Debating Evil

Using Word Embeddings to Analyze Parliamentary Debates on War Criminals in The Netherlands

Milan M. van Lange, Ralf D. Futselaar
NIOD, Institute for War, Holocaust and Genocide Studies
Herengracht 380, 1016CJ Amsterdam, The Netherlands
m.van.lange@niod.knaw.nl, r.futselaar@niod.knaw.nl

Abstract

We are proposing a method to investigate changes in historical discourse by using large bodies of text and word embedding models. As a case study, we investigate discussions in Dutch Parliament about the punishment of war criminals in the period 1945-1975. We will demonstrate how word embedding models, trained with Google's Word2Vec algorithm, can be used to trace historical developments in parliamentary vocabulary through time.

Keywords: War Criminals, Penal History, Word2Vec, Word Embedding Models

1. The case: War Criminals

Soon after German forces in the Netherlands surrendered in May of 1945, the question arose how the hundreds of suspected war criminals and thousands of Nazi collaborators in Dutch custody were to be treated. For the next five decades, this question caused a series of heated political controversies. The debates in Dutch parliament about the punishment, penalty reduction, or release of these people are not only among the longest debates in Dutch parliamentary history, but are generally considered to have been the most emotionally charged (Bootsma and van Griensven, 2003; Futselaar, 2015; Tames, 2013).

1.1. Discourse and controversy

In this paper, we use an implementation of word embedding models (WEMs) to analyze parliamentary discussions concerning incarcerated war criminals and Nazi collaborators after the end of the German occupation. At peak, in the summer of 1945, more than a hundred thousand people were incarcerated. They were accused of a variety of crimes, all committed during the occupation of the country: political and military collaboration, war crimes, and (complicity in) genocide. The overwhelming majority of these prisoners were released quickly, but a small and dwindling number remained in prison until 1989. After the 1960s, all remaining prisoners were former German officials and officers whose initial death sentences had been commuted to life in prison. As long as they remained behind bars, political controversy about plans for their release continued to resurface (Tames, 2013; Piersma, 2005).

We map the language used in Dutch parliament to discuss this specific case during a relatively short historical period. The results will enable us to track the preferred vocabularies in these discussions through time. In other words, we use the words spoken in plenary sessions of the Dutch parliament as a reflection of the vocabulary used. This vocabulary, in turn, we assume reflects the changing discourse about incarcerated war criminals in Dutch society. Thus, we aim to link these developments in parliamentary vocabulary to actual historical events, developments concerning the post-war dealing with war criminals, and discursive shifts in Dutch society (Olieman et al., 2017). Specifically, we aim to investigate the changing political attitude towards incarcerated war criminals and use our findings to test established notions prevalent in Dutch historiography.

The published proceedings provide us with a dataset comprising of all the words spoken in plenary sessions in both houses of parliament. The completeness of the parliamentary dataset allows us to investigate the changing parliamentary vocabulary through time, and in the context of different discussions. This vocabulary changed, and we use these changes to investigate, ultimately, the changing discourse in postwar Dutch society.

We here focus on two questions directly related to the treatment of these delinquents in the Dutch penal system. The first of these concerns the focus on the identification of the wronged party: did politicians focus on crimes against the dutch nation as a whole, or against specific groups of individual victims? The second concerns the appropriateness of harsh punishments, specifically whether or not life imprisonment was considered a just alternative for the death penalty. These questions both derive directly from historiography and serve to answer an overarching question: can we assess the validity of traditional scholarship using unsupervised text mining?

2. Parliamentary proceedings

In this investigation, we rely entirely on parliamentary proceedings, known in Dutch as the *Handelingen der Staten-Generaal*. The Handelingen are available in machine-readable form. The minutes of both houses of parliament for the period 1814-1995 were first digitised by the Royal Library of the Netherlands and made available to the public in 2010. The dataset for the period since 1946 was dramatically improved in the *political mashup* project that ran from 2012 to 2016. This improved and enriched dataset

is freely available, on request, from DANS, the Dutch national repository of research data. The dataset consists of a large collection of XML files containing the complete minutes of all the meetings of the lower and upper chambers of parliament, separated by date, speaker, political affiliation, etc. This makes it an excellent corpus for various forms of automated text analysis.[1]

3. Word Embedding Models and Historical Research

We investigate the vocabularies used in parliament to discuss a broad category of inmates that could be described as political delinquents, as well as the changes of these vocabularies through time. This is a fairly normal investigation undertake in traditional historical research, that is to say without computational analyses. Historians typically work by reading the relevant texts. This approach has several disadvantages. In this particular case the corpus to be assessed is enormous, making manual encoding of text problematic. More importantly, the traditional research process is highly vulnerable to the biases of the reader/researcher. When studying ethically charged controversies in the relatively recent past, this vulnerability to bias is evidently problematic.

3.1. Words in vector space

A WEM provides a possible solution to these problems. WEMs are techniques to investigate words, and relations between words, in large text corpora. More specifically, WEMs are based on the calculation of the average distance of unique words to all other unique words in a corpus. This results in a list of numerical values, that make up the 'vector' for each word. In principle, the number of values, also referred to as 'coordinates', or 'dimensions' of the vector, is the same as the the number of unique words in the text, minus one. The complete trained corpus, or 'spatial model', is often referred to as a vector space. Within this space, the position of a specific word relative to all other words, is described by its vector.

Since the position of unique words relative to other words is an average calculated on the basis of all occurrences in the text, WEMs are exceptionally effective at investigating relations between relatively frequent words in a sufficiently large text corpus. The method does not prioritize any particular words; the position of each unique word is investigated. Obviously, many close relationships occur only once or a few times. Other relationships appear frequently. Some words are synonyms or near synonyms, have very similar usages (tea and coffee, for example) or often appear in combination (New and York). The analytical possibilities of WEMs, as we will demonstrate below, go far beyond mere closeness. With WEMs we are able to identify associations between words that are not self-evident.

[1]Maarten Marx, Johan van Doornik, Andre Nusselder and Lars Buitinck, *Dutch Parliamentary Proceedings 1814-2012, nonsemanticized*, (October 10, 2012), Distributed by DANS EASY, https://doi.org/10.17026/dans-xk5-dw3s

3.2. Limitations of WEMs

WEMs also have an important downside that is particularly relevant to historical research. Since the training of the model determines the position of a word relative to all other words in that specific corpus, its vector is meaningless in any other model. Word vectors, hence, can only be compared with other word vectors within the same spatial model. For historians, this means that comparisons between different moments in time are likewise impossible, because each period in time would result in a different 'bag of words' and hence a different, and incomparable, spatial model. This means that, while WEMs are perfectly adequate tools for fulfilling the first of our aims, investigating vocabularies, they are virtually useless for the second aim, investigating change through time. Since change through time is the core of virtually all historical research (including this investigation), this presents us with a major problem; how can we compare outcomes for different WEMs, for different periods in time? We have, however, developed a workaround to enable us to use WEMs to investigate changing ways to talk about certain topics through time, about which more below.

3.3. Word2Vec

For this investigation, we have have used the relatively popular Word2Vec implementation of WEMs to train and analyze word embedding models. Word2Vec was developed by a team of Google engineers and published in 2013. It has been shown to be a particularly effective implementation. This algorithm, however, was developed with a different aim than the one for which we are using it. Initially, Word2Vec was a tool to investigate natural language itself, for example to identify (near) synonyms. In our, historical, investigation, the statistical modeling of language as such is not the objective. Rather than trying to identify linguistic regularities to investigate language, we focus on linguistic irregularities and patterns to identify the influence of political and historical change on changes in the language used in political speech.

For researchers using the R language, a package is readily available to analyse texts. This package, created and maintained by Benjamin Schmidt, has been used in this investigation as well (Schmidt, 2015). Our method, however, is in no way dependent on this particular platform and could also be used in Python or any other environment. Neither is the method reliant on the Word2Vec algorithm. It would work broadly in the same way with another implementation of word embeddings. Here, however, we have chosen to use a popular WEM implementation in a relatively user-friendly and accessible environment, with the added benefit of using open-source, free software.

4. Analytical process

Text analysis with WEMs involves two necessary steps. The first of these, the training of the corpus, creates the spatial model, the WEM itself. The second step is the analysis of the positions of specific words or word clusters within the virtual space of the model.

The corpus of the Handelingen is vast by the standards of historical research, but not very large for the kind of anal-

ysis we are undertaking. For the purpose of WEMs, the size is barely adequate. Therefore we have trained our dataset with a Skip-Gram Word2Vec model, which has anecdotally been shown to yield better results on smaller samples (Gelbukh, 2015).

Within the model, the vectors of different words can be compared by using cosine similarity. Within a vector space, any two vectors by can be described, by definition, as lying within a horizontal plane. Cosine similarity calculates the angle between these vectors. Perfectly overlapping vectors would result in a cosine similarity of 1, a perfectly opposite relationship -1. In practice, WEMs consist only of positive space, which means that scores fall between 0 (low, or no similarity) and 1 (high, or perfect) similarity (Singhal, 2001).

4.1. Training the models

The first step of our workaround is to train two WEMs (more than two is equally feasible), based on two corpora (in this case 1945-1955 and 1965-1975). Each of these corpora contains ten years of parliamentary speeches. (When using this approach, it is necessary to use relatively similar corpora, both in terms of size and in terms of language use. For historical research into relatively short periods of parliamentary history, this is not particularly problematic.) For reasons of efficiency, we have limited ourselves to unique words that appear at least five times in the corpus and we have limited the number of dimensions of each vector to one hundred. This allows this investigation to be undertaken, and repeated, using fairly normal office-grade hardware. We have experimented with more dimensions (several hundreds), but more vectors appear only to be useful with larger files and require far more computational power.

4.2. Analyzing word vectors

Within each spatial model, we have identified the 250 words with the highest cosine similarity to the Dutch terms for 'war criminal' (singular and plural, see table 1). With these 250 nearest neighbors, we have defined the time-specific vocabulary used in the discussion of war criminals. Obviously, these are not the same 250 words in each model.

To identify changes in the discussions surrounding our topic, we calculated the cosine similarity of each of the 250 nearest-neighbor words in each model to two different terms that are present in each of the two corpora. This allows us to compare the position of the vocabulary of the discussion on our topic (war criminals) in relation to, in this case, two stable concepts. The selection of these concepts is crucial for our investigation and for this method. It is here that we translate our research question into a formal, computational inquiry.

For now, we have chosen a two-dimensional implementation of this technique. This is not theoretically necessary, but it allows us to visualize and analyze results more easily in two dimensions. What is important is that concepts used to investigate the relative position of each investigated word are the same in each of the models to be compared. It is also necessary that the concepts are relatively stable through time. Since concepts are represented by words in the corpus itself, words that shift meaning dramatically, such as the English word 'gay' are less suitable than 'cheerful' or 'homosexual', which have not undergone such dramatic change over time.

When discussing concepts, the number of possible words referring to the same concept is often greater than one. Since our investigation focuses on concepts that may be described with multiple words, we need to create a so-called combined vector. We used synonyms and plurals to create a cluster of words with the shared meaning of the concept of interest. This cluster was used as a combined vector in the model by calculating the mean of all the vectors of the cluster words. That is to say that this word set was treated as a single term, resulting in a vector of similar length to a single-word vector. This combined vector allows us to investigate our corpus using all synonyms and near-synonyms of terms as if they were a single term, with a single vector.

After selecting two concepts that are present in each of the two corpora, we can calculate the relative similarity of other terms in the corpus to each of them. Although vectors between the two trained WEMs are not comparable, the relative distance to two or more other vectors can be compared very well across several models, provided the underlying concepts are historically stable. When the terms used to estimate the relative position of vocabularies are related and dissimilar, or even perfectly opposite, an historically meaningful analysis becomes viable.

Concept	**Concept represented by combined vector of the Dutch words:**
Death penalty	'doodstraf' and 'doodstraffen'
Life imprisonment	'levenslang', 'levenslange', 'vrijheidsstraf', 'gevangenisstraffen', 'gevangenisstraf', 'opsluiting', 'hechtenis'
Treason/traitor	'landverrader', 'landverraders', 'verrader', 'verraders' and 'landverraad'
Victim	'slachtoffer' and 'slachtoffers'
War criminal	'oorlogsmisdadiger' and 'oorlogsmisdadigers'

Table 1: Word sets used in Debating Evil

Using two concepts allows us to plot our 'vocabulary', that is the top 250 war-criminal-related words in each of the two periods, in a two-dimensional space. Figure 1 and 2 show the similarity scores of each of the 250 word vocabularies relative to one concept that serves as the y axis, and another on the x axis. Each point represents one of the 250 words that form the war-criminal vocabulary for a specific time period. They are plotted based on their cosine similarity score to the combined vector of the concept 'victim' (x) and 'treason' (y) in figure 1, and to 'life imprisonment' (x) and 'death penalty' (y) in figure 2. The average scores of

all 250 war criminal words on the two dimension are shown as horizontal and vertical lines. Thus, we have arrived at a visual representation that allows for a comparison of word embedding results for more than one corpus and hence for a comparison through time. (In this case, between two distinct periods.)

5. Results

Here, we present only two examples using four concepts and two time periods (1945-1955 and 1965-1975). Specifically, we try to identify differences in the way incarcerated war criminals and collaborators were discussed in the immediate aftermath of the Nazi occupation of the Netherlands, and at the height of controversies surrounding the intended release of a number of German war criminals from Dutch prisons - namely Kotälla, Aus der Fünten, and Fischer (Piersma, 2005).

Obviously, the discussions in the two periods refer to different groups of perpetrators. In the immediate aftermath of the Nazi occupation the population of inmates was large and diverse, consisting of small-time war profiteers, minor collaborators and their families, but also mass murderers. In the second period, only a handful of elderly foreigners were left, whose crimes were relatively similar and also similarly egregious.

For this investigation, however, our primary aim is not to unearth radically new insights into postwar penal policy in the Netherlands, but to confront the results of an unsupervised, 'distant' reading of parliamentary records to an established historiography. Such an historiography is available for the case at hand; Dutch historians have identified a number of trends in the thinking about political delinquents that (if true) should be reflected in these discussions. Two changes have been identified in particular:

1. The shifting focus from the nature of the crime committed and the person of the perpetrator towards the lasting, psychological damage endured by the victims (de Haan, 1997; van der Heijden, 2011).

2. A decline in the support, both public and political, for harsh, vengeful punishments, exemplified here in the discussions about the propriety of the death penalty. Although the death penalty was (again) abolished in the 1950s, it remained a point of discussion with regard to war criminals in custody. (Futselaar, 2015; Smits, 2008).

5.1. Historical case

Over the course of three decades, attitudes to incarcerated war criminals, as represented by the vocabularies used to discuss them, changed. In the first period the emphasis lay on crimes against the collective, whereas the focus changed to the plight of individual victims. As can be seen in figure 1, the initial emphasis on crimes against the nation (treason) in debates about war criminals clearly declined. The average cosine similarity between war-criminal words and treason words (horizontal lines) decreased significantly when we compare 1945-1955 to 1965-1975. At the same time, we observed increased levels of closeness in vector space between war criminal related words to words associated with (individual) victims, as can be seen in figure 1.

This observation is in line with the relevant historiography. Several authors have emphasized the sharp rise of interest into the mental health of individual war victims and and their families as a decisive factor in policy making and the formation of political opinion. This also indicates a shift in discourse from focusing on the initial crimes, committed by the war criminals, to the consequences of their deeds for individual people involved (de Haan, 1997; van der Heijden, 2011; Smits, 2008; Withuis, 2002).

This development can not be considered a mere discursive change: the observed shifts in parliamentary vocabulary represent actual historical developments in the postwar dealing with war criminals. In the early 1970s, the only war criminals remaining in Dutch prisons were German nationals. Whereas in 1945, main part of the more than hundred thousand incarcerated war criminals were Dutch citizens. Evidently, the accusation of treason was only applicable to the latter group. Hence, if we compare the two periods, it is not surprising that the discursive element of 'treason' evaporates from the war criminal vocabulary in Dutch parliamentary debate between 1965 and 1975.

It remains imperative to remain aware of the possible pitfalls of this type of investigation. This is evident in the sharp rise of references to the death penalty in war criminal vocabulary that we observed (see figure 2). During the second period under scrutiny, capital punishment had long been discontinued in the Netherlands and could not have been discussed as a serious penal option. Closer scrutiny of the data revealed that in many discussions, capital punishment was not advocated, but merely used as a reference point. The war criminals in question had originally been condemned to die, but their punishment had been commuted into life imprisonment. Several members of parliament felt that a pardon would mean that the original verdict (death penalty) would be watered down twice. In these discussions, capital punishment was often referenced, even when its use was not a viable (or even legal) option.

6. Conclusion

This paper outlines a method for studying discursive changes in history. We trained WEMs and calculated cosine similarities between two opposite or related concepts for specific periods. This enabled us to compare WEMs for different periods. This opens the door for the use of word embeddings as a tool for historical research, because it enables us to investigate change through time in sufficiently large and consistent historical datasets. Parliamentary records are perhaps the best example of such datasets. As such, this method holds considerable promise in a period when parliamentary proceedings and other historical sources are increasingly made available in machine readable form.

We have shown how developments in vocabulary can be considered reflective of discursive changes. These changes coincide with related historical events and developments in the post-war dealing with war criminals in Dutch society. The war criminal vocabulary shifted from focusing on the act of crime committed by war criminals towards the consequences of these deeds for victims and relatives. We also showed how actual historical developments regarding the

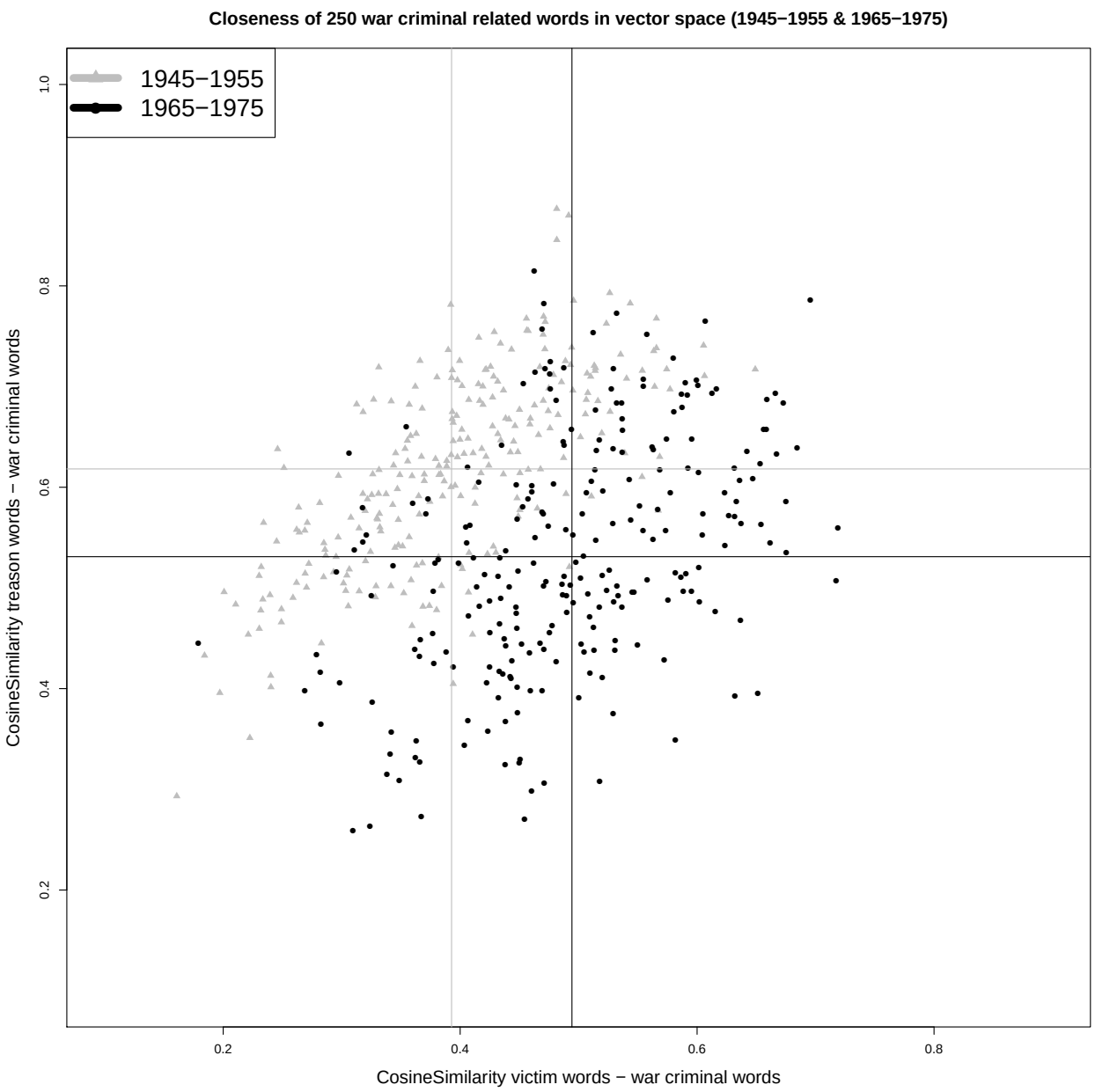

Figure 1: Top 250 war criminal related words 1945-1955 (grey) and 1965-1975 (black) plotted by their cosine similarity to victim (x) and traitor (y) words.

type of war criminals incarcerated in the Netherlands were reflected by a discursive shift, in which closeness to 'treason' declines and gave way to an increasing focus on victims in debating war criminals.

We have also encountered examples of pitfalls of an overly enthusiastic reliance on word embeddings as an analytical tool. Capital punishment was mentioned particularly frequently in the 1970s, but not because the possibility of executing the war criminals was seriously entertained. Distributional semantics are a powerful new tool for historians, but they do not not remove the need for hermeneutical awareness.

In this paper, the method is itself the main object of inquiry. We believe we have shown that it possible, feasible, and useful to develop and implement a coherent and widely applicable method for investigating historical change using WEMs.

7. Discussion

7.1. Method evaluation

For this paper, we have used two corpora of ten years to train our WEMs on. More interesting, from a research perspective, would be to find out how stable our results are when using smaller, overlapping windows of corpora over time, say with one-year steps. It is likely (but not certain) that using more fine-grained windows will reveal similar developments and shifts in language use over time. Repeating the analysis with more data points has the potential to gain more insights in the graduality and the pace of the

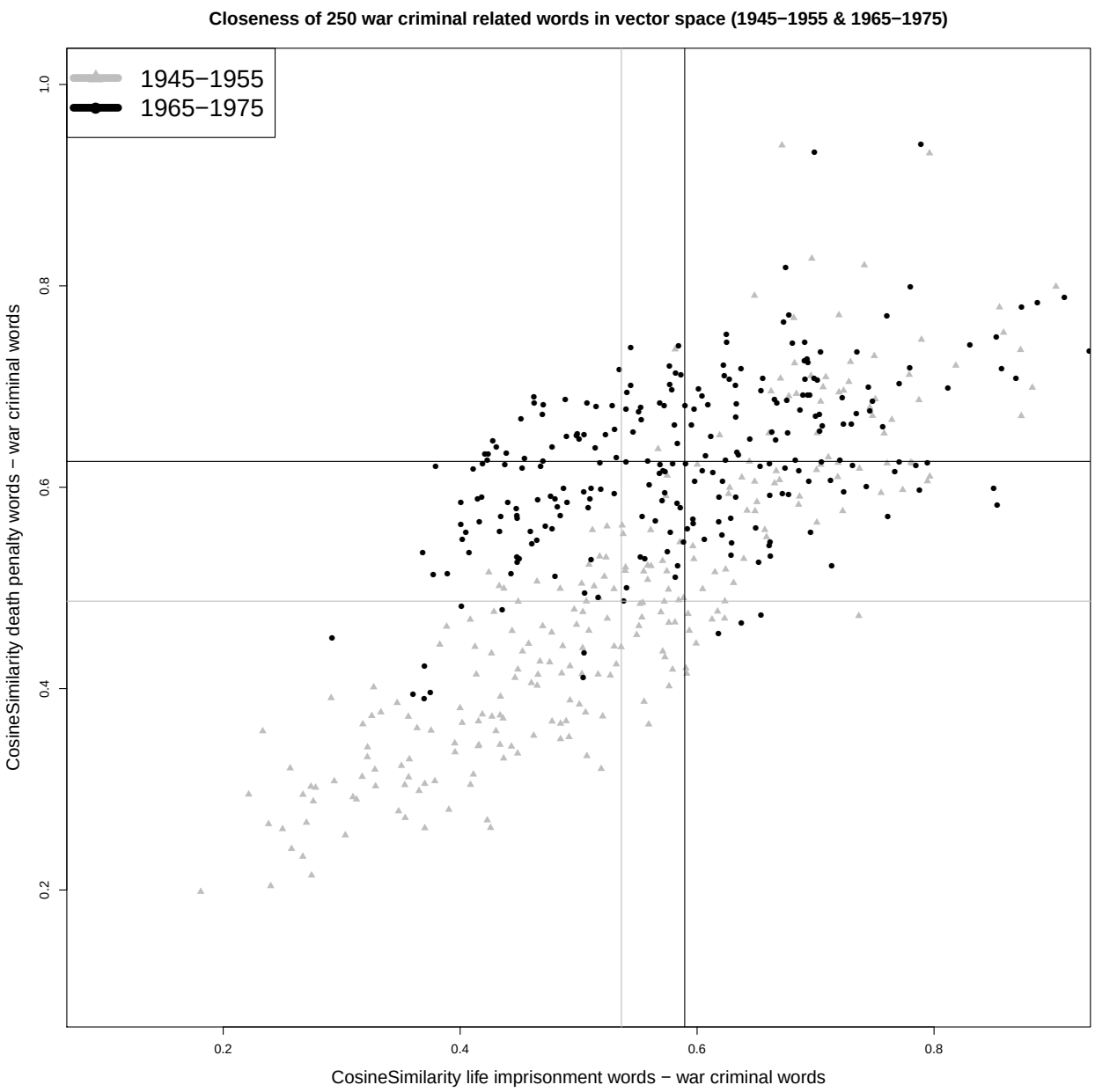

Figure 2: Top 250 war criminal related words 1945-1955 (grey) and 1965-1975 (black) plotted by their cosine similarity to life imprisonment (x) and death sentence words (y).

observed shifts in language used. That said, there is a potential trade-of between detail and precision given that the size of the corpora available to historians are mostly modest in size.

A second ambition is to look more seriously into the distribution of the cosine similarity scores, and the changes in these distributions over time. It will be interesting to measure, visualize, and statistically evaluate these distributions more closely, and to see whether they can be linked to, for example, unanimity and/or homogeneity in parliamentary discussions.

7.2. Historical evaluation

Another remaining ambition is to compare the parliamentary vocabularies used to discuss 'domestic' collaborators and foreign (usually German) war criminals. Furthermore, we also hope to position the war criminal debates in a broader context: how distinct are they from other war-related debates, and from other discussions about penal law or criminals in a more general sense?

Just as a closer investigation of different categories of perpetrators is viable and useful, different groups of war victims who were discussed in parliamentary debates also license further investigation. These may have included first and second generation victims of wartime violence and persecution, former forced labourers, holocaust survivors and the children of holocaust victims, etc. Given the emphasis on the protection of war victims mentioned above, we are interested to see if there have been changes in the groups emphasized in political speech about the topic.

8. Acknowledgements

We are grateful to the participants of our Text Mining workshop at the Luxembourg Centre for Contemporary and Digital History (C2DH) in Esch-sur-Alzette (June 2018), for their comments, input, and criticism.

9. Bibliographical References

Peter Bootsma and Peter van Griensven. 2003. 'teleurstelling is mijn opperste emotie': Vragen over emotie in de politiek aan a.a.m. van agt. In C. van Baalen et al., editor, *Jaarboek Parlementaire Geschiedenis, 2003. Emotie in de Politiek*. SDU Uitgevers, Den Haag.

Ido de Haan. 1997. *Na de ondergang: de herinnering aan de Jodenvervolging in Nederland 1945-1995*. Nederlandse cultuur in Europese context. Sdu Uitgevers.

Ralf Futselaar. 2015. *Gevangenissen in oorlogstijd 1940-1945*. Boom uitgevers Amsterdam.

Alexander Gelbukh. 2015. *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings*. Number dl. 1 in Lecture Notes in Computer Science. Springer International Publishing.

Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. Good applications for crummy entity linkers? the case of corpus selection in digital humanities. *CoRR*, abs/1708.01162.

Hinke Piersma. 2005. *De Drie Van Breda: Duitse oorlogsmisdadigers in Nederlandse gevangenschap, 1945-1989*. Balans, Amsterdam, 1st edition.

Benjamin Schmidt. 2015. Vector space models for the digital humanities. *Ben's Bookworm Blog*, oct.

Amit Singhal. 2001. Modern information retrieval: a brief overview. *BULLETIN OF THE IEEE COMPUTER SOCIETY TECHNICAL COMMITTEE ON DATA ENGINEERING*, 24:2001.

Hans Smits. 2008. *Strafrechthervormers en hemelbestormers: opkomst en teloorgang van de Coornhert-Liga*. Aksant, Amsterdam, 1st edition.

Ismee Tames. 2013. *Doorn in het vlees: foute Nederlanders in de jaren vijftig en zestig*. Balans, Amsterdam.

Chris van der Heijden. 2011. *Dat nooit meer: de nasleep van de Tweede Wereldoorlog in Nederland*. Atlas Contact, Uitgeverij.

Jolande Withuis. 2002. *Erkenning: van oorlogstrauma naar klaagcultuur*. De Bezige Bij, Amsterdam.