

(De)constructing Binarism in Journalism: Automatic Antonym Detection in Dutch Newspaper Articles

Alie Lassche^{1,†}, Ruben Ros^{1,†} and Joris Veerbeek^{2,†}

¹Leiden University, Institute of History, Doelensteeg 16, 2311 VL Leiden, The Netherlands

²Utrecht University, Department of Media and Culture Studies, Drift 13, 3512 BR Utrecht, The Netherlands

Abstract

Binary oppositions, since their introduction by Claude Levi-Strauss and other structuralists in the seventies, are under pressure, especially because they legitimize societal power structures. Deconstruction of binary oppositions such as man/woman, black/white, left/right, and rich/poor is therefore increasingly encouraged. The question arises of what kind of effect the debate about binary oppositions has had on its linguistic use. We have therefore detected antonyms in a corpus of Dutch newspaper articles from the period 1990-2020, to study the development of binarism in journalism. Our method consists of two parts: the use of a good-old lexicon, and the finetuning of a BERT model for antonym detection. In this paper, we not only present our results regarding the (de)construction of binary oppositions in Dutch journalism, but we also reflect on the two methodological stages and discuss their gain.

Keywords

antonym detection, BERT, binary opposition, journalism, newspapers

1. Introduction


According to Claude Levi-Strauss and other structuralists, binary oppositions form the basic structure of all human cultures. Everyone everywhere thinks and structures their worlds in terms of pairs of opposites (raw/cooked, light/dark, left/right, man/woman) [8]. The binary opposites are considered inseparable in their opposition because the one term only has meaning as the negation of the other term [7]. Furthermore, in every pair, one term is favoured over the other by being marked as positive, while the other term is considered negative [3]. Because of these characteristics, binary oppositions have been criticized over the last sixty years in areas such as deconstructionism, post-structuralist feminist theory, queer theory, post-colonialism, and critical race theory. Moreover, in the public debate, criticizing the legitimization of societal power structures in which a specific majority is favoured has become more important than ever. Deconstruction of binary oppositions such as man/woman, black/white, left/right, and rich/poor is increasingly encouraged. However, this does not have to mean that the use of binary opposition is diminishing: the debate could just as well lead to an increase in its usage. The question arises of what kind of effect the debate about binary oppositions has had on its


CHR 2023: Computational Humanities Research Conference, December 6–8, 2023, Paris, France

[†]These authors contributed equally.

✉ a.w.lassche@hum.leidenuniv.nl (A. Lassche); r.s.ros@hum.leidenuniv.nl (R. Ros); j.veerbeek@uu.nl (J. Veerbeek)

ORCID 0000-0002-7607-0174 (A. Lassche); 0000-0002-5303-2861 (R. Ros); 0000-0001-5110-0720 (J. Veerbeek)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

linguistic use. How has the discourse surrounding the critique and deconstruction of binary thinking influenced the utilization of binary oppositions in texts?

This paper aims to detect binary oppositions in a corpus of Dutch newspapers and analyse patterns in their use in Dutch journalism between 1990 and 2020. We chose this genre and period because we expect a change here in the use of binary oppositions, as a reflection of the developments in its public debate. To operationalize the detection of binary oppositions, we look for antonyms in our corpus. Equating these two concepts has important consequences: while an antonym pair consists of two words that have opposite meanings, a binary opposition consists of two opposing words that often have a connotation of contrast, conflict, or tension. Therefore, not every antonym pair is a binary opposition. In the remainder of this paper, we will further reflect on this methodological choice. The method we propose consists of two layers, and can thus be considered a two-stage rocket: we start with creating a good-old lexicon of antonym pairs in the Dutch language. Afterwards, we use this as training data to finetune a BERT model for automatic antonym detection. In what follows, we will discuss relevant related work concerning the extraction of antonyms from text. We will then discuss the methodological pipeline we propose in more detail. We will not only present our results regarding the use of binary oppositions in Dutch journalism, but we will also reflect on the two methodological stages, and discuss the gain of each of them.

2. Related work

Detecting antonyms – words with an opposite meaning – is a task often undertaken by linguists. However, differentiating antonyms from synonyms is challenging due to similar usage. Linguists use pattern-based and co-occurrence models to distinguish them. *Pattern-based models* assume that antonymous word pairs co-occur in some antonym-indicating lexico-syntactic patterns. Examples are patterns such as *from A to B*, *between A and B*, and *either A or B*. Roth and Schulte im Walde combined patterns with discourse markers for classifying paradigmatic relations between words such as synonymy and antonymy. Schwartz, Reichart, and Rappoport presented a symmetric pattern-based model for word vector representations in which antonyms were assigned to dissimilar vector representations. More recently, a novel pattern-based neural method *AntSynNET* to distinguish antonyms from synonyms was presented [9].

In *co-occurrence models*, each word is represented by a weighted feature vector, where features typically correspond to words that co-occur in particular contexts. Yih, Zweig, and Platt introduced a vector space representation where antonyms were positioned on opposite sides of a sphere. Scheible, Schulte im Walde, and Springorum showed that the differences in the contexts of synonymous and antonymous pairs could be identified with a simple word space model. Santus, Lu, Lenci, and Huang introduced an Average-Precision-based measure for the unsupervised discrimination of antonymy from synonymy. They argued that synonyms are expected to have a broader and more salient intersection of their top-K salient contexts than antonyms.

The recent introduction of pre-trained large language models such as BERT has largely improved how the task of antonym detection is addressed. In a recent paper by Church, Cai, and Bian, an earlier proposed mixture of experts (MoE) method [14] was combined with dLCE

Table 1

Characteristics train, validation, and test data.

	train	test	val	total
synonym pairs	1705	197	207	124,740
antonym pairs	1659	220	230	2,109
random pairs	16882	2114	2094	N/A

embeddings [9]. Its performance was compared with the performance of a BERT model that was finetuned using different datasets. The highest performance (0.947) was gained with a model that was trained and tested on a subset of Samuel Fallows’ thesaurus of synonyms and antonyms [5].

3. Corpora and methods

The methodology we use is inspired by the work of Church, Cai, and Bian on finetuning a BERT model for antonym detection. However, since we want to discuss the gain of using a large language model in this task, we will compare its outcome with the more basic approach of using a lexicon of antonym pairs, in order to find binary oppositions. We thus propose the following methodology to detect antonyms in Dutch text corpora:¹

- 1. Preparing a word list with antonyms and synonyms.** We extracted antonymous and synonymous word pairs from the website www.mijnwoordenboek.nl and the online dictionary Van Dale.² To ensure a balanced representation of these two classes, we down-sampled the synonym pairs. Additionally, to enable the model to discern when two words are unrelated in both antonymous and synonymous senses, we introduced random word pairs. These random pairs form the majority class, as we assume that the majority of word pairs in our data are not related. To achieve this, we sampled these random pairs at a rate ten times the size of our synonym and antonym pairs, as shown in Table 1.
- 2. Finetuning BERT model for antonym detection.** We tested five different BERT models, both Dutch and multilingual.³ After following Devlin, Chang, Lee, and Toutanova for standard hyperparameter tuning, which entails optimizing the learning rate, epochs, and batch size, the multilingual model `mdeberta-v3-base` [6] yielded the highest performance, hence our choice to continue with this model. Overall, the model achieved an accuracy of 0.90 on the test set. On the antonym class specifically, the model achieved an *f1*-score of 0.79.
- 3. Preparing word pairs from newspaper data set.** Our dataset consists of articles from the Dutch newspaper *NRC* from the period 1990-2020, which are all available on their

¹Please refer to the git repository for the full code: <https://github.com/rubenros1795/antonym-detection>.

²www.vandale.nl. Van Dale does not provide any information on selection criteria for their synonym and antonym dictionary. Mijn Woordenboek states on its website that synonyms are licensed from Van Dale, Kernerman Dictionaries, and Interglot, unless otherwise specified. An active group of volunteers and users continuously contributes and verifies words.

³These include: `bert-base-dutch-cased`, `robert-v2-dutch-base`, `xlm-roberta-base`, `mdeberta-v3-base`, and `bert-base-multilingual-cased`.

Table 2
Characteristics newspaper data set.

	#
articles	589,739
words	346,909,375
filtered words (N, Adj)	93,278,902
word pairs	25,248,589

website.⁴ *NRC* is among the major newspapers in the Netherlands, with a liberal orientation. Formerly known as *NRC Handelsblad*, it has a strong focus on business and international news. We preprocessed the data by excluding advertisements, job postings, and news index pages. We kept only nouns and adjectives that occur more than 10 times in the full corpus. We then created pairs for every possible combination between two words within one sentence. This resulted in 25,248,589 unique word pairs. To reduce the number of word pairs that had to be annotated by the model, we used a threshold of 0.4 for the cosine similarity between the words in a pair, leaving 2,471,340 to be annotated by the model. We trained a `fastText` model on our dataset to define these similarities.⁵ Characteristics of the corpus and the word pairs can be found in Table 2.

4. **Annotating antonyms in newspaper word pairs.** We applied our finetuned BERT model to the newspaper word pairs. We considered two given words antonyms if the probability was higher than 50%. Although our model consists of three classes, we are only interested in antonyms in our analysis. To further limit the number of false positives, we opted for this threshold of 50%, which is conventional in a binary classification task. This resulted in 128,294 word pairs that were classified as antonyms by our model.

4. Results

4.1. Binary oppositions in newspapers

In Figure 1, we present the averaged adjusted PMI for the antonym pairs extracted from the corpus of newspaper articles. There appears to be a modest decline in the co-occurrence of antonym pairs between 1990 and 2020. We use Pointwise Mutual Information (PMI) as a target metric for estimating the joint probability of an antonym pair appearing together [2]. We use the sentence as the context for establishing whether the antonym words are used together. However, as shown in Figure 2, sentence length decreases monotonically over time. The figure also shows a steep decrease in article length until 2012, and a slow increase in the number of sentences within an article. In other words: sentences become shorter over time, while articles first become shorter, but longer after 2012, due to more sentences within articles.⁶ The decrease

⁴www.nrc.nl.

⁵The average cosine similarity for antonyms and synonyms was 0.44.

⁶The peaking article length in 2020 is largely caused by the appearance of the so-called *corona live blogs*, a daily live blog in which all COVID-related news was collected.

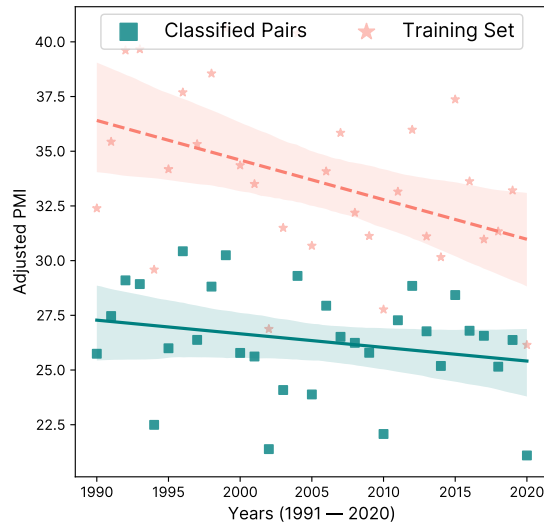


Figure 1: Mean adjusted Pointwise Mutual Information for antonyms in the newspaper data. The red line shows the scores for the antonym pairs in the training set.

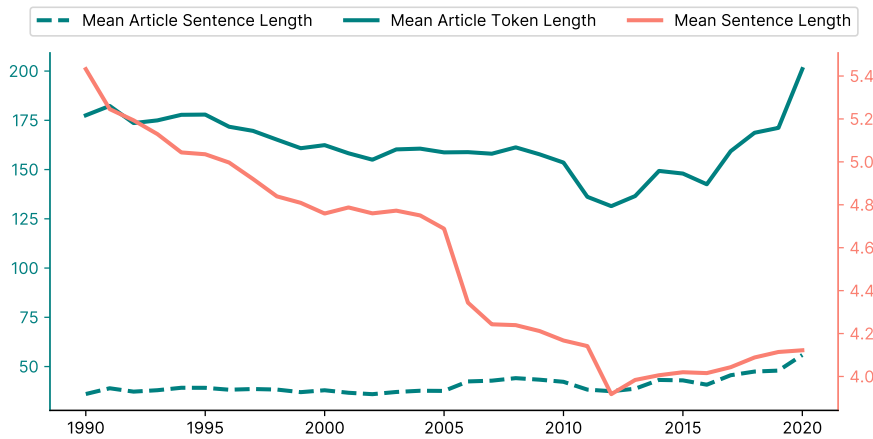


Figure 2: Average article length, sentence length, and number of sentences in articles, aggregated per year. The left y-axis is related to the average number of sentences and the average article length, while the right y-axis is related to the average sentence length.

in sentence length is likely to affect antonym PMI scores. To address this issue, we augment the PMI values with a decay function, which can be found in the appendix.

The antonym pairs with our classifier’s highest score are included in the appendix in Table 3. Three dominant clusters can be detected when looking at these most frequent antonym pairs and pairs lower on the frequency list. There is a clear cluster of economic/financial word pairs, including *euro-procent*, *euro-dollar*, *procent-totaal*, and *jaar-kwartaal*. Secondly, there are word pairs related to geopolitics: *europes-amerikaans*, *amerikaans-nederlands*, *amerikaans-iraaks*, *russisch-oekraïens*, *turks-syrisch*, and *amerikaans-russisch*. A third cluster includes antonym

pairs related to social topics and relations: *vrouw-man*, *vader-moeder*, *jong-oud*, *zoon-vader*, *meisje-jongen*, and *kind-ouder*.

Figure 3 shows the adjusted PMI of the 15 highest-scoring antonym pairs that appear in at least ten years.⁷ By filtering on word pairs occurring over a minimum period of ten years, we aim to exclude word pairs pertinent to brief or ephemeral events within the domains of politics, economics, or society. Almost all of these word pairs are from the political domain. Several refer to continuous political relations of the United States (*amerikaans-koreaan*s, *amerikaans-italiaans*, *amerikaans-japans*, *amerikaans-mexicaans*). Other word pairs are more conceptual, yet undeniably stem from the realm of political discourse: *illegaal-legaal*, *tegenstander-voorstander*, *meerderheid-minderheid*, *binnenlands-buitenlands*, *integratie-immigratie*, *conservatief-progressief*, and *internationaal-regionaal*. It demonstrates that the employment of antonyms predominantly prevails in political contexts.

We are furthermore interested in antonym pairs that show a clear development over time. We therefore applied the Mann-Kendall test to detect the pairs with the most consistent monotonic upward and downward trend. In Figure 4, the development of the 15 antonym pairs with the clearest decrease is shown.⁸ Two of these pairs belong to the earlier defined geopolitical cluster of word pairs: *amerikaans-israëli*sch and *amerikaans-japans*. The decrease of the word pair *zwart-blank* (black-blank) represents the shift from the last years to replace the word *blank* with *wit* (white). The word *blank* sounds more positive in comparison to *zwart*, while *wit* and *zwart* are considered neutral. Two other pairs similarly show the decline of political concepts frequently juxtaposed in the twentieth century, such as *katholiek-protestant* and *werkgever-werknemer*.

The 15 antonym pairs with the most pronounced upward trend are shown in Figure 5.⁹ Selecting the top 15 thus results in trends marked by limited increase. The decrease of one pair does not seem to directly increase another pair. It reflects the declining trend as depicted in Figure 1. What stands out is that several of these word pairs are related to the third cluster we distinguished earlier, with word pairs concerning social topics and relations: *school-ziekenhuis*, *baby-ouder*, and *eigen-collectief*.

Apart from the (trans)national events and developments to which these patterns can be linked, they might also reflect a change in the journalistic style and scope of the newspaper *NRC*. In 2006, as an addition to the evening newspaper *NRC Handelsblad*, the morning newspaper *nrc.next* was launched, which targeted younger readers. In 2017, *NRC Handelsblad* and *nrc.next* became two editions of the same newspaper, respectively a morning and afternoon edition. Since 2022, only the morning edition exists, bearing the name *NRC*. The website from which our corpus originates includes articles from *nrc.next*, *NRC Handelsblad*, and *NRC*. The sharp decrease in sentence length (Figure 2) coincides with the launch of *nrc.next* in 2006. This, together with the earlier decline in co-occurrence of antonym-pairs (Figure 1), suggests that targeting a younger audience results in a change in journalism that is both reflected in style and content. Based on the qualitative review of the declining pairs, we suspect that *nrc.next* was part of a more general decline of political-economic coverage relative to issues around

⁷The translations of these antonym pairs are listed in Table 4.

⁸The translations of these antonym pairs are listed in Table 5.

⁹The translations of these antonym pairs are listed in Table 5.



Figure 3: Adjusted PMI over time for the 15 antonym pairs with the highest score in the classification model.

lifestyle and social issues, which resulted in a decline of antonym pairs in the former area.

Finally, in Figure 6, we have visualized how the development of a certain word pair relates to the development in frequencies of the distinctive words in that pair. We normalize the Adjusted PMI scores, as well as the relative frequencies for both words in the pair between 0 and 1 for comparability. For most antonym pairs, these time series assume the shape of the *eigen-ander* pair, with a high positive correlation between $P(w1) / P(w2)$ and *PMI*. The example above of a decreasing *PMI* between *zwart* and *blank* occurs in the context of the general decrease in the frequency of both terms. There is only one pair that stands out with a clear negative correlation: the *vrouw-man* pair. The words in this pair show an upward trend. Their joint appearance,

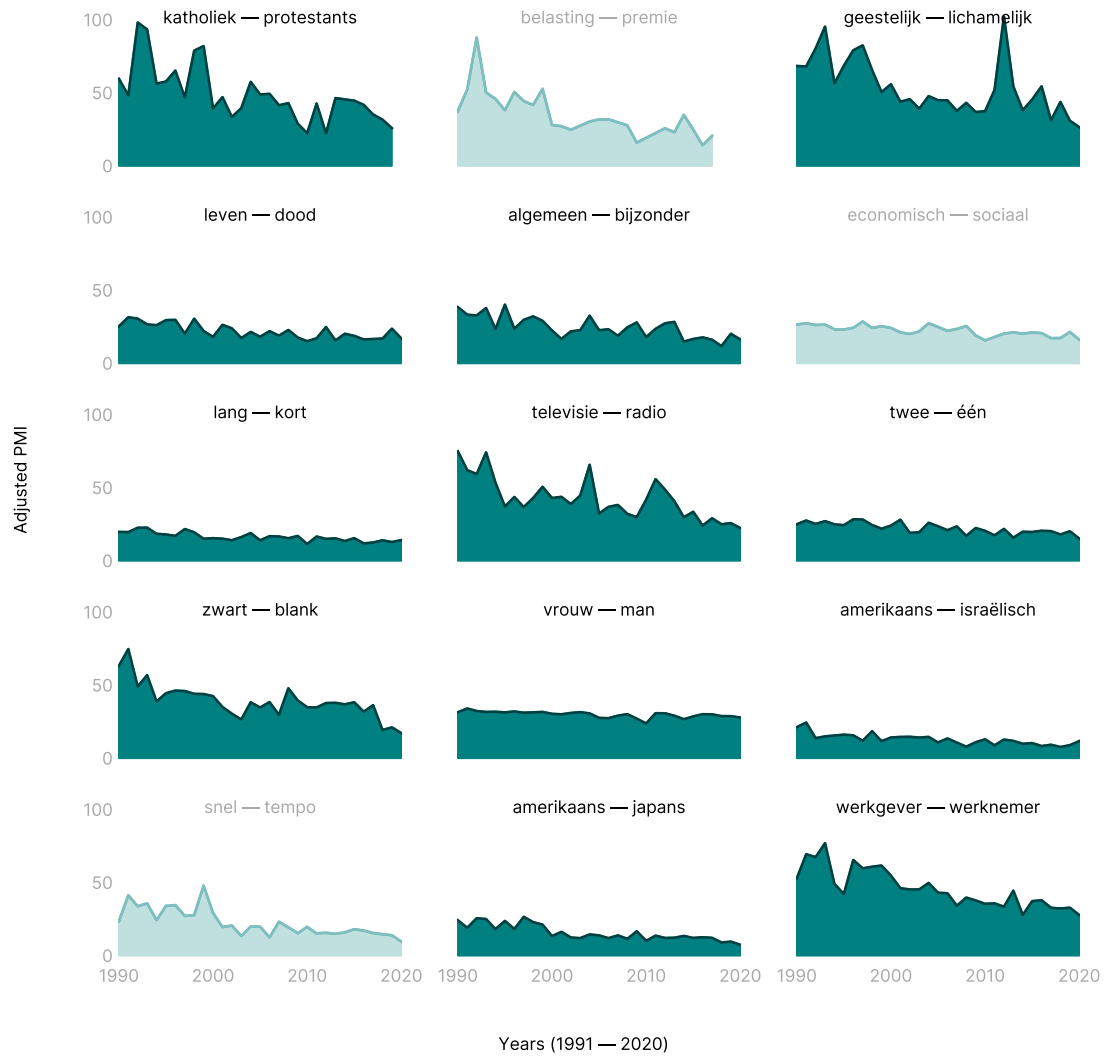


Figure 4: Adjusted PMI over time for the top 15 most decreasing antonym pairs. Decrease is measured with the slope of the regression line. We select the top 15 based on p -values derived from the Mann-Kendall. Pairs manually identified as false positives are plotted in a lighter tint.

however, decreases over time, which means that both terms are increasingly used separately, as visible in a Pearson correlation coefficient of -0.49 between $P(w1)$ and PMI . A similar but weaker divergence is visible in the *rechts-links* pair and the *vader-kind* pair. Although this does not show the end of binary oppositions, specific cases like these do clearly decline over time.

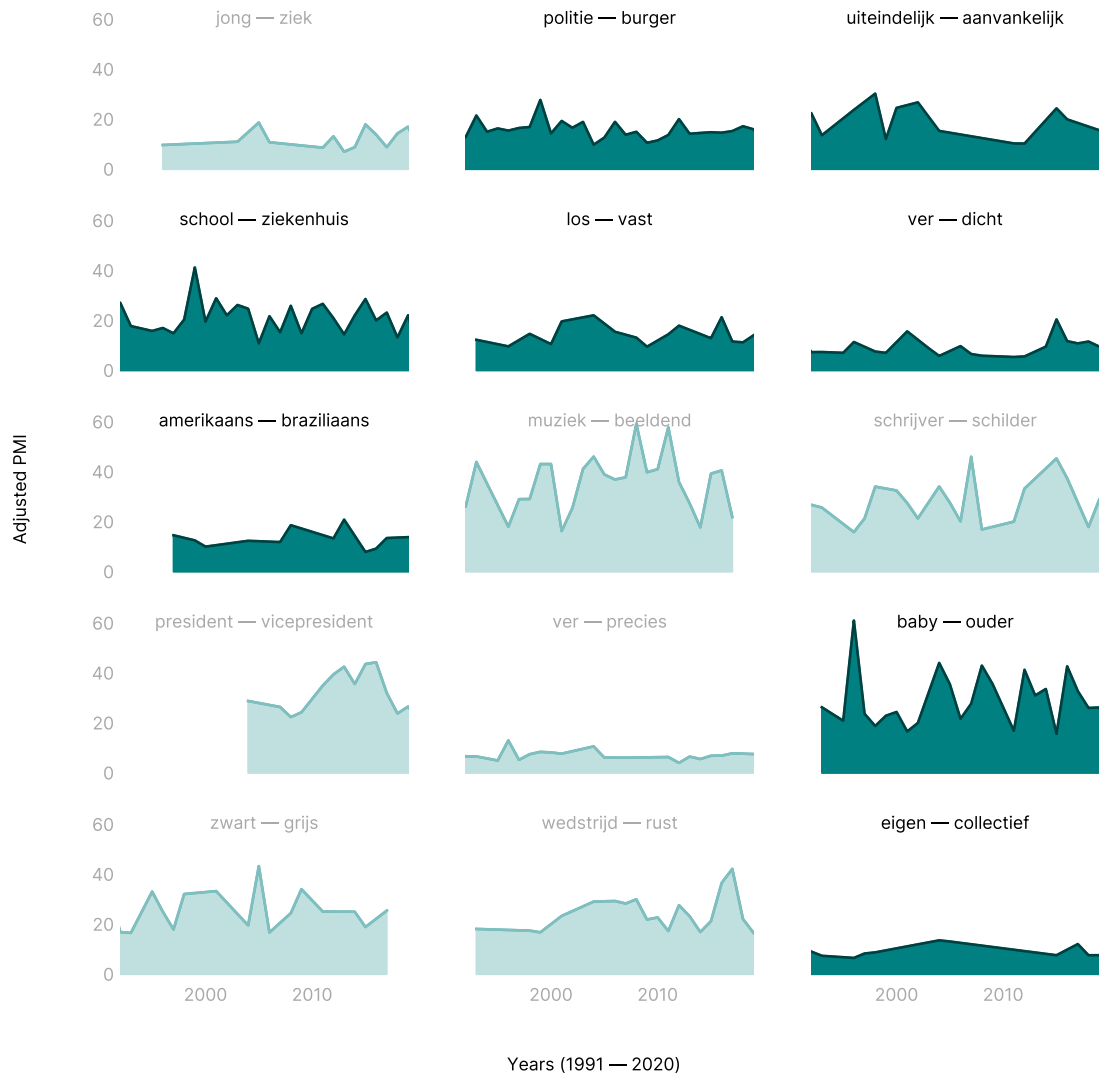


Figure 5: Adjusted PMI over time for the top 15 most increasing antonym pairs. Pairs manually identified as false positives are plotted in a lighter tint.

4.2. Methodological considerations

The utilization of BERT models enabled us to discover a broad spectrum of antonyms. At the same time, we also observed a notable incidence of false positives during our analysis. Therefore, in this paragraph, we reflect on the methodological gains and pitfalls of using a machine-learning approach to discover antonyms in texts and contrast that with the use of a simple lexicon.

Our initial list of antonyms consisted of 2,109 pairs. Using the methodology we presented in this paper, we found 128,294 unique pairs of antonyms in our corpus. Surprisingly, only

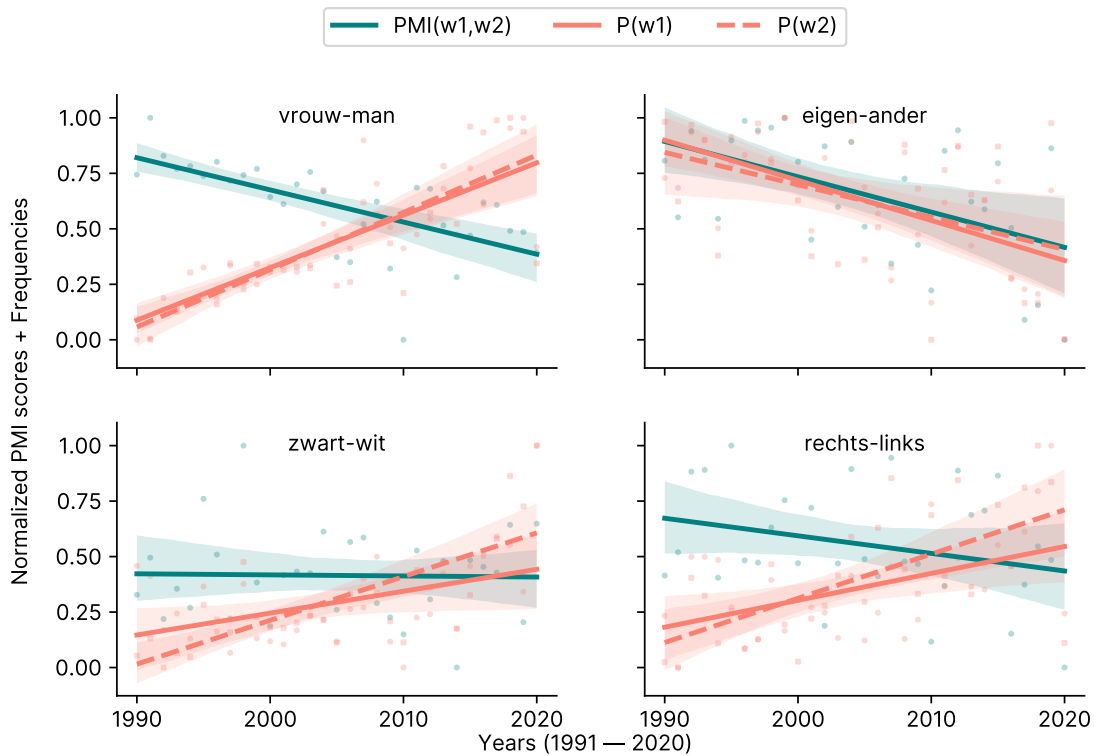


Figure 6: Adjusted PMI values and relative frequencies for four antonym pairs over time. For comparability, we normalized the PMI values and frequencies between 0 and 1.

603 of these pairs overlapped with the ones from our initial list, indicating that the majority of antonym pairs in the initial list were not present in the corpus. With the discovery of 127,691 new antonyms, the use of BERT models significantly expanded our analysis scope.

When we examine the antonyms with the highest probability (see Table 6 for the top 15), we observe that the model can identify a wide range of antonyms. Firstly, there are quite a few antonyms where one of the words begins with a negative prefix (in Dutch: *non-*, *on-*, *anti-*, *dis-*, *niet-*). 7,680 (6%) of the discovered pairs contain such a negative prefix. Incorporating all these words into a lexicon would be a labour-intensive task. However, a possible solution might involve combining lexicons with rule-based systems. Secondly, surprisingly, we also observe a lot of antonyms that refer to abstract ideological or artistic movements, such as *naturalistisch-surrealistisch* (naturalistic-surrealistic), *communisme-individualisme* (communism-individualism) and *anarchistisch-kapitalistisch* (anarchist-capitalist). Additionally, during our analysis, we found numerous antonyms opposing two countries, such as *duits-nederlands* (german-dutch), or *amerikaans-russisch* (american-russian). The utilization of BERT models thus results in a significant number of pairs that would be considered false negatives in a lexicon approach. These antonyms describe new forms of thinking or emerging societal developments, which is of particular importance in a journalism context.

At the same time, the use of a machine learning approach also produces a considerable number of false positives, which contaminate the analysis. Ideally, these false positives would be manually removed from the list, but at this scale, it would be quite time-consuming. The most evident form of false positives we found during our analysis are pairs where the words are not actually opposites but frequently occur near each other, such as *euro-procent* (euro-percentage), *ministerie-volksgesondheid* (ministry-public health), *eeneig-tweeling* (identical-twin). In other cases, the false positives included misclassified synonyms (*monetair-economisch*, monetary-economic) and, surprisingly, even different spellings of the same words (*amerikaans-amerikaanse*, american-american).

In the examples provided above, the classification of these pairs as antonyms is clearly incorrect. However, we also encountered numerous borderline cases that challenge the boundaries of how we define an antonym. For instance, is a geologist the opposite of a psychologist? Is a tomato antonymous to a bell pepper? Or a cyclist to a pedestrian? In all these cases, the answer depends on the context; on whether these words were used in an antonymous or complementary manner ('Cyclists are becoming a growing concern for pedestrians' vs. 'Drivers must be vigilant about pedestrians and cyclists.'). Although we utilized contextual language models, our setup did not fully incorporate the contextual aspect since we trained our models at the word level. To unlock the full potential of BERT models, a possible improvement to our setup would be to move beyond the word level and train models on a token sequence level, which would require a training dataset where words are tagged as antonyms in their context. This way, we can better capture the nuances and contextuality of antonymous relationships between words. However, curating such a dataset was beyond the scope of this paper.

5. Conclusion

Are binary constructions constructed or deconstructed in Dutch newspaper articles of the last thirty years? The question does not have a straightforward answer, as evidenced. It is challenging to determine when something can be considered an antonym pair, let alone a binary opposition. Nevertheless, our initial exploration yielded some noteworthy results. We have observed a modest decline in the use of antonym pairs in our corpus. Detected patterns in frequency changes could not only be linked to (trans)national events, but also to developments in the journalistic style, scope, and target groups of the newspaper *NRC*. Moreover, we have shown that using a BERT model in this task has led to promising results. Intriguing binary oppositions such as man-woman, black-white, and employer-employee emerged through the use of the trained classifier and our subsequent analysis. Utilizing an LLM has also confronted us with many new challenges. Pursuing and obtaining high-performance scores in training and finetuning large language models is no guarantee for success. Nevertheless, we are optimistic that further exploration of the potential of these models can lead to a more profound insight into the use of binary oppositions in Dutch newspapers and beyond.

References

- [1] K. Church, X. Cai, and Y. Bian. “Training on Lexical Resources”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 6290–6299.
- [2] K. Church and P. Hanks. “Word association norms, mutual information, and lexicography”. In: *Computational linguistics* 16.1 (1990), pp. 22–29.
- [3] J. Derrida. *Positions*. Chicago, Ill.: University of Chicago Press, 1981.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [5] S. Fallows. *A Complete Dictionary Of Synonyms And Antonyms*. 1898.
- [6] P. He, J. Gao, and W. Chen. “Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing”. In: *arXiv preprint arXiv:2111.09543* (2021).
- [7] M. Klages. *Literary Theory: A Guide for the Perplexed*. London: Bloomsbury Publishing Plc, 2007.
- [8] C. Lévi-Strauss. *The raw and the cooked*. New York: Harper and Row, 1969.
- [9] K. A. Nguyen, S. Schulte im Walde, and N. T. Vu. “Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 76–85. URL: <https://aclanthology.org/E17-1008>.
- [10] M. Roth and S. Schulte im Walde. “Combining Word Patterns and Discourse Markers for Paradigmatic Relation Classification”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 524–530. DOI: 10.3115/v1/P14-2086.
- [11] E. Santus, Q. Lu, A. Lenci, and C.-R. Huang. “Taking Antonymy Mask off in Vector Space”. In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. Phuket, Thailand: Department of Linguistics, Chulalongkorn University, 2014, pp. 135–144. URL: <https://aclanthology.org/Y14-1018>.
- [12] S. Scheible, S. Schulte im Walde, and S. Springorum. “Uncovering Distributional Differences between Synonyms and Antonyms in a Word Space Model”. In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, 2013, pp. 489–497. URL: <https://aclanthology.org/I13-1056>.
- [13] R. Schwartz, R. Reichart, and A. Rappoport. “Symmetric Pattern Based Word Embeddings for Improved Word Similarity Prediction”. In: *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China: Association for Computational Linguistics, 2015, pp. 258–267. DOI: 10.18653/v1/K15-1026.

- [14] Z. Xie and N. Zeng. “A Mixture-of-Experts Model for Antonym-Synonym Discrimination”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, 2021, pp. 558–564. DOI: 10.18653/v1/2021.acl-short.71.
- [15] W.-t. Yih, G. Zweig, and J. Platt. “Polarity Inducing Latent Semantic Analysis”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 1212–1222. URL: <https://aclanthology.org/D12-1111>.

A. Methods

We use a decay function of the form:

$$\text{Adjusted PMI}(x, y) = \text{PMI}(x, y) \cdot \exp(c \cdot \text{Sentence Length})$$

Where:

Adjusted $\text{PMI}(x, y)$ represents the adjusted PMI score for word pair x and y .

$\text{PMI}(x, y)$ is the standard PMI score for the word pair x and y .

c is the decay constant that controls the rate of decay.

Sentence Length refers to the length of the sentence in which the word pair is found.

The decay constant (c) is estimated through a curve-fitting process. We fit the decay function to the PMI scores and sentence lengths in the dataset using a nonlinear curve fitting. The estimated constant is determined based on this fitting process. This adjusted approach enhances the reliability of PMI-based analyses, offering a more consistent representation of word associations even as sentence lengths vary.

B. Antonym pairs and their translations

Table 3

Most frequent antonym pairs in the corpus.

original	translation	#	original	translation	#
vrouw - man	woman - man	22,182	vader - moeder	father - mother	5639
jaar - maand	year - month	14,174	ander - één	other - one	5581
groot - klein	large - small	14,073	jaar - kwartaal	year - quarter	5526
jaar - week	year - week	10,974	europes - amerikaans	european - american	5365
euro - procent	euro - percentage	10,752	werkgever - werknemer	employer - employee	5171
kind - ouder	child - parent	10,665	jaar - kort	year - short	4703
nieuw - oud	new - old	9526	amerikaans - nederland	american - dutch	4406
twee - één	two - one	8559	ander - oud	other - old	4222
eigen - ander	own - other	7822	goed - slecht	good - bad	4199
euro - dollar	euro - dollar	7695	duits - nederlands	german - dutch	4150
hoog - laag	high - low	7186	uur - week	hour - week	4060
keer - één	time - one	6751	internationaal - nederlands	international - dutch	3915
dag - week	day - week	6517	zoon - vader	son - father	3915
procent - totaal	percentage - total	6349	amerikaans - brits	american - british	3861
jong - oud	young - old	6226	politie - justitie	police - justice	3849
rechts - links	right - left	6101	ministerie - defensie	ministry - defense	3660
economisch - sociaal	economic - social	5904	minister - president	minister - president	3604
zwart - wit	black - white	5793	meisje - jongen	girl - boy	3596
vraag - antwoord	question - answer	5724	oud - kind	old - child	3591
europes - nederlands	european - dutch	5677	nederlands - buitenlands	dutch - foreign	3577

Table 4

Translations of most frequent antonym pairs in Figure 3.

original	translation
amerikaans - koreaans	american - korean
illegaal - legaal	illegal - legal
tegenstander - voorstander	opponent - supporter
zeker - onzeker	sure - unsure
amerikaans - italiaans	american - italian
amerikaans - japans	american - japanese
positief - negatief	positive - negative
meerderheid - minderheid	majority - minority
binnenlands - buitenlands	national - foreign
snel - langzaam	fast - slow
amerikaans - mexicaans	american - mexican
chemisch - biologisch	chemical - biological
integratie - immigratie	integration - immigration
conservatief - progressief	conservative - progressive
internationaal - regionaal	international - regional

Table 5

Translations of most decreasing and increasing antonym pairs in Figure 4 and Figure 5.

original	translation	original	translation
katholiek - protestants	catholic - protestant	jong - ziek	young - ill
belasting - premie	tax - premium	politie - burger	police - citizen
geestelijk - lichamelijk	mental - physical	uiteindelijk - aanvankelijk	ultimately - initially
leven - dood	life - death	school - ziekenhuis	school - hospital
algemeen - bijzonder	general - particular	los - vast	loose - fixed
economisch - sociaal	economic - social	ver - dicht	far - close
lang - kort	short - long	amerikaans - braziliaans	american - brazilian
televisie - radio	television - radio	muziek - beeldend	music - visual
twee - één	two - one	schrijver - schilder	writer - painter
zwart - blank	black - white	president - vicepresident	president - vice-president
vrouw - man	woman - man	ver - precies	far - precise
amerikaans - israëliësch	american - israelian	baby - ouder	baby - parent
snel - tempo	fast - pace	zwart - grijs	black - grey
amerikaans - japans	american - japanese	wedstrijd - rust	game - break
werkgever - werknemer	employer - employee	eigen - collectief	own - collective

Table 6

The top 15 antonyms with the highest probability.

original	translation	p_{ant}	cos sim	in train ds
globalisering - individualisering	globalization - individualization	0.99	0.70	×
geoloog - psycholoog	geologist - psychologist	0.99	0.49	×
helder - onhelder	clear - unclear	0.99	0.59	×
individualisering - mondialisering	individualization - mondialization	0.99	0.76	×
burgerschap - ministerschap	citizenship - ministry	0.99	0.45	×
degelijk - ongelijk	sound - uneven	0.99	0.47	×
excommunistische - socialistisch	excommunist - socialist	0.99	0.66	×
doordacht - ondoordacht	thoughtful - thoughtless	0.99	0.78	☒
democratisch - militaristisch	democratic - militaristic	0.99	0.55	×
onrechtvaardigheid - rechtvaardigheid	injustice - justice	0.99	0.85	☒
begrip - onbegrip	understanding - misunderstanding	0.99	0.67	×
juistheid - onjuistheid	correctness - incorrectness	0.99	0.89	×
biologisch - homeopathisch	biological - homeopathic	0.99	0.50	×
bioloog - criminoloog	biologist - criminologist	0.99	0.52	×
desintegratie - integratie	disintegration - integration	0.99	0.72	☒