DHQ: Digital Humanities Quarterly
2016
Volume 10 Number 3

Mining Public Discourse for Emerging Dutch Nationalism

Maarten van den Bos <M_dot_J_dot_A_dot_vandenBos_at_uu_dot_nl>, Utrecht University
Hermione Giffard <H_dot_Giffard_at_uu_dot_nl>, Utrecht University

Abstract

Historians have argued that nationalism spread from elite groups to larger populations through public media, yet this has never been empirically proven. In this paper, we use digital tools to search for expressions of nationalism in Dutch newspaper discourse in the late nineteenth century by text mining in large newspaper repositories. The absence of emotional nationalist rhetoric in Dutch national newspapers suggests that nationalism in the late nineteenth century was much more subtle than the literature based on elite discourse tells us.

Introduction

In the nineteenth century, the French philosopher Ernest Renan took up an important question of his day in a lecture at the Sorbonne. His classical text, which resulted from that lecture, *Qu'est-ce qu'une nation?,* is seen as one of the first attempts to define nations scientifically. Renan disagreed with many before him who had defined the nation by criteria such as race or ethnic group or common characteristics. Instead, he defined a nation by the desire of people to belong together, their willingness to share a mutual past and a common future. Therefore, he declared, the existence of the nation was based on daily approval of this shared desire by its members [Renan 1992, 41–4]. Historians since then have offered numerous theories explaining what made allegiance to the nation-state, or nationalism, possible among groups. Key among the mechanisms cited are newspapers. 1

This paper contributes to the existing literature on nationalism by testing the thesis that asserts the importance of newspapers to rising nationalism. It will do this by approaching the question in a new way using digital tools. Using text mining to answer this question is not an attempt to define or redefine nationalism. Instead we rely on existing literature to establish the outlines of our investigation including the hypotheses regarding the role of newspapers in nationalism that we set out to test. Existing literature has primarily based its study of the nation on elite discourse. As a result, the creation of the modern nation has been described as the process of the transmission of high culture to society at large. Because digital tools allow us unprecedented access to the contents of digitized newspapers, in this paper, we use the affordances of digital tools to study a less elite view of nationalism in the late nineteenth century than is found in existing academic literature on the topic. 2

Indeed, adopting this viewpoint answers the call of many historians. In a recent essay, John Breuilly highlighted the elite bias of nationalism studies as a “major deficit” in the field [Breuilly 2012, 23]. According to Breuilly, existing studies focus overwhelmingly on either elite discourse or on politics. Often, the two concerns are merged into a single narrative that stresses the importance of elite discourse in shaping the mind-sets of nationalist politicians. Insofar as a nationalism studies considers the popular meaning of nationalism and its development within a broader sphere, this tends to be based on unproven inferences. Yet in the evidence that existing studies cite – high turnouts at nationalist festivals and ceremonies, the popular backing of nationalist politicians or the high numbers of volunteers for military service – one catches glimpses of a larger public audience that not only responded to but also shaped the meanings of nation and nationalism. We are concerned in this paper with tracing nationalism as presented to a larger audience through newspapers. Following Hobsbawm's earlier assertion, that nationalism “is constructed essentially from above, but (…) cannot be understood unless also analyzed from below, that is in terms of the assumptions, hopes, needs, longings an 3 interest of ordinary people" [Hobsbawm 2012, 10], we will use the power of digital methods to investigate nationalism from below.

Existing theories, such as those by Anderson, Hobsbawm, and Gellner, lead us to expect nationalism to have been spread in the late nineteenth century through the frequent presence of telling phrases in newspapers that invoke the greatness of a nation and its history [Anderson 1983] [Hobsbawm 2012] [Gellner 2008]. According to Benedict Anderson, newspapers with national circulation were crucial to the creation of imagined national communities by disseminating the idea of nationhood. In his view, the nation should be understood not as a geographical or political entity, but as an imagined community united by a "deep, horizontal comradeship" whereby national co-fellows are believed to constitute a bounded, natural entity. Anderson argues that the nation emerges out of imaginative contexts of social and cultural experience. Regular, synchronic readings of daily or weekly newspapers produced the idea among readers that they shared a set of interests and were part of an imagined community. Newspapers were thus a chief way that " the imagined world is visibly rooted in everyday life" [Anderson 1983, 7, 25, 36]. Although, as Tim Edensor has argued, Anderson's focus on printed media to the exclusion of all else is a very reductive view of culture, which encompasses much more than just newspapers, newspapers have consistently been singled out as important to rising nationalism in the late nineteenth century [Edensor 2002, 7–17]. So it is very much worth testing the hypothesis that newspapers contributed to nationalism. Our digital examination of the content of Dutch newspapers is a crucial contribution to understanding the nature of the "nationalist" content of newspapers. By studying the spread of nationalist sentiment through public discourse, we hope to be able to better understand the role of newspapers in the spread of nationalism. 4

Theories of nationalism suggested by other authors offer an alternative to the emphasis on emotional rhetoric. Another model of how nationalism was spread, for example, is offered by the British social psychologist Michael Billig. He argues nationalism was spread, not via the impassioned rhetoric of elite discourse, but by building a common identity through everyday expressions. Instead of the "passionate and exotic examples" often given of the spread of nationalism, Billig draws attention to the routine and mundane reproduction of the idea of the nation that is necessary to maintain it over time. He points, for example, to the use of national flags in everyday contexts, differentiating between what he calls the "waved" flags at national sporting events and the "unwaved flags" that he sees as a more important daily reminder of national belonging. Thus a person walking into the supermarket and seeing a national flag on a milk carton, even if they are looking for something else, will be unconsciously reminded of his national identity. Billing makes the same argument for the increasing use of pronouns like "we" that are used as a signifier of "us" as members of the nation and indeed the increasing use of nations' names in the public domain. Nationalism, for Billig, is marked by the fact that phrases like "the economy", "the government" and "the countryside" are immediately and unconsciously translated by readers as "our economy, government and countryside". This then constitutes an important part of the way in which nations are *naturalised*, or absorbed into a common-sense view about the way the world is. In this way they are also invested with moral values, which elevate the national over other social groupings [Billig 1995] [Skey 2009]. Billig's notion of banal nationalism supports the argument of Niek van Sas, who asserts in his well-respected study of nationalism in the Netherlands, that nationalism was much more subtle in the years around 1900 than we might expect [van Sas 2004, 161–2]. Others have cited the development and further standardization of language and the use of certain words and concepts to describe certain events, customs or practices as national as important contributions that newspapers have made to forging a national identity [Baycroft 2004, 9–10] [Lodge 1993, 2–3]. By analysing popular discourse in national newspapers using digital methods, we want to judge whether emotional nationalist rhetoric or more banal mechanisms were more important to promulgating the notion of national belonging through newspapers. 5

Using digital methods, this paper tests two hypotheses drawn from secondary literature about the nature of nationalist discourse (emotional or banal) in the late nineteenth century in national Dutch newspapers. To carry out our test, we first used topic modelling to try to capture the content of newspaper articles. We used topic modelling for its ability to produce pseudo-semantic information in the presence of *OCR* errors (which beset our corpus and made tools from corpus linguistics, as discussed below, difficult to use). The fact that we did not find emotional, nationalist rhetoric in Dutch newspapers in the late nineteenth century was very revealing, for it suggests that nationalism, which the literature agrees was present at this time, was expressed in newspapers not using grandiose, emotional appeals, but in more 6 subtle ways. This also implies that the emotional nationalism that is so often commented on — and blamed for leading to the First World War [Clark 2013] — was formed outside of newspapers and reflected in them, rather than being formed in newspapers. Newspapers, in other words, were important to spreading nationalism, but as a reinforcement of things that happened elsewhere. To check the reasonableness of our findings from topic modelling, we both close read a significant number of articles and tried to carry out a linguistic analysis of our corpus.

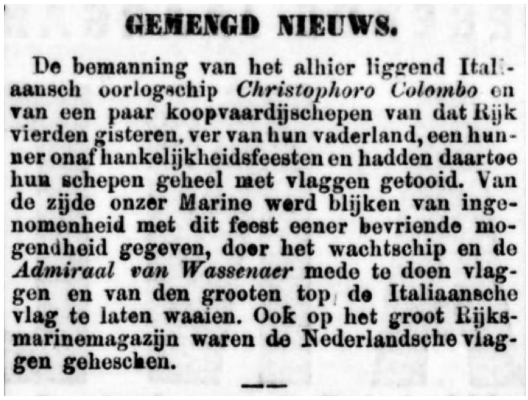
Sources

Our corpus was of Dutch newspapers in the Dutch national library, accessible via depher.nl. Based on existing literature, we decided to study newspapers published in the second half of the nineteenth century. Not only does existing literature emphasize the importance of nationalism in the late nineteenth century, but it is also the period when public debate about Dutch national identity was most strongly polarized [Aerts and te Velde 1999]. We chose 1870 as our start date to avoid the most serious effects of the national newspaper tax in the Netherlands, which was abolished in 1869. The tax served to restrict access to newspapers as well as discourage their publication. Yet over time, average incomes went up, the price of papers declined and literacy rates went up. After 1870, the Dutch media landscape became more and more fragmented and journals and newspapers became the most important vehicles of new views and opinions [van Vree 2000] [Gunn 1999]. We avoided the problem of inferring mass views from those of a limited political elite by looking directly at public discourse in newspapers, as digital tools makes possible on a large scale. Although opinion is divided about the precise relationship of the public sphere to the newspaper reporting, the use of newspapers as sources brings with it the promise of being able to chart public discourse on a scale much larger than existing studies [Broersma 2011] 7

Studying public discourse in the Netherlands offers a new perspective to a literature that has been focused on the creation of large nations within Europe: France, Germany, Italy and the United Kingdom. Although the Dutch nation had existed as a political or organizational entity for many decades before 1870 – the first national anthem was chosen in 1815 – a national identity was weak among the Dutch early on. Privileges like raising taxes, setting rules, organizing their own political and religious organisations were guarded jealously by local cities and communities. It was in the nineteenth century that the liberal elite of the Netherlands came to the conclusion that that strengthening national identity was a good way to prevent the country from falling apart into different fractions, as the Catholics and Orthodox Protestants were making strong claims to their rights as religious communities. Although neither Catholic nor Protestant elites willingly tried to subvert national solidarity, the country’s liberal elite felt threatened by each group’s divergent views of the history of the Netherlands and their claims about the country’s future. The elite therefore became shepherds of national unity, campaigning to define and reinforce a Dutch national identity. Central to this effort was the promotion of royalty: the House of Orange, and especially the young princess, Wilhelmina, who was crowned on 6 September 1898, a few days after her eighteenth birthday [Bank and van Buuren 1992, 21–89] [te Velde 1992]. Niek van Sas argues that the coronation played a critical role in stimulating national identity in the Netherlands [van Sas 1991, 600]. 8

The newspaper collection of the Dutch National Library is both a rich and a difficult collection to use. It is rich because of the quantity of digital titles that it contains, although its coverage is not uniform. In the period 1870-1914, the collection grows from 2500 newspaper issues published in 1870 to more than 6000 in 1914. While this quantity is helpful for research, mining the corpus poses three major difficulties. First, the data is messy. The collection contains a mixture of national, regional and colonial newspapers in the same dataset, although the metadata generally specifies their area of circulation, so they can be analysed separately. Second, the quality of the OCR is poor. This makes it hard to analyse these newspapers using sophisticated techniques that rely on the semantic information provided by full sentences. Figure 1 shows a sample from 1880 that epitomizes this problem. 9

GEMENGD NIEUWS.
De bomanning van het alhier ligtrcnd Italiaansch oorlogschip Christo]>horo Colombo on van een paar koopvaardijschepen van dat Kijk vierden gisteren, ver van hun vaderland, een hunner onnl' hiiukcl jjkheidsfeesten en haddon daartoe hun schepen geheel met vlaggen getooid. Van do zijde onzer Marino werd blijken van ingenomenheid met dit feest ooner bovrionde niogonühoid gegeven, doer het wachtschip en do Admiraal ran Waaneiiaer modo to doen vlaggen en van den grooten top de Italiaanscho vlag te laten waaien. Ook op het groot Kijksmarinemagazija waren do Kedcrlaudsuhe vluggen gehoseken.



GEMENGD NIEUWS.

De bemanning van het alhier liggend Italiaansch oorlogschip *Christophoro Colombo* en van een paar koopvaardijschepen van dat Rijk vierden gisteren, ver van hun vaderland, een hunner onafhankelijkheidsfeesten en hadden daartoe hun schepen geheel met vlaggen getooid. Van de zijde onzer Marine werd blijken van ingenomenheid met dit feest eener bevriende mogendheid gegeven, door het wachtschip en de *Admiraal van Wassenaer* mede te doen vlaggen en van den grooten top de Italiaansche vlag te laten waaien. Ook op het groot Rijksmarinemagazijn waren de Nederlandsche vlaggen geheschen.

Figure 1. Sample showing the quality of OCR derived from the pdf of a newspaper article typical of the corpus that we used. This example is taken from De Standaard (a Dutch newspaper with national circulation), dated 8 June 1880. About half of the short article contains OCR errors, making linguistic analysis challenging. Source: Delpher.nl.[1]

Third, the open source tools available to researchers to mine the collection are relatively simple. The chief way to access the collection is through the website. But its search engine only allows researchers to do keyword searches. In our research, we extensively used the tool known as *Texcavator*, developed at Utrecht University in collaboration with the University of Amsterdam. This tool not only enables more sophisticated searches, but also allowed us to export search results in bulk for further analysis, such as topic modelling [van Eijnatten et al. 2014]. The digitization of newspaper archives is prone to errors and biases that bring a whole range of problems for digital researchers, which must be must be taken into account by a good researcher, just as they would be with any other source [Milligan 2013]. 10

Creating A National Community

The late nineteenth century truly seemed to be the age of the nation. It marked the starting point of four major trends that would come into their own in the twentieth century: nation building, democratization, bureaucratization and the rise of the welfare state. At the start of the nineteenth century, there were no nation states in Europe except France (and perhaps the United Kingdom). Europe was divided into large empires, some constitutional monarchies and a large number of small fiefdoms with their own rulers and laws. Slowly, the nation-state became the central entity in European politics [Mazower 2012, 11] [Osterhammel 2014, 573–4, 631]. When, mainly in the second half of the century, the old empires fell and national entities like Germany and Italy were created, people became citizens of countries with presumably distinct national identities, united by shared customs, a standardized language and preferably a national history in which the diversity of the past was neglected in favour of a story that the national entity now created had always existed on a cultural level [Maier 2012]. There is no question that the creation of such states was an artificial process, but how was it achieved? 11

The Kingdom of the Netherlands was proclaimed in 1813 but lacked any real national identity. The Dutch Republic of the eighteenth century was a union based purely on military needs. It was only in times of war that the seven provinces that made up the Republic deliberated together on strategy and finance. Outside of wartime, regional and local communities were mostly self-governing. In 1795, the Batavian Revolution abolished the republic entirely and created a central state. The revolution declared national unity for the first time. After an interlude as part of France, the Kingdom of the Netherlands was proclaimed and inherited this unity. After this, there was a tendency towards a single national identity, but local identities remained strong and local customs, laws and regulations were hard to circumvent [Skinner 1989] [Prak 1999]. Indeed, in the 1830s, the kingdom fell apart into current-day Belgium, Luxemburg and the Netherlands, and each region faced the problem of defining a unique national identity. In 1848, the northern region of 12 the “nation” adopted a liberal constitution and local identities were slowly forged into a singular national identity. A new field of “national politics” became the focus of both public and political debates seeking to define, limit, and challenge the dominant vision of Dutch identity [de Haan 2003] [de Rooy 2005]. The Dutch nation increasingly became a factor in the daily life of its citizens, especially in the last quarter of the nineteenth century, when the national government took on ever more responsibilities for the wellbeing of its citizens [Wolffram 2003].

The nation’s liberal elite were key actors in the newly active national frame of the Netherlands. One of their publishing platforms was *De Gids*, a leading cultural journal established in 1837. An 1872 article by Charles Boissevain argued that a new reassessment of “public life” and “national sentiment” was needed in Dutch public discourse. The author called on his fellows to renew their commitment to defending their liberal values publicly, as classic liberal ideas on the state, public life and the nation were under fierce attack from orthodox Protestants and Catholics, who were actively rewriting Dutch national history to support their disparate discourses on national identity [Bossevain 1872] [Raedts 2011, 227–76]. Later historians, like Remieg Aerts, have demonstrated that articles in *De Gids* on Dutch national identity not only grew in number after 1870, but also changed in tone, especially from the 1880s, as a new nationalist discourse emerged [Aerts 1997, 388–424]. Aerts’ prizewinning thesis is a key study of Dutch identity in the late nineteenth century, which we used to contextualize our results. 13

Limiting A Corpus By Events

The first way that we tried to locate nationalist sentiment in public discourse was by studying the discourse around events identified in the historical literature as “nationalist”. Events were useful in our search for relevant articles because they offered a focus point for public discourse about nationalism – and a way to translate a complex concept into something that a computer could find. We used a time period that spanned two years before and after each event in question. Although we avoided the need to define nationalism ourselves by drawing on existing literature, the choice of events nevertheless implicitly defined it. Such events might have been instigated by elites, but they undoubtedly involved the populace. Because, following Aerts, we expected to see a change in the tone of Dutch nationalism over time, we compared articles reporting on a nationalist event in 1872 with articles about an event in 1898. The first was the commemoration of the landing at Den Briel. The landing at Den Briel took place in April 1572, so 1872 marked the tercentenary of the invasion of Den Briel, which was in 1572 under Spanish control - an important turning point in the Eighty Year’s War. The second event that we used was the inauguration of Queen Wilhelmina in November 1898. 14

Topic modelling, the method of analysis that we chose, is a method of text mining that has been much hyped in recent years [Blei et al. 2003]. It uses statistics to identify groups of words that frequently occur together. These groups are known as topics, but have to be identified by the user. The method is particularly good at identifying groups of related key words in data that is unstructured and may contain large numbers of transcription errors [Walker et al. 2010]. The key words belonging to each generated “topic” are “significant” in so far as they are frequently used together in the corpus. We were interested to see what words were used frequently together in discourses that we would label, using domain knowledge, nationalist: topics about nations or national events, the royal house and potentially words indicating membership in a larger group. 15

Introducing time – a crucial concern for historians - to topic modelling is an area of active research, but no group has yet managed to do this in a convincing way [Wang et al. 2008] [Qiaozhu and ChengXiang 2005]. We explored the possibility of plotting topic percentages over time using the time connected to each of our “documents” or articles, but the result was unconvincing (See Figure 2). Because there was no better way to see a change over time, we chose events at either end of our time period. (Although one could argue that these events are not equivalent. See below.) Choosing to observe historical changes this way introduces a host of other problems – each event will have different groups of words associated with each other, for example, and might share no words in common – but we judged these problems to be outweighed by the importance of looking for changes over time. 16

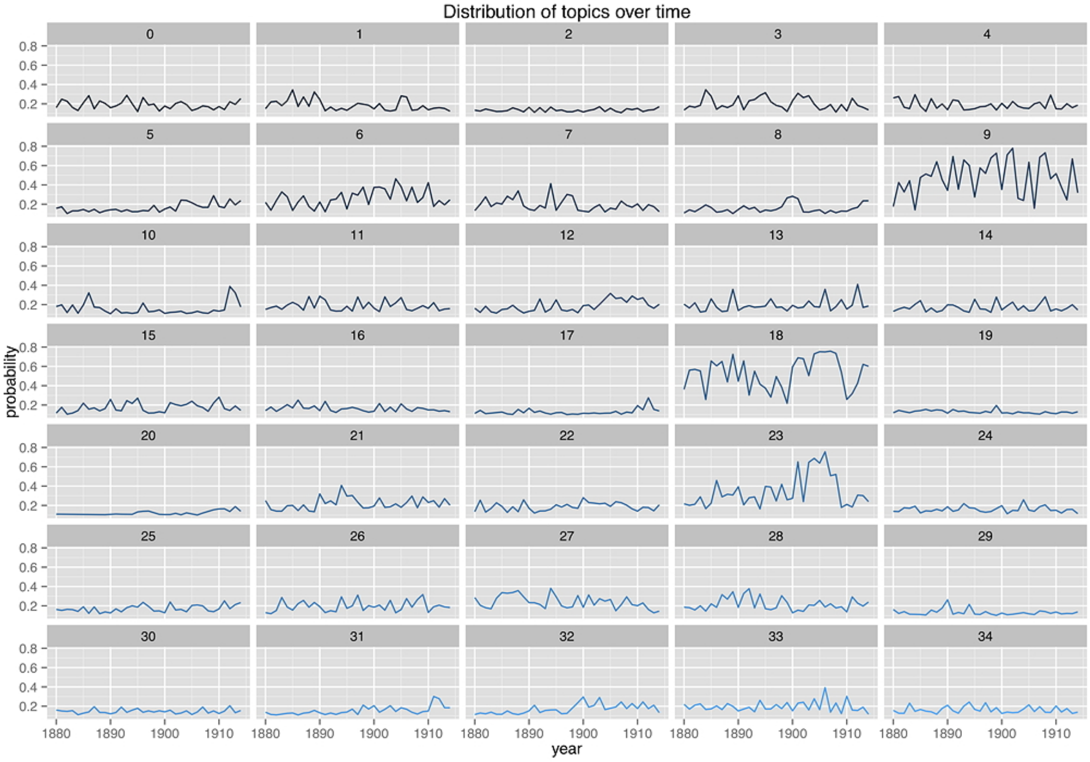


Figure 2. Plots showing each of our topics over time; the plot shows the probability attached to each topic according to the dates of our articles – this is calculated during topic modelling. The only topic that has a clearly meaningful change over time is topic number 28, but it has nothing to do with nationalism; its words describe crime. Topics 15, 23 and 31 consisted only of individual letters.

Limiting the corpus of articles modelled has two advantages: first, it makes the number of articles more manageable, and second, it emphasizes the themes of interest to historians – particularly in a corpus like ours made up of many (potentially short) newspaper articles. A key problem with topic modelling is that most computers are not able to model large numbers of documents containing a large number of words. This restricts the ability of historians to use this tool because it forces them to rely on others with larger computers. Using an undifferentiated corpus, while desirable from a numerical point of view (the desire to use as many documents as possible to increase the method's accuracy), is also problematic because common statistical measures of significance like *tf/idf* (topic frequency – inverse document frequency) are not necessarily useful for inferring historical meaning and may incorrectly emphasize or deemphasize words in shorter articles, as are typical in a newspaper [Giffard 2015]. Furthermore, the use of a large, undifferentiated corpus may lead to the elimination of topics that are of interest to historians if other topics are discussed more frequently. Because topic modelling an entire corpus might not allow researchers to explore obscure topics, researchers should use a more limited corpus. Such a corpus can help researchers avoid this problem because a limited corpus can be chosen so as to contain the concept of interest to a degree that is of statistical import. Defining a smaller corpus also gives the researcher greater control over what sources are included in it, enabling the researcher to take steps to avoid bias, for example [Blaxill 2013, 320–1]. 17

We chose to define a sub corpus of articles about each event by using key word searches in a time period about two years before and two years after the event in question. To create two corpuses, we used the two search queries: "(*Briel* or *Brielle*) and *Nederland**" and "*inhuldiging*" [inauguration].[2] To facilitate comparison between two event sub-corpuses, we chose search terms such that we'd have a similar number of articles for each time period. The best settings for topic modelling are hard to define and are different for each case; we used the same settings for our LDA topic analysis in each case (same number of iterations, topics, topic size, and same importance threshold) in order to reduce one more variable from our analysis. These choices could pose additional interpretive problems because more newspapers – and therefore more articles – came into being over time (so one would expect more "recent" events to be discussed in a greater number of articles), but we judged the subset that our search returned to be representative. In all of our tests, 18 we relied on the existing metadata in the national collection to define newspapers with “national” or “regional” circulation as well as “articles” rather than “advertisements”. For each event corpus, we modelled the content of the article including its title, but removed other metadata such as the title of the newspaper. The results are given below:

Search	Dates	Articles	Nationalist Topic Produced
“(Briel or Brielle) and Nederland”*	01.01.1870 to 31.12.1874	3,462 articles (1.79 million words)	*Vereeninging des feest koning Nederland leden warden twee oranje briel eene heeren waar uur prins*[3]
“Inhuldiging”	01.01.1896 to 31.12.1901	2,330 articles (1.5 million words)	*Koningin volk oranje majesteit groote waar willem moeder uwe onze waarop prins gouden Wilhelmina allen*[4]

Table 1. The topic from each event corpus that mentions the Dutch royal family. The settings that were used with the Mallet GUI were for 10 topics, 500 iterations, 15 words per topic, 0.1 significance threshold.

Because the two events we chose are not equivalent, we were surprised to see similar language used in the topics for both events, and thus words used in the articles mentioning these events. The first event split opinion particularly among Catholic and Protestant communities who were divided on whether the people who landed at Den Briel were heroes or not, whereas the second event was framed mainly by the liberal political elite to reemphasize national unity and reaffirm tolerance as the core of Dutch national identity by stressing the people’s ability to cope with societal and religious differences. 19

The similarity in language supports the argument, made by Frans Groot, that newspapers sought to pacify conflicts and differences by using a more neutral vocabulary [Groot 1995]. Seeing similar language used a quarter of a century later to describe the coronation of the new Dutch queen suggests that the newspapers adopted a similarly understated stance to later nationalistic events. This result suggested that we could be confident that including articles over a longer time period in our topic modelling would not compromise meaning. 20

Topic Models Over a Longer Time Period

Studying nationalism using a corpus covering a longer time period required us to find a different way of limiting our corpus than by events. Tracing a concept is naturally more complicated than following the use of a word over time, and finding a non-event based corpuses was similarly more complicated. We wanted to avoid defining “nationalism” too strictly in order to let the sources speak, so we looked for quasi-objective keywords to limit our corpus. Literature on Dutch nationalism tells us that the terms *“vaderland”*, *“volk”*, and *“natie”* did not change much in meaning over the last quarter of the nineteenth century [Aerts and te Velde 1999]. So we adopted these words as search terms. We began by creating a “nationalist” sub-corpus defined by articles containing the keyword *“vaderland”* in national newspapers published from 01 January 1880 to 01 January 1900. 21

To see that the selection focused our search, we compared this sub-corpus to one generated using a keyword that did not have to do with nationalism according to the literature. For the more general sub-corpus, we sought a corpus of similar size (in articles) to the “vaderland”-corpus. Our selection (we tried search words like “cow”, “boat”, “rain” and “tree” that might reflect frequent Dutch concerns, the number of articles returned for each is shown in Table 2.) fell eventually on *“bier”* [beer]. As we discovered, the Dutch word “bier” was often incorrectly OCR’d as the Dutch word “hier” [here], thus adding numerous articles to the corpus that did not actually have to do with beer. This was judged to be an acceptable mix-up in this case because we were searching for a general corpus rather than specific information about beer. 22

Sub-corpus defined by search term	Articles
vaderland (fatherland)	**28,807**
koe (cow)	122
boot (boat)	18,609
kaas (cheese)	20,764
Regen (rain)	21,479
Bier (beer)	**26,449**
Vrouw (woman)	58,385

Table 2. The number of articles in national Dutch newspapers containing the words “vaderland” or another keyword for the years 1880-1900.

23 Using the two sub-corpuses thus created (*“vaderland”* and *“bier”*), we were able to begin our search for nationalist sentiment. We began by doing collocation analysis on the two sub-corpuses. We found that the *“vaderland”*-corpus uses the word "*ons*" [our] within three words to the left of *“vaderland”* 7,386 times; it uses the word *“vaderland”* 36,791 times. So about 21% of the occurrences of *“vaderland”* in the *“vaderland”*-corpus are connected to ons. The *“bier”*-corpus, in contrast, uses *“ons”* up to three words before *“vaderland”* 370 times compared with 2,003 occurrences of the word *“vaderland”*, or 19% of the time.

	Bier	Vaderland
“Vaderland”	2,003	36,791
“Ons…. Vaderland”	370 (19%)	7,386 (21%)
“Volk”	6,259	20,901
“Ons…. Volk”	484 (8%)	2,463 (12%)

Table 3. The number of times that the words “vaderland” and “volk” are used in newspapers articles from national Dutch newspapers in the sub-corpus generated using the word “bier” and using the word “vaderland” for the years 1880-1900. We performed the same collocation test for “volk”. The differences between the two are not very large although the words vaderland and volk occur in the “vaderland”-corpus many more times.

24 These findings suggests that using the sub-corpus did indeed allow us to focus our search on a “nationalist” theme, because the absolute number of times the word *“vaderland”* appears is much greater than in the more general *“bier”*-corpus. Nevertheless, while the statistics suggest that the nationalist theme (represented by these words) was more present in the *“vaderland”*-corpus, we can also see by the similar percentages for the appearance of *“ons”* and *“vaderland”* that the sub-corpus did not over-represent, percentage-wise, nationalist themes. Limiting the corpus in this way did not mean that we expected only nationalist themes to emerge from the nationalism-sub-corpus (not least due to OCR errors); just that we expected nationalism to be a key theme among those discussed in the articles chosen – and indeed it appears to be.

25 We then sought to test our corpuses for the influence of time. In order to do this, we created additional limited corpuses that covered the years from 1875 – 1880 (early) and 1895-1900 (late). In comparing early to late, we saw no great increase in the use of the modifier *“ons”* before either *“vaderland”* or *“volk”* in either corpus. This suggests that nationalism was persistent in Dutch newspaper articles between 1875 and 1900 but low-level. Table 4 gives a summary of the results.

	Bier		***Vaderland***	
	Early	Late	Early	Late
Articles	2,877	16,841	6,118	11,332
“Vaderland”	399	1,008	8,100	14,172
“Ons… vaderland”	94 (24%)	155 (15%)	1,522 (19%)	2,526 (18%)
“Volk”	926	3,662	3,772	8,251
“Ons… volk”	60 (7%)	273 (8%)	405 (11%)	772 (9%)

Table 4. A similar analysis to that described in Table 3 for corpuses defined by the terms “bier” and “vaderland” from 1875 – 1880 (early) and 1895-1900 (late) using national Dutch newspapers. Again the differences are not striking, implying relatively little linguistic change between the early and late periods.

26 We next applied topic modelling to both the *“vaderland”*- and *“bier”*-sub-corpuses. We modelled the corpus connected to each event in two ways: firstly, treating each article as a single document and secondly, treating each paragraph as a single document, reasoning that paragraphs would be defined by coherent themes. Interestingly, the first produced more interpretable results, justifying our decision to choose our corpus based on numbers of articles.

27 After applying topic modelling to both corpuses, we used our domain knowledge based on existing literature to classify the topics. Looking at the topics produced by the language-indepdendent Mallet GUI, we found that both sub-corpuses had topics in common. We took this as evidence that the topics were of importance in Dutch newspapers generally, rather than just in our *“vaderland”*-corpus. But what we saw on comparing the topics generated for each corpus, like the Boerwar topic shown in Table 5, was that the corpuses used different language to discuss the same topics. Again, limiting the corpus did not mean that we expected nationalism to be the only topic discussed in these articles. We expected “nationalism” to be present in the sub-corpus, guaranteeing that we could find it, but not making it the only theme. The bold words are unique to each corpus.

Topic	**“Vaderland”-corpus**	**“Bier”-corpus**
Boerwar (1899-1902)	**Engeland** zuid **engelsche** afrika boeren Transvaal **land the engelschen amerika** oorlog **republiek** president **londen britsche groote hen lord Holland zullen**[5]	**General** oorlog boeren zuid afrika **troepen man** president Transvaal **brief warden colonel dag twee Pretoria leger ontvangen vijand berichten telegram**[6]
Trade	Land handel Nederland **onze** landbouw groote **nijverheid** nederlandsche **zeer belang duitschland** industrie **waar thans** landen **vaderland buitenland groot werk** alle[7]	Nederland handel **bier** land **artikelen firma waarde** landen alle **enz goederen** nederlandsche **invoer gebruik** industrie **jaar** groote landbouw **welke engeland**[8]

Table 5. Two topics generated from each sub-corpus; the topics were identified as similar by the authors. The bold words are unique to each corpus. The words above, drawn from nineteenth century newspapers, are in some cases spelled differently from modern Dutch. The settings we used in the Mallet GUI were for 50 topics, 500 iterations, 20 words per topic, 0.1 significance threshold.

28 The first topic shown in Table 5 shows that the Boerwar (1899-1902) was important in both corpuses. Yet the language used to describe the war was clearly different in the two sub-corpuses: the first focused more on “ethnic” descriptors rather than the more “news”-like words of the second corpus. The second topic we chose was trade, another important theme with regard to the Netherlands. In the *“vaderland”*-corpus, we found that the topic was expressed using “national” terms – particularly referring to foreign nations, which we would expect to see in a world beginning to organize itself by nation-states [Anderson 1983, 95–8]. The most similar topic in the *“bier”*-corpus, in contrast, uses more general words. It was interesting that the two corpuses had unique vocabularies for discussing similarly important events. This too suggests that nationalism in newspapers was expressed through the constant presence of more banal mechanisms that suggest a larger change in how the world was organised and understood. Although the *“vaderland”* corpus was thus more concerned with nations than the more general *“bier”* corpus, we did not see any emotional, nationalist rhetoric.

Linguistic Analysis

29 Our topic modelling suggested that larger historical changes were reflected in newspapers not through the presence of emotional rhetoric, but in the use of a vocabulary that changed over time. Indeed, many authors have pointed to linguistic changes, such as the increasing use of the term “national” or references to national organizations, that indicate the growth of national identity, so we tried to use these changes to track the process of national identity building in Dutch newspapers. Seeing an increase of such changes over time would indicate increasing intensity of personal association with a national grouping, in this case the Netherlands, as opposed to local groupings, such as the Dutch provinces. Although our attempts at linguistic analysis were suggestive but ultimately unsuccessful, we will shortly describe the major tests that we carried out and the problems that we faced in trying to come to conclusive answers.

30 Firstly, we used ngrams to track the usage of keywords that refer to nation states over time. The frequency of words like *“Nederland”* [the Netherlands], but also *“Tweede Kamer”* [the Dutch parliament], *“Premier”* [prime minister], *“nationaal”* [national], *“vaderland”* [fatherland], and *“Nederlands”* [Dutch] did indeed rise over the nineteenth century. We hoped to be able to compare this increasing usage with the number of references made to local organizational forms, since we would expect local organizational forms to be replaced by national ones. Words that refer to local communities, however, like the names of the different Dutch provinces, hardly decreased over the period in question (See Figure 3).

31 Although trying to search for these words did not yield the results we hoped for, it suggested that we might find something if we compared word usage in newspapers with local as opposed to national circulation. The results were disappointing, however. The term *“provincie”* [province] for instance appears almost as frequently in national as regional newspapers. Between 1870 and 1914, the term *“provincie”* appears 85,583 times in national papers and 112,307 in regional papers – but more regional newspapers have been digitized and this number is constantly changing, one reason why using a limited sub-corpus gives the researcher more control. It would perhaps be more fruitful to examine how many newspaper column inches were devoted to national as opposed to regional news or discourse, but that relies on being able to use a computer to distinguish between national and regional news. Ultimately, however, although a graph of linguistic use over time seems to suggest an increase in reference to national units, this is difficult to translate into information about personal identity as opposed to just bureaucratic organisation. That the national should be increasingly referred to was not unexpected given that citizens would be expected to increasingly have intercourse with national organizations as these are established. These results, while suggestive, show the difficulty of translating historical questions into questions that can be answered using digital methods – or conversely, of limiting our historical questions to questions that can be answered using digital methods.

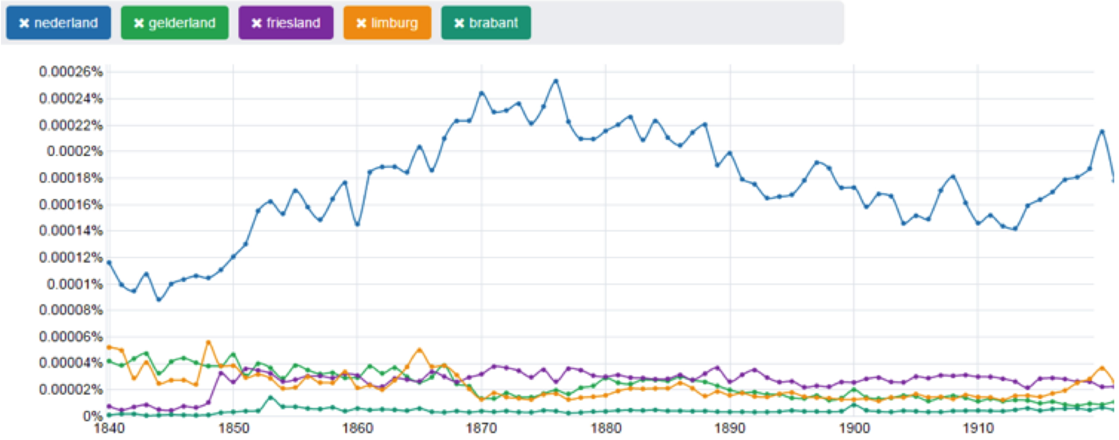


Figure 3. A graph showing the difference in usage between the words “nederland” (blue) and the Dutch provinces “gelderland” (green), “friesland” (purple), “limburg” (orange) and “brabant” (tourquoise) in all Dutch newspapers. The timelines are normalized to take into account differences in the number of newspapers that were published or digitized per year – so the y-axis shows a relative frequency, or the percentage of all the ngrams in a given year represented by a given search term. Source: Political Mashup KB n-gram viewer (http://kbkranten.politicalmashup.nl/).

Despite approaching mining newspapers from many different angles, none of our experiments turned up evidence of emotional, nationalist rhetoric in Dutch newspapers in the last quarter of the nineteenth century. Instead, we saw indications that nationalism was spread through much more quotidian methods. So, following Billig, we looked particularly at the use of pronouns in newspaper articles. Because such subtle expressions of nationalism would be expected to occur in any and all types of article – whether about national politics, sports or the weather – our examination used the whole corpus (so all articles for the time period under study). But subtle changes such as the references of pronouns are difficult to mine for. It is easy to see that the number of articles that contain both the word *“Nederland”* [Netherlands] and the pronoun *“wij”* [we] was on the rise in the nineteenth century, for example (See Figure 4). Also, one finds that nationalist word combinations like Billig’s *“ons land”* [our land] were used more frequently in 1900 than 1870 (See Figure 5). Like the previous linguistic tests, the graphs given in Figures 4 and 5 suggest an intensification of identification with the nation-state over time. Unfortunately, to be able to really assess whether the use of pronouns changes over this period, we would need to be able to determine the exact correlation between noun and pronoun. 32

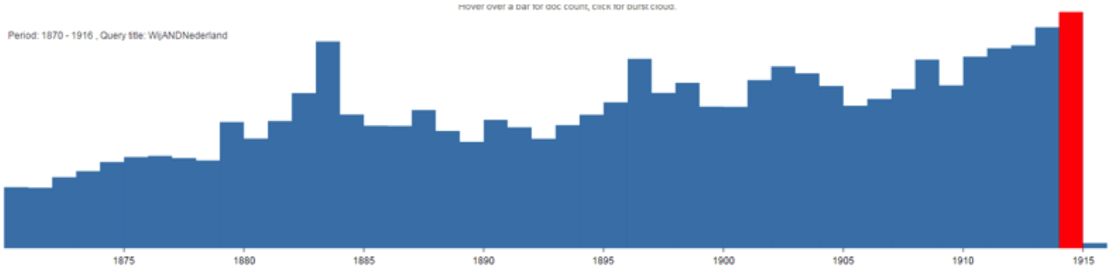


Figure 4. Time line showing the occurrence of “wij AND Nederland” in all Dutch newspapers between 1870 and 1914. Normalized. Source: Texcavator.

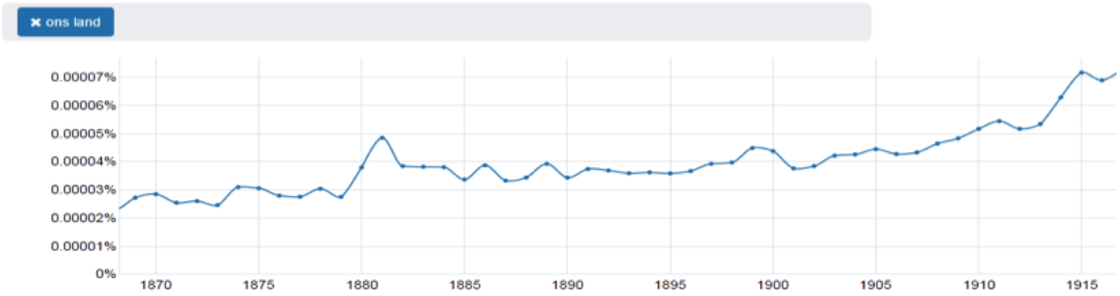


Figure 5. A graph of the n-gram “ons land” over time in all Dutch newspapers, normalized for the number of papers published in a given year. There is indeed a “burst” or peak of articles in 1898. Normalized. Source: Texcavator.

Examining the linguistic structure of a large amount of messy data poses fundamental technical difficulties – even when the OCR quality of the corpus is perfect, which was not true in our case. For the same reason – the lack of semantic information in our corpus (i.e. the guarantee that contiguous words are correctly transcribed in a large proportion of instances) – we were unable to use sentiment mining as a tool in our research, although it would have shed important light on some aspects of the “nationalism” question, such as whether individual newspapers viewed the nation positively or negatively at different points in time. In so far as linguistic changes are signals of wider historical changes, we want to be able to find them on a large scale, but we need this to be theorized by historians as well as by linguists and sociologists. It remains an open question about the potential of such analysis in a collection of nineteenth century newspapers such as we used in our study. 33

An important part of the source criticism that we wanted to do for digital searches in our data is to measure or estimate the level of error or uncertainty (and conversely certainty) that accompanies each search. We hardly expect the results to be 100 per cent accurate (given the tendency to false positives as well as false negatives), but some levels of error 34 are acceptable for our research, while others are too high. Not knowing even the order of magnitude is a serious stumbling block [Traub et al., n.d.][Milligan 2013]. We recognize that errors are generated as much from computational quirks as from digitization, and studying the errors of these methods will undoubtedly benefit from the input of trained statisticians. Until this occurs, it will be difficult to use such tools for anything more than suggestive research.

Conclusion and Outlook

The experiments described in this paper focused on finding evidence of nationalism, a complex topic that is common in the professional, historical literature. Our study of newspapers in a small, European country was particularly important because the secondary literature claims that they were very important for spreading nationalism in the late nineteenth century. We first tested whether the nationalist language present in newspapers was more of an emotional or banal sort, thereby testing existing hypotheses about how ideas about national belonging reached the public via newspapers. We used digital methods to interrogate our sources because they allowed us to analyse large corpuses of newspaper articles quickly. What we found — not in elite publications but through a much broader analysis of thousands of newspaper articles and millions of words made possible by digital tools — suggests that the majority of the Dutch population would have been exposed to nationalist thinking through subtle expressions of national belonging rather than emotional rhetoric. Although our findings are not by any means conclusive, they do demonstrate that inferring general views about nationalism from discourse in elite publications is problematic. 35

Although Michael Billig used a present-day perspective in his book and based his arguments about banal nationalism mostly on sociological literature, we found that his work establishes a useful conceptual framework in which we can understand different forms and expressions of nationalism in the public media. With regard to Dutch nationalism in particular, our findings also support and extend the argument of Van Sas. The problem, as critics of the linguistic turn in history have pointed out, is that observing a change in language in newspapers does not allow us to make a claim about causality or who the actors were who were driving a given process. Indeed, our study indicates that we need to take seriously the possibility that Western Europe’s experience with nationalism in the twentieth-century has caused historians to look for (and highlight) similarly inflamed rhetoric in earlier periods, when such passionate appeals may have been rare rather than representative. 36

Through applying digital methods to test a hypothesis found in the historical literature, this paper has contributed to both the study of nationalism and the development of digital methods for concept mining. Although our use of digital methods let us analyse many more sources than is traditional in nationalism studies, thereby enriching the field, we found that major methodological development — not to mention theorisation — is still necessary to apply existing tools to our problem. We found it both necessary and desirable to use smaller corpuses of sources to answer targeted questions, using computers as both search tools and tools of analysis. Using a limited sub-corpus may make it harder to use digital analysis techniques that require training computers using large numbers of documents, but not doing so risks that less popular themes will be drowned out by other discourses. Using a sub-corpus, as others have done [Blaxill 2013], also increases a researcher’s ability to describe the corpus and thus to carry out source criticism. 37

Unsurprisingly, creating a valid and useful sub-corpus requires composing a good search query. What this is depends on the research question. In this case, in order to study public discourse, we looked for neutral search terms (terms that the literature has concluded were neutral) over the time period that we wanted to study. While we chose terms that would point to events or topics that were publicly visible, thus providing both an opportunity and a reason for public discourse about national identity, we were careful not to choose terms that were themselves charged with meaning or possessed a changing meaning over the time period that we studied. We used multiple methods in order to check the plausibility of results from particular methods. Nevertheless, we found, as others have suggested, that external domain knowledge was indeed crucial in interpreting our results in each step of the process. The desire to use existing digital tools to answer our research question forced us to think about nationalism in many different ways, including especially how it was manifested linguistically, although OCR errors in our nineteenth century newspaper corpus made analytical methods that relied on semantic information inconclusive. We need to be careful, however, because reliance on digital methods can result in elevating certain types of meaning and tests that we as historians are not perhaps competent at using or that we would not otherwise rely on. We found, as others have, that seeking to test existing hypotheses 38 produced interesting leads for further investigation (probably using traditional historical methods — although we did not want to rely on close reading, we did end up close reading some 400 articles to check that our results were not absurd, since the methods that we used are still being developed).

With its successes and failures, this project represents an important step towards developing a method to mine public discourse from public media – both the content of and the degree of public support for particular debates over time. We will continue to explore how we can answer questions about identity formation using digital methods to analyse large repositories of historical documents. Developing our ideas further will require developing new, advanced methods for analysis as well as for corpus selection. Yet we have already seen that each and every digital output must be subjected to historical interpretation. Using digital methods is messy, but so is all historical research.

Notes

[1] Article about a celebration on board a Dutch ship.

[2] The asterisk is a multiple character wildcard search that returns matches with 0 or more additional characters. The search “Nederland*” would thus also return *“Nederlanders”* and *“Nederlandisch”* as well as words with the given stem but OCR errors in the ending like *“Nederlandens”*.

[3] union the party king Netherlands members values two orange Briel one misters were hour prince

[4] queen people orange majesty big where Willem mother your our when prince gold Wilhelmina all

[5] England south English Africa farmers Transvaal country the English America war republic president London British large the lord Holland shall

[6] general war farmers south Africa troops person president Transvaal letter wards colonel day two Pretoria army received enemy messages telegram

[7] country trade Netherlands our agriculture large hard-working Dutch very importance Germany industry where now countries fatherland foreign-country large work all

[8] Netherlands trade beer country articles company worth countries all etc. goods Dutch import use industry year large agriculture which England

Works Cited

Aerts 1997 Aerts, Remieg. 1997. *De Letterheren. Liberale cultuur in de negentiende eeuw: het tijdschrift De Gids*. Amsterdam.

Aerts and te Velde 1999 Aerts, Remieg and Henk te Velde. 1999. “*De taal van het nationaal besef, 1848-1940.*” In *Vaderland. Een geschiedenis vanaf de vijftiende eeuw tot 1940*, edited by N.C.F. van Sas, 391-454. Amsterdam.

Anderson 1983 Anderson, Benedict. 1983. *Imagined communities*. London.

Bank and van Buuren 1992 Bank, J. and M. van Buuren. 1992. *1900. Hoogtij van burgerlijke cultuur*. Den Haag.

Baycroft 2004 Baycroft, Timothy. 2004. *French Flanders in the Nineteenth and Twentieth Century.* Rochester.

Billig 1995 Billig, Michael. 1995. *Banal Nationalism*. London.

Blaxill 2013 Blaxill, Luke. “Quantifying the Language of British Politics, 1880–1910.” *Historical Research* 86.232 (2013): 313–41.

Blei et al. 2003 Blei, David, A.Y. Ng and M.I. Jordan. 2003. “Latent Dirichlet Allocation. ” *Journal of Machine Learning Research* 3: 933-1022.

Bossevain 1872 Bossevain, Charles. 1872. “*Iets over de tentoonstelling in Arti.*” *De Gids* 36: 529-562

Breuilly 2012 Breuilly, John. 2012. “What does it Mean to Say that Nationalism is ‘Popular’?” In *Nationhood from below. Europe in the Long Nineteenth Century*, edited by Maarten van Ginderachten and Marnix Beyen, 23-43. Basingstoke.

Broersma 2011 Broersma, Marcel. 2011. “*Nooit meer bladeren? Digitale krantenarchieven als bron.*” *Tijdschrift voor Mediageschiedenis* 14.2: 29-55.

Clark 2013 Clark, Christopher. 2013. *The Sleepwalkers. How Europe Went to War in 1914.* London.

de Haan 2003 de Haan, Ido. 2003. *Het beginsel van leven en wasdom. De constitutie van de Nederlandse politiek in de negentiende eeuw.* Amsterdam.

de Rooy 2005 de Rooy, Piet. 2005. *Republiek van rivaliteiten. Nederland sinds 1813.* Amsterdam.

Edensor 2002 Edensor, Tim. 2002. *National Identity, Popular Culture and Everyday Life*. Oxford.

Gellner 2008 Gellner, Ernest. 2008. *Nations and Nationalism* Ithaca, NY.

Giffard 2015 Giffard, Hermione. 2015. “Mining Newspapers. A Plea for Significance.” Beyond Methods of Mining, Utrecht, 14-15 September 2015. Conference report: asymenc.eu.

Groot 1995 Groot, Frans. 1995. “*De strijd rond Alva's bril. Papen en geuzen bij de herdenking van de inname van Den Briel, 1572-1872.* ” *BMGN–Low Countries Historical Review* 110.2: 161-181.

Gunn 1999 Gunn, S. 1999. “The public sphere, modernity and consumption: new perspectives on the history of the English middle class. ”In *Gender, civic culture and consumerism: Middle class identity in Britain 1800-1914*, edited by A. Kidd en D. Nichols, 12-29. Manchester.

Hobsbawm 2012 Hobsbawm, Eric J.. 2012 [1990]. *Nations and Nationalism Since 1780: Programme, Myth, Reality*. Cambridge University Press: Cambridge.

Lodge 1993 Lodge, A.R. 1993. *French: from Dialect to Standard.* London.

Maier 2012 Maier, Charles S.. 2012. “Leviathan 2.0: Inventing Modern Statehood. ”In *A World Connecting, 1870-1945*, edited by Emily Rosenberg, 29-282. Cambridge, MA.

Mazower 2012 Mazower, Mark. 2012. *Governing the World. The History of an Idea*. London.

Milligan 2013 Milligan, Ian. 2013. “Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010. ” *The Canadian History Review*. 94.4: 540-569.

Osterhammel 2014 Osterhammel, J. 2014. *The Transformation of the World. A Global History of the Nineteenth Century.* Princeton.

Prak 1999 Prak, Maarten. 1999. *Republikeinse veelheid, democratisch enkelvoud. Sociale veranderingen in het revolutietijdvak, ’s-Hertogenbosch 1770-1820*. Nijmegen.

Qiaozhu and ChengXiang 2005 Qiaozhu M. and Z. ChengXiang. 2005. “Discovering evolutionary theme patterns from text: an exploration of temporal text mining. ”In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 198-207.

Raedts 2011 Raedts, Peter. 2011. *De ontdekking van de Middeleeuwen. Geschiedenis van een illusie*. Amsterdam.

Renan 1992 Renan, Ernest. 1992. *Qu’est qu’une nation? Et autres essais politiques*, texts chosen and presented by Joël Roman. Paris.

Skey 2009 Skey, Michael. 2009. “The national in everyday life: A critical engagement with Michael Billing’s thesis of Banal Nationalism. ” *The Sociological Review* 57: 331-346.

Skinner 1989 Skinner, Q.. 1989. “The State. ”In *Political Innovation and Conceptual Change*, edited by T. Ball, J. Farr and R.L. Hanson, 90-132. Cambridge.

te Velde 1992 te Velde, H. 1992. *Gemeenschapszin en plichtsbesef. Liberalisme en nationalisme in Nederland, 1870-1914*. Den Haag.

Traub et al., n.d. Traub, Myriam C., Jacco van Ossenbruggen, and Lynda Hardman. “Impact Analysis of OCR Quality on Research Tasks in Digital Archives”, n.d.

van Eijnatten et al. 2014 van Eijnatten, Joris, Toine Pieters and Jaap Verheul. 2014. “TS Tools: Using Texcavator to map public discourse.” *Tijdschrift voor Tijdschriftstudies* 35: 59-65.

van Eijnatten et al. 2013 van Eijnatten, Joris, Toine Pieters and Jaap Verheul. 2013. “Big Data for Global History. The

Transformative Promise of Digital Humanities. ” *BMGN - Low Countries Historical Review* 128.4: 55-77.

van Vree 2000 van Vree, Frank. 2000. *De politiek van de openbaarheid. Journalistiek en de publieke sfeer.* Groningen.

Walker et al. 2010 Walker, Daniel D., William B. Lund and Eric K. Ringger. 2010. “Evaluating Models of Latent Document Semantics in the Presence of OCR Errors. ”In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* 240–50.

Wang et al. 2008 Wang, Chong, David Blei, and David Heckerman. 2008. “Continuous Time Dynamic Topic Models. ” In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, 579-586.

Wolffram 2003 Wolffram, D.J. 2003. *Vrij van wat neerdrukt en beklemt. Staat, gemeenschap en sociale politiek, 1870-1918*. Amsterdam.

van Sas 2004 van Sas, Niek. 2004. *De metamorfose van Nederland. Van oude orde naar moderniteit, 1795-1900.* Amsterdam.

van Sas 1991 van Sas, Niek. 1991. “*Fin-de-Siècle Als Nieuw Begin. Nationalisme in Nederland Rond 1900.*” *BMGN - Low Countries Historical Review* 106.4: 595-609.