999 Nisbet, H.B. 1999. “Herder: The Nation.” In *Approaches to the Writing of National History in the North-East Baltic Region*, edited by Michael Branch, 78–96. Helsinki: Finnish Literature Society.

Nolan 2012 Nolan, Mary. 2012. *The Transatlantic Century: Europe and America, 1890–2010*. Cambridge: Cambridge University Press.

OED 2019 *Oxford Dictionary of English*. 2019. Oxford University Press. https://doi.org/10.1093/acref/9780199571123.001.0001.

Osterhammel 2014 Osterhammel, Jürgen. 2014. *The Transformation of the World: A Global History of the Nineteenth Century*. Princeton: Princeton University Press.

O’Rourke 1999 O’Rourke, Kevin H. 1999. *Globalization and History: The Evolution of a Nineteenth-Century Atlantic Economy*. Cambridge, Mass: MIT Press.

Pocock 2016 Pocock, J. G. A. 2016. *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition*. Princeton: Princeton University Press.

Pumfrey et al. 2012 Pumfrey, Stephen, Paul Rayson, and John Mariani. 2012. “Experiments in 17th Century English: Manual versus Automatic Conceptual History.” *Literary and Linguistic Computing* 27 (4): 395–408.

Recchia et al. 2017 Recchia, Gabriel, Ewan Jones, Paul Nulty, John Regan, and Peter de Bolla. 2017. “Tracing Shifting Conceptual Vocabularies Through Time.” In *Knowledge Engineering and Knowledge Management*, edited by Paolo Ciancarini, Francesco Poggi, Matthew Horridge, Jun Zhao, Tudor Groza, Mari Carmen Suarez-Figueroa, Mathieu d’Aquin, and Valentina Presutti, 10180:19–28. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58694-6_2.

Riedl and Bieman 2013 Riedl, Martin, and Chris Bieman. 2013. “Scaling to Large3 Data: An Efficient and Effective Method to Compute Distributional Thesauri.” *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, no. October: 884–890.

Rosie et al. 2004 Rosie, Michael, John MacInnes, Pille Petersoo, Susan Condor, James Kennedy, M.J. Rosie, J. MacInnes, P. Petersoo, S. Condor, and J. Kennedy. 2004. “Nation Speaking Unto Nation? Newspapers and National Identity in the Devolved UK.” *The Sociological Review* 52 (4): 437–58. https://doi.org/10.1111/j.1467-954X.2004.00490.x.

SChultz and Schultz 2012 Schultz, Duane P., and Sydney Ellen Schultz. 2012. *A History of Modern Psychology*. 10th ed. Belmont, CA: Wadsworth.

Salmi 2013 Salmi, Hannu. 2013. *Nineteenth-Century Europe: A Cultural History*. Cambridge, Mass.: Polity.

Skinner 1969 Skinner, Quentin. 1969. “Meaning and Understanding in the History of Ideas.” *History and Theory* 8 (1): 3. https://doi.org/10.2307/2504188.

Skinner 1978 — — — . 1978. *The Foundations of Modern Political Thought*. Cambridge: Cambridge University Press.

Skinner 2012 — — — . 2012. *Liberty Before Liberalism*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139197175.

Smith 2008 Smith, Anthony D. 2008. *The Ethnic Origins of Nations*. Malden, MA: Blackwell.

Smith et al. 2015 Smith, David A., Ryan Cordell, and Abby Mullen. (2015) “Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers.” *American Literary History*, 27(3), pp. E1-E15.

Steinmetz 2016 Steinmetz, Willibald. 2016. “Forty Years of Conceptual History: The State of the Art.” In: *Global Conceptual History: A Reader*, edited by Margrit Pernau and Dominic Sachsenmaier, 339–66. London ; New York: Bloomsbury Academic.

Uimonen 2000 Uimonen, Minna. 2000. *Hermostumisen Aikakausi: Neuroosit 1800- Ja 1900-Lukujen Vaihteen Suomalaisessa Lääketieteessä*. Helsinki: Finnish Literature Society.

Viola and Verheul 2020 Viola, Lorella, and Jaap Verheul. 2020. “One Hundred Years of Migration Discourse in The Times: A Discourse-Historical Word Vector Space Approach to the Construction of Meaning.” *Frontiers in Artificial Intelligence* 3 (September): 64. https://doi.org/10.3389/frai.2020.00064.

Weitz 1988 Weitz, Morris. 1988. *Theories of Concepts: A History of the Major Philosophical Tradition*. London: Routledge.

Wells 2015 Wells, Wyatt. 2015. “Rhetoric of the Standards: The Debate over Gold and Silver in the 1890s.” *The Journal of the Gilded Age and Progressive Era* 14 (1): 49–68. https://doi.org/10.1017/S153778141400053X.

Wevers and Koolen 2020 Wevers, Melvin, and Marijn Koolen. 2020. “Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, May, 1–18. https://doi.org/10.1080/01615440.2020.1760157.