Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990

Melvin Wevers
DHLab - KNAW Humanities Cluster
melvin.wevers@dh.huc.lab.nl

Introduction

This paper examines gender bias in word embeddings of historical newspapers by comparing the strength of association between male and female dimensions of gender on the one hand, and words that represent occupations, psychological states, or social life, on the other. After the Second World War, Dutch society was stratified according to ideological and religious "pillars", a phenomenon known as pillarization. Newspapers were often aligned to one of these pillars [3]. The newspaper *Trouw*, for example, has a distinct Protestant origin, while *Volkskrant* and *De Telegraaf* can be characterized as, respectively, Catholic and neutral. **Did newspapers associated with specific pillars exhibit particular gender biases with respect to particular aspects of society, behavior, or culture?**

Materials and Methods

The data set consists of six Dutch national newspapers with ideological backgrounds ranging from liberal, social-democratic, neutral/conservative, Protestant, to Catholic. We incorporate external lexicons as target words: (1) approximately 12.5k job titles from the HISCO data set (2) Positive and Negative emotion words from Cornetto. (3) The Dutch translation of LIWC2001, which contains lists of words to measure psychological and cognitive states

Per newspaper, we train four word embedding models, one per decade between 1950 and 1990.[1] The size of the vocabulary approximately doubles for some newspapers between 1950 and 1990 (Figure 1). The variance of the targets words, however, was small ($\mu \approx 0.003$) and constant ($\sigma[1.3^{-9}, 2.9^{-9}]$), indicating model stability.

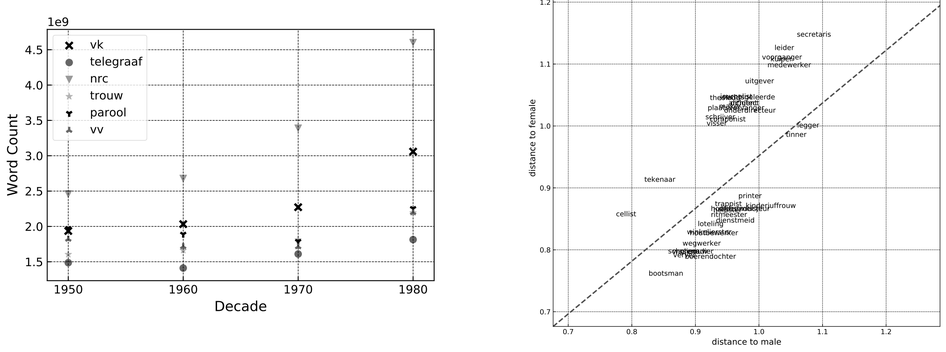

Figure 1: Left: vocabulary size; Right: gender Bias related to job titles

Following [2], we construct two vectors representing the gender dimensions, by creating an average vector that includes words referring to male ('man', 'his', 'father', etc.) or female as well as the most popular first names in the Netherlands for the period 1950-1990. Next, we calculate the distance between each gender vector and every word in a list of target words from our lexicons, for example, words that denote occupations [1]. The difference between the distances for both gender vectors represents the gender bias (Figure 1). Lastly, We apply Bayesian linear regression to determine whether the bias changed over time. The linear model is formulated as:

$$\mu_i = \alpha + \beta * Y_i + \epsilon,$$

with μ_i the bias for each decade (i) and Y_i the coefficient related to each decade (i). The likelihood function is: $X \sim \mathcal{N}(\mu, \sigma)$ with priors defined: $\alpha \sim \mathcal{N}(0, 2)$, $\beta \sim \mathcal{N}(0, 2)$, and $\epsilon \sim \text{HalfCauchy}(\beta = 1)$. We compute a combined linear model based on all newspapers for the all the target words groups. Then, for the same categories, we compute individual linear models for each newspaper.

Results

The combined linear models generally display minimal shifts in bias. Partly, the weak trends are related to opposing shifts in the individual newspapers, cancelling each other out. Nonetheless, the bias associated with the categories 'TV', 'Music', 'Metaphysical issues', 'Sexuality' navigate toward women, with all of them starting from a position that was clearly oriented toward men. Conversely, 'Money', 'Grooming', and Negative Emotion words move toward men, which in the 1950s were all more closely related to women.

For the category Job Titles, we see a slight move toward women, while words from the LIWC category Occupation move marginally in the direction of men. This suggests that job titles might be more closely related to women, while the notion of working gravitates toward men.

[1]Code: `https://github.com/melvinwevers/historical_concepts`
Word Embedding Models: `http://doi.org/10.5281/zenodo.3237380`

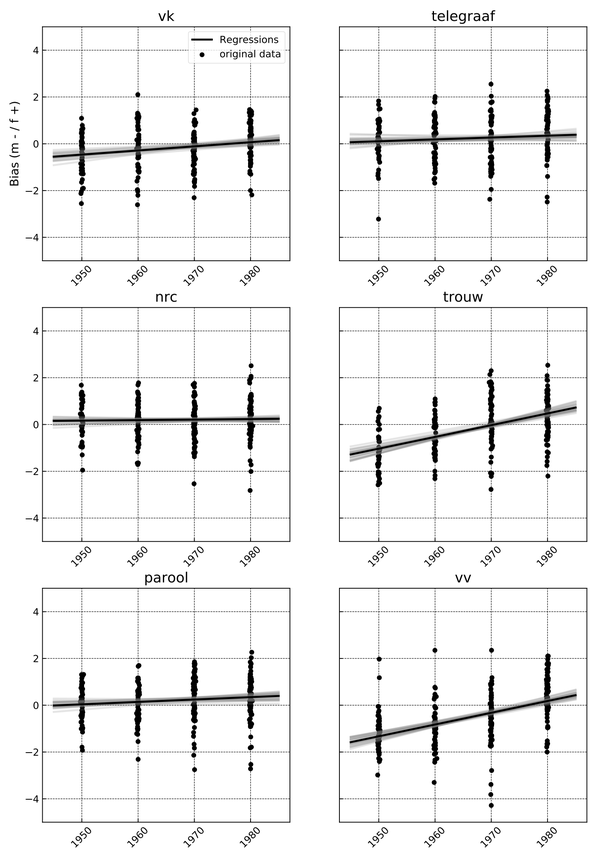

Figure 2: Individual newspaper model 'Sexuality'

The linear models for the individual newspapers demonstrate distinct differences between the newspapers (Figure 2). First, *Volkskrant* is the most stable newspapers with 56% of the target word categories not changing. When bias changes in this newspaper, it moves toward women 9 out the 11 changing categories. *Telegraaf*, *NRC*, and *Parool* generally move toward men, respectively (84%, 92%, and 80% of categories). The biases of *Trouw* and *Vrije Volk*, contrarily, move toward women (both 72%). A noteworthy result is that in all newspapers the bias shifts toward men in the category 'money'. Moreover, they also all exhibit a move toward women for the category 'sexuality', with the clearest shift in *Volkskrant*, *Trouw*, and *Vrije Volk*.

Conclusions

1. While newspaper discourse as a whole is fairly stable, individual newspapers show clear divergences with regard to their bias and changes in this bias.
2. In relation to themes such as sexuality and leisure, we see the bias moving toward women, whereas, generally, the bias shifts in the direction of men.
3. Even though Dutch society became less stratified ideologically (depillarization), we found an increasing divergence in gender bias between religious and social-democratic on the one hand and liberal newspapers on the other.
4. newspapers with a social-democratic (*Vrije Volk*) and religious background, either Catholic (*Volkskrant*) and Protestant (*Trouw*) demonstrate the clearest shift in bias toward women.
5. The liberal/conservative newspapers *Telegraaf*, *NRC Handelsblad*, and *Parool*, on the contrary, orient themselves more clearly toward men.
6. Despite increasing female employment numbers in the Netherlands, the association with job titles moves only gradually toward women, while words associated with working move toward men.

References

[1] Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518. ACM, 2017.

[2] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):3635–44, April 2018.

[3] Huub Wijfjes. *Journalistiek in Nederland, 1850-2000: beroep, cultuur en organisatie*. Boom, Amsterdam, 2004.

Acknowledgements

I would like to thank Folgert Karsdorp for his feedback. This research was part of the project "Digital Humanities Approaches to Reference Cultures: The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990", which was funded by the Dutch Research Council (NWO)