"Vaderland", "Volk" and "Natie": Semantic Change Related to Nationalism in Dutch Literature Between 1700 and 1880 Captured with Dynamic Bernoulli Word Embeddings

Marije Timmermans* GBBO, De Bilt, The Netherlands marije.r.t@gmail.com Eva Vanmassenhove Department of CSAI Tilburg University The Netherlands e.o.j.vanmassenhove@uvt.nl

Dimitar Shterionov Department of CSAI Tilburg University The Netherlands *d.shterionov@uvt.nl*

Abstract

Languages can respond to external events in various ways - the creation of new words or named entities, additional senses might develop for already existing words or the valence of words can change. In this work, we explore the semantic shift of the Dutch words "natie" ("nation"), "volk" ("people") and "vaderland" ("fatherland") over a period that is known for the rise of nationalism in Europe: 1700-1880 (Jensen, 2016). The semantic change is measured by means of Dynamic Bernoulli Word Embeddings (Rudolph and Blei, 2018) which allow for comparison between word embeddings over different time slices. The word embeddings were generated based on Dutch fiction literature divided over different decades. From the analysis of the absolute drifts, it appears that the word "natie" underwent a relatively small drift. However, the drifts of "vaderland" and "volk" show multiple peaks, culminating around the turn of the nineteenth century. To verify whether this semantic change can indeed be attributed to nationalistic movements, a detailed analysis of the nearest neighbours of the target words is provided. From the analysis, it appears that "natie", "volk" and "vaderland" became more nationalisticallyloaded over time.

1 Introduction

The nineteenth century is often characterized as the era of modernity and nationalism (Leerssen, 2006; Hobsbawm, 2012; Jensen, 2016; Gellner, 1983). However, the development of the modernist mindset did not happen overnight. Brunner et al. (1972) call this cultural transition period from the early modern period to the modern period the *Sattelzeit* or saddle period. In this period, from roughly 1750 to 1850, the reading public expanded, people became used to thinking about the past and the future, ideologies such as nationalism arose, and abstract concepts became more politically applicable.

This paper aims to contribute to the study of the development of the cultural thought of nationalism during the Sattelzeit in Dutch society by researching fiction literature from 1700 to 1880. By employing a dynamic word embedding model we examine whether the (literary) contexts of three target words "natie" ("nation"), "volk" ("people") and "vaderland" ("fatherland") have changed over the course of the eighteenth century and nineteenth century. The dynamic word embedding model allows us to measure to what extent the contexts might have changed by quantifying the semantic drift leveraging the target words' embeddings. By doing so we aim to establish whether there is indeed a measurable semantic change or drift that coincides with the upcoming cultural and political thoughts of the era.

2 Related Work

In history, studying how concepts have changed over time is called *Begriffsgeschichte* or conceptual history. Influential in conceptual history are the works from Kosselleck (2002), Foucault (1970) and Skinner (2002). Van Sas (1999) studied different representations of the Dutch nation by looking at words expressing concepts related to nationalism over the centuries, using political texts and literature from the fifteenth century to 1940.

In the field of digital humanities, dynamic word embeddings can be employed to measure how words change over time. Word embeddings are distributional representations of words constructed based on their distribution in texts, i.e. these embeddings quantify how often words co-occur with other words in (preferably large) corpora. This idea is based on the distributional hypothesis, which presumes that the meaning of a word can be derived from its linguistic context (Firth, 1957). Semantic representations can be learned using Natural Language Processing techniques, such as Word2Vec (Mikolov et al., 2013), that automatically learn associations by leveraging information from large corpora. These distributional methods have been proven suitable to capture (broad) semantic changes in large generic corpora (Hamilton et al., 2016; Kutuzov et al., 2018; Tahmasebi et al., 2021). More recently, word embeddings have also been used to investigate semantic shifts in historical contexts, e.g. shifts in gender bias in historical newspapers (Wevers, 2019), changes in gender and ethnic stereotypes (Garg et al., 2018), evolution of concepts (Orlikowski et al., 2018), study of parliamentary debates (Van Lange and Futselaar, 2018) and others.

A practical difficulty with dynamic word embeddings arises when attempting to compare the embeddings over different time periods. This problem is referred to as the alignment problem and different solutions have been proposed (Di Carlo et al., 2019; Hamilton et al., 2016). A second challenge, especially when dealing with historical data, is the fact that large corpora are required to train word embeddings (e.g. the model of Hamilton et al. (2016) required a dataset of 100,000,000 words per time slice). Kim et al. (2014), Bamler and Mandt (2017), Yao et al. (2018) and Rudolph and Blei (2018) proposed a dynamic word embedding model for handling such sparse data.

For the current research we employ the Dynamic Bernoulli Embedding model of (Rudolph and Blei, 2018). Rudolph and Blei (2018) demonstrated that Dynamic Bernoulli Embeddings give good predictive performance for time windows with sparse data. Moreover, their method is able to capture changes of rare words. Both Dynamic Filtering of Skip-Gram and Dynamic Bernoulli Embedding are able to detect drifts within very sparse datasets. However, the Dynamic Bernoulli Embedding has been shown to keep words that do not change over time more stable (Montariol and Allauzen, 2019). In our experiments, we apply the Dynamic Bernoulli Embedding model to Dutch literature to study the semantic shift of words related to nationalism, with the goal to contribute to the analysis of the historical discourse on nationalism.

3 Experimental Setup

3.1 Dataset

The data is retrieved from the Digital Library of Dutch Language (DBNL)¹ which contains thousands of literary texts as well as secondary litera-

ture and additional information (e.g. biographies and portrayals) from The Netherlands and Belgium. We limited the data collected to fiction, since these works are more widespread than non-fiction and given that the content of popular genres in the nineteenth century had nationalistic tendencies. The historical novel romantically celebrated the nation's past, while the rustic novel and the realistic novel showed their readers the social and moral representation of the nation (Rigney, 2020; Leerssen, 2020). While rhyme and other stylistic specifics can have an effect of the position and context of the target words, poetry is also included since it is makes up a large percentage of literary works in the DBNL. This is in particular true for the earlier decades of the time period of interest.

The final dataset compiled from DBNL consists of 414 fiction books, such as prose, plays and youth literature from the time period between 1700 and 1880. To capture change over time, the data is sliced into bins per decade, based on their publication dates. The data is divided in a training (80%) and a validation (20%) set.

3.2 Preprocessing Steps

We applied spelling normalization based on the work by Braun (2002). Stop words were removed using the NLTK package (Bird et al., 2009) and the word frequency in texts from the target time period (1700-1880). Additionally, words that are less than two characters/numbers were pruned as well as words occurring less than ten times in the documents. These steps ensure a compacter dictionary for the model.

3.3 Dynamic Bernoulli Embeddings

We employ the Dynamic Bernoulli Embedding model (Rudolph and Blei, 2018). Rudolph and Blei (2018). This model is a type of exponential family embeddings that captures sequential changes in the data representations. It extends Bernoulli embeddings for text which provide a conditional model for individual text entries to text data over time. This model has a good predictive performance for time windows with sparse data. It has been proven that this model captures changes of (rare) words while keeping words that do not change relatively stable (Montariol and Allauzen, 2019). The number of passes over the data is ten, with an additional first pass, or zeroth pass, where the embedding vector is trained on all the time slices, for initialization. The dimension of the embeddings are set to 100,

¹https://www.dbnl.org/

and the number of negative samples is set to 20. These settings are based on the settings of Rudolph and Blei (2018). After 100 mini batches, the positive likelihood (L_{pos}) is calculated on the validation set and saved. The context size employed is six.

Model	values
context size	6
passes over data	10 + 0th
dim. of embeddings	100
Hyperparameters	values
minibatch	100, 300, 500
learning rate	0.2, 0.02, 0.002, 0.0002
drift	1, 5 and 10

Table 1: Model and hyperparameter settings that were explored for the experiments (based on the optimal hyperparameters identified by Rudolph and Blei (2018))

The hyperparameters that need to be determined are the batch size, the learning rate and the precision of the random drift. We limited the search for optimal hyperparameters based on the experiments described in Rudolph and Blei (2018). The model is expensive to run, so for efficiency reasons, instead of testing a combination of all the settings, we first determine the optimal batch size, while keeping the other hyperparameters on their default setting. We keep the batch size setting that gives the highest L_{pos} on the validation set, for comparing different learning rates. This is repeated for the last setting, the precision of the random drift. Then, the model with the one with the highest L_{pos} is chosen as the final model.

4 Results

The Dynamic Bernoulli Embedding models are evaluated with the Bernoulli positive likelihood on the validation set, or L_{pos} . This metric is used to select the hyperparameter settings. The results of the experiments are represented in Table 2.

In Table 3 the absolute drift of the target words are given. The absolute drift is the metric used to measure how much the context, or the usage of a word changes over time. According to the final model, the target word that has the largest absolute drift and thus changed the most over time is "volk" (0.1800), followed by "vaderland" (0.1128). The target word "natie" shows the smallest absolute drift (0.0644). As a reference frame, the word with the largest absolute drift in our dataset is the word "we" (informal form of the Dutch 1^{st} per-

Minibatches	Learning rate	Drift	Lpos val
100	0.002	1	-6854684
300	0.2	1	-11756400
	0.02	1	-6560637
	0.002	1	-6281450
	0.002	5	-7143437
	0.002	10	-6064922
	0.0002	1	-7961574
500	0.002	1	-6287337

Table 2: Overview of hyperparameter settings that were explored together with the L_{pos} on the validations set for every setting.

son plural pronoun "wij", "we" in English), with an absolute drift of 0.3864. Looking at the position of the words with the largest absolute drift, the words "volk", "vaderland", "natie" are in the 157^{st} , 882^{nd} and 3711^{th} place of a total of 61114 terms. We ought to note that many of the words with large drifts are either words with old spelling forms or names. First, words with old spelling have a hight absolute drift since they go out of use in later decades - their position in the word embedding does only rely on the drifting prior mechanism. Although we implemented a spelling normalization step as a pre-processing step, not all spelling inconsistencies were successfully corrected. Second, names used in fictional literature are also among the words with the largest drifts due to the fact that they only appear in some books (and in some decades).

Target Word	Embedding
Mean absolute drift	0.0253
"natie" ("nation")	0.0644
"volk" ("folk")	0.1800
"vaderland" ("fatherland")	0.1128

 Table 3: Absolute drifts for the target words and the mean absolute drift of all words

Figure 1 illustrates the drift of the target words over the different time slices. This graph shows that while "natie" has a small absolute drift, the drift becomes a bit larger over time, but it stays below the average drift of words.

The word "vaderland" shows three peaks in their drift over time. The first peak coincides roughly with the emergence of the word "vaderland" in Dutch book titles (Kloek, 1999). The second peak is around 1780, which is the decade of the politi-

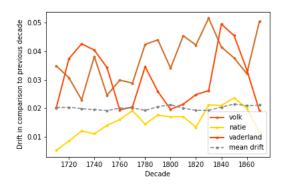


Figure 1: The drift of target words and neutral words, in comparison to their position in the previous decade

cization of the Dutch enlightenment (Kloek, 1999). The third peak of drift happens in the decade of the Dutch constitutional reform of 1848 and the revolutions in Europe of the same year.

Aside from the absolute drift and the drift over time, the nearest neighbours of the target words in the embedding of a specific time slice were analyzed. The nearest neighbors can be understood as the word most often used in a similar context of the target words, and are thus considered semantically close according to the distributional hypothesis. Due to the page-limit, we restrict ourselves to a brief illustration of the nearest neighbors of the word "vaderland". We allude at some of the findings we observed for the word "volk" and "natie".

The word "volk" changes fast from the last quarter of the eighteenth century onwards. The nearest neighbors of "volk" in the earlier decades of the eighteenth century are mainly related to biblical themes. De Kruif (2001) explains that biblical literature was popular in the eighteenth century. The interpretation of the word "volk" changes from "people of Israel" towards the meaning of "people as a mob" in later decades, which explains the larger drift from the 1780s onwards.

For "vaderland", the nearest neighbors are "geboorteland" ("country of birth") and "geboortegrond" ("place of birth"). Among the neighbors are also some more affectionate words such as "dierbare" ("dear") and "dierbaarst" ("dearest") and "vrijgevochten" ("free-spirited"). Aside from that, among the top 10 nearest neighbours, we can find terms alluding at a fatherland's past: "wapenroem" ("fame of arms"), "onafhankelijkheid" ("independence"), and, specific to the Dutch past, "bataven" ("batavian(s)"). From 1780 onwards, the words "nederland" ("The Netherlands") and "vlaanderland" ("Flanders") are present in the top 10. We give an example of the top 10 nearest neighbours of the word "vaderland" in Table 4.

For "volk", the word "oproerig" ("rebellious") is the nearest neighbor for every decade except the last one. Other words like "oproer" ("rebellion"), "muitziek" and "muitzucht" ("mutinious") emphasize the dangerous/negative connotation of "people" ("people as a mob"). These words furthermore appear more frequently and get a higher position in the top 10 nearest neighbors in the later decades. "Natie" showed almost no variation over time, as was expected by the low absolute drift. For "natie", many of its nearest neighbours across the different time slices were words referring to institutions

While "natie" didn't undergo a traceable semantic shift according to the final mode, we nevertheless looked into the nearest neighbours over the decades. "Handeldrijvende" ("trading") is the nearest neighbor in all decades, followed by "naäpen" ("copying", as in what a copycat does) and "Nationaliteit" ("nationality"). Further down the neighbouring words we encounter institutions (universities ("universiteiten"), courts ("gerechtshoven", "rechtbanken"), governments ("gouvernments"), people's government ("volksregering"), and republic ("republiek").

5 Conclusions and Future Work

In this study the emergence and development of nationalism in Dutch culture is studied by looking at the semantic change of the target words "natie", "volk" and "vaderland". This is done by applying the Dynamic Bernoulli Model, proposed by Rudolph and Blei (2018), to Dutch fiction literature between 1700-1880, during the emergence and development of nation building and nationalism (Gellner, 1983). Furthermore, through the analysis of the nearest neighbours we show how the contextual meaning of the target words changed. To the best of our knowledge, this is the first research that uses Dynamic Bernoulli Embeddings to contribute to an analysis of historical discourse, which in this case is the debate on the origins and spread of nationalism in the Netherlands. The results of this study show that there are measurable changes in the dynamic word embeddings of words related to nationalism over the course of the eighteenth and nineteenth century during a period that is known as the Sattelzeit.

We want to acknowledge certain limitations of

1760	1770	1780	1790	1800	1810
geboorteland	geboorteland	geboorteland	geboorteland	geboorteland	geboorteland
geboortegrond	volksbestaan	volksbestaan	volksbestaan	geboortegrond	geboortegrond
volksbestaan	geboortegrond	geboortegrond	geboortegrond	volksbestaan	volksbestaan
dierbaarst	dierbaarst	dierbaarst	dierbaarst	dierbaarst	dierbaar
eendrachtsband	eendrachtsband	eendrachtsband	eendrachtsband	dierbaar	dierbaarst
vrijgevochten	onafhanklijkheid	vrijgevochten	dierbaar	eendrachtsband	eendrachtsband
onafhanklijkheid	vrijgevochten	dierbaar	onafhanklijkheid	onafhanklijkheid	onafhanklijkheid
wapenroem	wapenroem	bataven	vrijgevochten	vaderlande	vaderlande
vaderlande	bataven	onafhanklijkheid	bataven	bataven	vrijgevochten
roemvol	dierbaar	vlaanderland	vlaanderland	vlaanderland	nederland

Table 4: Top 10 nearest neighbors for the target word "vaderland" for the time slice 1760 - 1810. The words "vlaanderland" and "nederland" (marked in bold) start showing among the 10 nearest neighbors from 1780.

our study. The method we employed in our study is sense-agnostic (e.g. homonyms) and works on a vocabulary common to all the investigated time slices, meaning that only words appearing in the entire corpus can contribute to the analysis of the target words. Additionally, historical sources often employ a non-standardized spelling. Despite our preprocessing to standardize the spelling using general rules, it is possible that some variations have been missed.

A more in-depth analysis of how semantic shifts of words can reflect the development of nation building, nationalism (or in general the arrival of certain ideologies) is a research direction we aim to further explore in future work. Future work could also benefit from a more in-depth exploration of the hyperparameter settings and their combinations.

References

- Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning, volume* 70 of Proceedings of Machine Learning Research, pages 380–389, Sydney, Australia. PMLR.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media, Beijing.
- Loes Braun. 2002. Information Retrieval from Dutch Historical Corpora. Master's thesis, Maastricht University.
- Otto Brunner, Werner Conze, Reinhart Koselleck, and Arbeitskreis für Moderne Sozialgeschichte, editors. 1972. *Geschichtliche Grundbegriffe*, 4. aufl edition. Number historisches Lexikon zur politisch-sozialen Sprache in Deutschland / hrsg. von Otto Brunner; Werner Conze; Reinhart Koselleck. [Hrsg. im Auftrag des Arbeitskreises für Moderne Sozialgeschichte e.V.]; Bd. 2 in Geschichtliche Grundbegriffe. Klett-Cotta, Stuttgart. OCLC: 246138897.

- José De Kruif. 2001. Classes of readers: Owners of books in 18th-century The Hague. *Poetics*, 28(5-6):423-453.
- Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training Temporal Word Embeddings with a Compass. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6326–6334.
- J.R. Firth. 1957. A synopsis of linguistic theory 1930-1955. In Frank Palmer, editor, *Selected Papers of J.R. Firth 1952-1959.* Longman.
- Michel Foucault. 1970. The archaeology of knowledge. *Social Science Information*, 9(1):175–185.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings* of the National Academy of Sciences, 115(16):E3635– E3644.
- Ernest Gellner. 1983. *Nations and nationalism*. New perspectives on the past. Blackwell, Oxford.
- William L. Hamilton, 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), pages 1489–1501. Association for Computational Linguistics.
- E.J. Hobsbawm. 2012. *Nations and nationalism since* 1780. Cambridge University Press.
- Lotte Jensen. 2016. The Roots of Nationalism: National Identity Formation in Early Modern Europe, 1600-1815. Amsterdam University Press.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

- J.J. Kloek. 1999. Vaderland en letterkunde. In Niek van Sas, editor, *Vaderland*, number I in Nederlandse Begripsgeschiedenis. Amsterdam University Press, Amsterdam.
- R. Kosselleck. 2002. The practice of conceptual history: Timing history, spacing concepts. Stanford University Press.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.
- J.T. Leerssen. 2006. National thought in Europe: A cultural history. Amsterdam University Press.
- J.T. Leerssen. 2020. Literary realism and the nation. In J.T. Leerssen, editor, *Encyclopedia of Romantic Nationalism in Europe*. Amsterdam University Press.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, volume 26, pages 3111–3119. Curran Associates, Inc.
- Syrielle Montariol and Alexandre Allauzen. 2019. Empirical Study of Diachronic Word Embeddings for Scarce Data. In *Proceedings of Recent Advances in Natural Language Processing*, pages 795–803, Varna, Bulgaria.
- Matthias Orlikowski, Matthias Hartung, and Phillipp Cimiano. 2018. Learning Diachronic Analogies to Analyze Concept Change. In *Conference: 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11, Santa Fe, New Mexico, United States of America.
- Ann Rigney. 2020. The historical novel. In J.T. Leerssen, editor, *Encyclopedia of Romantic Nationalism in Europe*. Amsterdam University Press.
- Maja Rudolph and David Blei. 2018. Dynamic Embeddings for Language Evolution. In *WWW*, pages 1003–1011.
- Quentin Skinner. 2002. *Visions of politics*. Cambridge University Press.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection.
- M. Van Lange and R. Futselaar. 2018. Debating evil: Using word embeddings to analyze parliamentary debates on war criminals in The Netherlands. In Proceedings of the Conference on Language Technologies & Digital Humanities, pages 147–153. Znanstvena založba Filozofske fakultete v Ljubljani.

- Niek Van Sas, editor. 1999. Vaderland. Een geschiedenis vanaf de vijftiende eeuw tot 1940. Amsterdam University Press.
- M. Wevers. 2019. Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990. pages 92–97.
- Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, pages 673– 681, New York, NY, USA. Association for Computing Machinery.