Tracing Colonial Discourse in Dutch Historical Newspapers

Jiaqi Zhu[1], Teresa Paccosi[1], and Marieke van Erp[1]

[1] DHLab, Humanities Cluster, KNAW, Netherlands

Abstract

The expansion of colonialism often involves categorisations of colonised people based on their race, language, social status, and perceived level of "civilisation". While such categorisations may seem fixed, the language used to describe colonised populations is dynamic, shifting in meaning and connotation as colonial relationships and power structures change over time. We investigate diachronic semantic changes in Dutch colonial terminology through newspaper articles from 1860 to 1960 using word embeddings. This period witnessed major colonial expansion, the implementation of "ethical imperialism" (the Dutch "Ethical Policy" promoting education and welfare while maintaining colonial control), and eventual decolonisation. We combine computational semantic change detection with a more fine-grained analysis of the associated adjectives, demonstrating how the combination of distant and closer techniques can reveal patterns of linguistic transformation that reflect broader societal and political changes in Dutch colonial discourse.

Keywords: Diachronic Semantic Change, Word Embeddings, Colonial Categorisations, Digitised Newspapers, Computational History

Disclaimer: This paper contains derogatory sentences and words. They are provided solely as illustrations of the research results and do not reflect the opinions of the authors or their organisations. In-text examples of derogatory and potentially offensive language are presented in *italics*.

1 Introduction

Throughout history, colonisers frequently categorised colonised people based on their race, occupation, religion, class, and legal status. These categorisations helped maintain social hierarchies in colonial societies that benefited the colonisers or constituted the legitimising discourse of colonialism [14]. During Dutch colonial rule, some words that already existed in the Dutch lexicon shifted from their original meanings in referring to a specific group of people under the colonial kaleidoscope. For example, *inboorling*, which originally means "someone born in the land", was for a very long time a neutral word in the Dutch language. However, since the 19th century, it has been used to refer to Indigenous people from colonies, who were considered primitive and wild by colonisers [13]. Meanwhile, some terms from the colonised societies' language entered the Dutch lexicon as they were adopted by colonisers such as *koelie* (meaning "day worker", which is thought to be derived from the Hindi word *quli*) [4].

Newspapers serve as a valuable source for studying such linguistic changes because they mirror societal attitudes of their time [1], making them ideal for tracing how certain terminology or concepts evolved in public usage. The press may have documented official colonial policies, public debates about colonial affairs, and everyday references to colonial subjects, providing a comprehensive record of how language adapted to changing colonial realities. This paper contributes to understanding Dutch colonial discourse by computationally analysing the semantic evolution of colonial terminology in newspaper discourse between 1860 and 1960. Our approach provides empirical evidence for semantic change patterns in terms identified as problematic by contemporary cultural heritage institutions, offering historical context for current decolonisation efforts.

Jiaqi Zhu, Teresa Paccosi, and Marieke van Erp. "Tracing Colonial Discourse in Dutch Historical Newspapers." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1292–1303. https://doi.org/10.63744/SwkybkCCvmsj.

We use *Words Matter: an unfinished guide to word choices in the cultural sector* [10],[1] a lexicon developed in 2018 by the National Museum of World Cultures in the Netherlands to identify problematic terms in Dutch Cultural Heritage collections. We choose to analyse newspaper discourse as, unlike museum collections that reflect curatorial practices, newspapers captures how these terms were actively used and evolved in real-time public consciousness. By applying this lexicon to historical Dutch newspaper data, we can trace the semantic evolution of these terms and understand their changing social meanings over time.

The remainder of this paper is organised as follows. In Section 2, we discuss related work, followed by a description of our data in Section 3. In Section 4, we present the methodology. We present the results of our analyses and discussion in Section 5. Conclusions, limitations, and directions for future work are discussed in Section 6.

2 Related Work

Computational semantic change detection has predominantly relied on distributional semantics, particularly word embeddings, to trace how word meanings shift across historical periods. Several studies have demonstrated the effectiveness of these methods for examining politically contentious terminology [11; 12; 17; 18; 20]. Park and Cordell [11]'s analysis of "slave" and "servant" related terminology exemplifies how newspaper data can reveal euphemistic language shifts that reflect broader social attitudes, employing similar methods to our investigation of Dutch colonial terms. Soni, Klein, and Eisenstein [17]'s study of abolitionist networks demonstrates how newspaper discourse captures the evolution of activist terminology over time, while Pedrazzini and McGillivray [12]'s examination of mechanisation vocabulary in British newspapers and Tahmasebi [18]'s diachronic analysis of Swedish newspaper language illustrate the temporal analytical potential that we apply to Dutch colonial discourse. Wevers' study [20], which used Dutch historical newspapers to analyse language changes in gender bias, is the most directly related to our work. It uses a selection of the same corpus we used in this paper, demonstrating its suitability for semantic change detection.

Research specifically addressing Dutch colonial categorisations has predominantly taken qualitative approaches, examining questions similar to ours but through different methodological lenses. These studies have explored how mixed-race classifications such as "Indo-European" functioned in colonial power structures and contemporary identity negotiations [3], how the measurement and classification of racial "mixedness" reflected broader colonial administrative practices [5], and how ethnographic encounters shaped colonial categorisation systems [7]. [2] represents the closest precedent to our work, conducting computational word co-occurrence analysis of sensitive terms in Dutch newspapers, sharing both our methodological approach and dataset source, though this analysis focused on a more recent period (1950s-1990s) and employed different analytical techniques.

This paper addresses the methodological and temporal gap by applying computational semantic change detection methods to Dutch colonial terminology across the 1860-1960 period, combining the computational approaches demonstrated effective in international contexts with the Dutch newspaper corpus used by previous studies, while extending the temporal scope to cover the critical colonial expansion and decolonisation period. The methodology is mainly inspired by Menini et al.'s work [8], addressing the olfactory domain. This study investigates the way specific olfactory objects were perceived over time, analysing evaluative changes from a perceptual, cultural, and historical perspective.

3 Data

The historical Dutch newspapers data used in this study were collected from Delpher,[2] the Dutch National Library (KB) digitised newspaper and magazine portal. We selected newspaper articles spanning the period 1860-1960. Although Dutch colonial rule began long before this period and semantic changes might

[1] https://amsterdam.wereldmuseum.nl/en/about-wereldmuseum-amsterdam/research/words-matter-publication

[2] https://delpher.nl

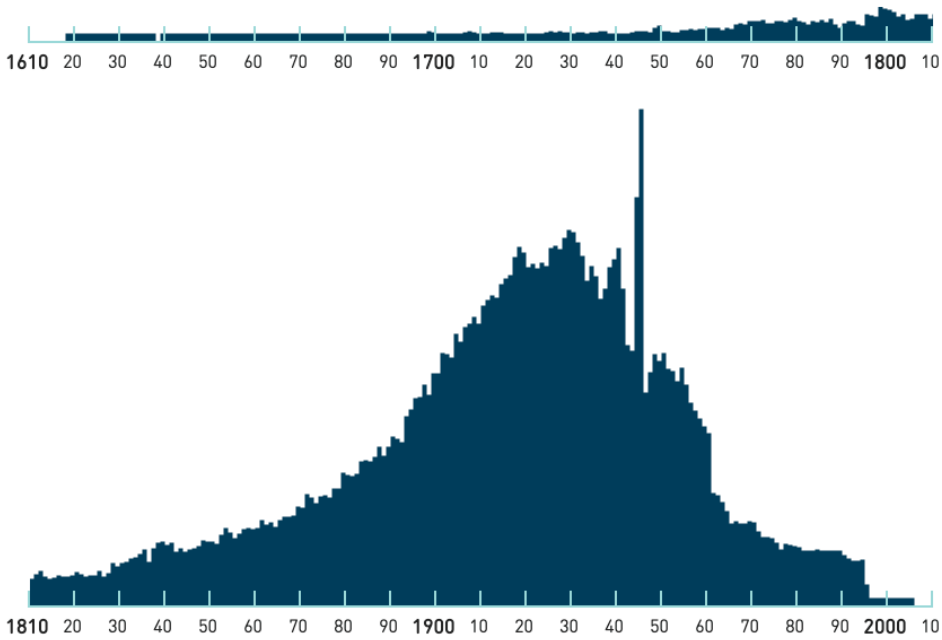

Figure 1: Number of digitised newspapers per year in the KB's *Delpher* collection (1618–1995). *Source: Koninklijke Bibliotheek (Delpher portal).* https://www.delpher.nl/nl/kranten

have occurred with earlier contact, our analysis starts from 1860 due to the availability and consistency of the digitised corpus. While the KB collection includes newspapers dating back to the seventeenth century, the number of digitised newspapers before the nineteenth century is extremely limited. As shown in Figure 1, the volume of digitised newspapers remains negligible until the mid-nineteenth century and only from 1860s onward the number of available newspapers become substantial and stable, exceeding 6,000 newspapers per year. Starting from 1860 thus ensures adequate data coverage and temporal consistency, providing a reliable basis for tracing linguistic and semantic patterns in newspapers.

We specifically target two major Dutch national newspapers, namely *Algemeen Handelsblad* and *De Telegraaf*, because of the contrast in their editorial positions during this period: *Algemeen Handelsblad* represents a more liberal perspective, *De Telegraaf* takes a more conservative to far-right direction, especially during World War II [6]. Furthermore, both newspapers maintained high circulation and similar distribution rates during the study period of 1860-1960, ensuring representative coverage of public discourse during this historical period. We only include articles while excluding other types of content, such as advertisements and announcements, to focus on editorial and news content rather than commercial material. See Figure 2 and Table 1 for statistics on the number of articles and tokens we collected.

We segment the collected dataset into three time periods: 1860-1899, 1900-1939, and 1940-1960. Due to the establishment of *De Telegraaf* in 1893, the temporal spans are not fully balanced across newspapers: *Algemeen Handelsblad* covers the full period (1860-1960), while *De Telegraaf* starting its coverage from 1893. The temporal imbalance between newspapers, with *De Telegraaf* missing the initial 33 years of the study period, may affect direct comparisons of early semantic patterns. However, we hypothesise that the number of tokens in this period is enough to draw some relevant conclusions from the data. We chose 1900 as a breaking point, as before this year, Dutch colonialism in the East Indies was fundamentally oriented toward commercial exploitation [15]. The implementation of the Ethical Policy in 1901 introduced new frameworks that marked a transition in official colonial ideology [19]. The Japanese occupation (1942-1945) and subsequent decolonisation struggles may have created contested semantic environments in which colonial terminology was challenged, redefined, or abandoned.

We applied text pre-processing procedures to remove noise and standardise the corpus, eliminating punctuation marks, symbols, and numerical characters, as well as filtering out words shorter than three characters to reduce the impact of abbreviations and typographical artefacts commonly found in histor-

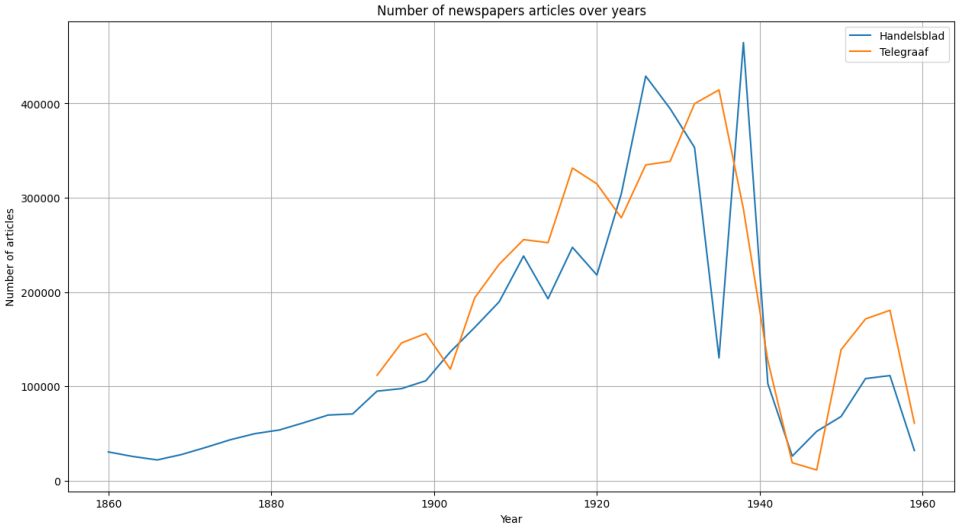

Figure 2: Number of articles over 1860-1960

Newspaper	1860–1889	1900–1939	1940–1960
De Telegraaf	54,908,714	362,705,143	73,323,824
Algemeen Handelsblad	215,533,509	510,985,442	108,149,025

Table 1: Newspaper tokens count per period

ical newspaper digitisation. We removed function words (articles, prepositions, conjunctions, auxiliary verbs) because they serve primarily grammatical rather than semantic functions, and their high frequency could mask the semantic relationships between content words that are central to our analysis of colonial terminology evolution.

4 Methodology

After the pre-processing step, we conducted two separate analyses, illustrated in Figure 3. We first use embeddings to examine the evolution of selected target words through changes in their neighbouring words (i.e., the words used in the most similar contexts). Secondly, we perform a connotative analysis of the adjectives most frequently used to describe the same words.

Our analysis focuses on terminology from the Words Matter lexicon, specifically terms that were used as race and social classifications. Table 2 presents our selection of Words Matter terminology, where word forms of each term are aggregated and collected under their stemmed forms (shown in bold italics). This aggregation approach accounts for morphological variations while maintaining semantic coherence. The frequency distribution of these selected terms across both newspapers and time periods is presented in Figure 4, which reveals considerable variation in usage patterns, with some terms appearing frequently throughout the corpus, while others show more sporadic occurrence.

4.1 Embeddings-based analysis

Following the methodology presented in [12], we trained three separate Word2Vec [9] models for each newspaper, one per period, testing different numbers of training epochs (3, 5, 10) and vector sizes (100, 300). To determine the optimal hyperparameters, we conducted a grid search evaluation comparing the quality of models trained with different parameter combinations. The best-performing configuration, which was selected for the final models, evaluated with a set of synonyms and spelling variations, used 3 training epochs and a vector size of 300. Since models for each temporal period are trained indepen-

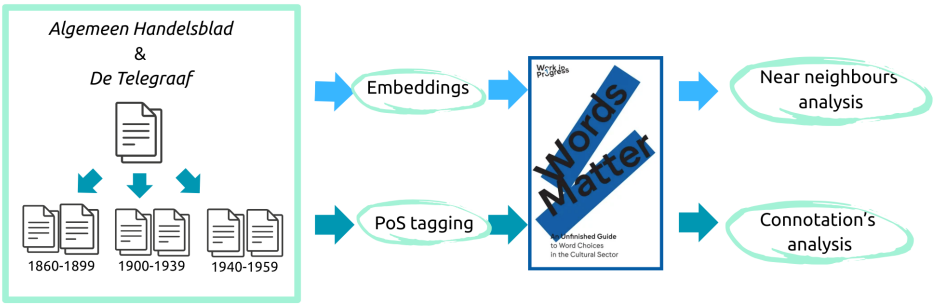

Figure 3: Analysis Workflow

dently, the resulting embeddings exist in different vector spaces and cannot be directly compared. We aligned the different semantic spaces using the Orthogonal Procrustes method [16]. After aligning all embedding spaces, we first measured semantic change by calculating the cosine similarity of the same word embedding across the different periods. Specifically, we set the cosine similarity in the first period to 1, using it as a reference point to observe how much the word representation changed in the subsequent periods. To assess whether the results obtained were not due to model artefacts, we conducted a control analysis using 10 neutral Dutch words, for which we did not expect significant changes. The average cosine similarity compared to the reference period for these words is 0.82 for *De Telegraaf* and 0.78 for *Algemeen Handelsblad*, while the average change in cosine similarity between period 1 (1900-1939) an period 2 (1940-1959) is 0.06 for *De Telegraaf* and 0.03 for *Algemeen Handelsblad*. These results show the stability of the meanings of these neutral words, and support the validity of the model. We then conducted a nearest neighbours analysis to further investigate this type of change.

We first analysed variations in **cosine similarity** of the selected keywords (see Table 2) across the chosen time periods, with the goal of identifying potential patterns of semantic change. Based on this analysis, we identified three main scenarios: *divergence*, *stability*, and *parallel* change across newspapers. Divergence occurs when the cosine similarity of a given keyword remains relatively stable in one newspaper but notably shifts in the other. We interpret this as an indication of diverging usage patterns, potentially influenced by the newspapers' different ideological orientations. In the case of stability, cosine similarity scores remain consistent in both newspapers, suggesting a possible, relatively stable semantic representation of the keyword over time. Parallel change refers to instances in which both newspapers exhibit similar degrees of semantic shift. This pattern may indicate the influence of broader historical phenomena on language use, independent of editorial perspective. To support our hypotheses, we also conducted a **nearest neighbours analysis** for the keywords in which we observed one of the semantic change scenarios. By comparing the nearest neighbours across time slices and newspapers, we sought to provide additional support to the patterns individuated with the cosine similarity analysis.

4.2 Connotation analysis

While the previous analyses provide an overview of which words have undergone relevant semantic change and the kinds of context in which they tend to appear, they do not offer information on their

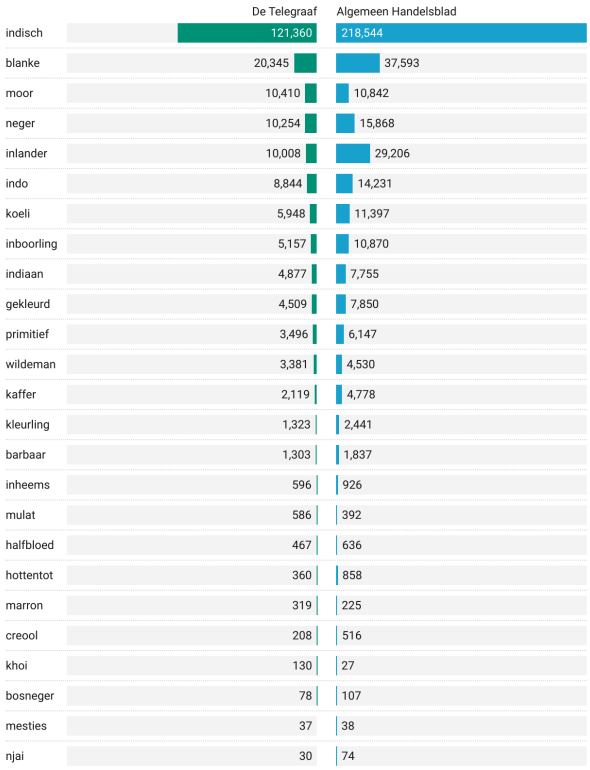

Figure 4: Term occurrences in *De Telegraaf* and *Algemeen Handelsblad* (sorted by frequency)

connotative dimension. To address this limitation, we conducted an analysis of the **connotative adjectives** (specification and evaluation) most frequently used to modify the identified keywords across the different time periods, aiming to explore potential connotative shifts of these concepts. Indeed, given that one of the main differences between the two newspapers lies in their ideological orientation, we expect connotation to play a key role in shaping the distinct uses of language. We therefore assigned part-of-speech (PoS) tags to both corpora, using SpaCy v3.7.[3] We isolated the selected keywords when used as nouns and extracted the adjectives occurring as their modifiers in each time period. We further identified connotative adjectives and analysed their frequency over time for each relevant keyword.

5 Results and Discussion

In this section, we present the results of our analyses on some selected colonial terms. We first present the result of the cosine similarity comparison in Subsection 5.1, quantifying the degree of semantic stability or change for each term across periods and newspapers, and providing a broader overview of which words underwent the most significant transformations. We then describe the nearest neighbours analysis (Subsection 5.2) to examine the specific semantic contexts and thematic associations surrounding these terms, addressing the observed changes from a qualitative perspective, to possibly support the results of the previous cosine analysis. Finally, we illustrate the connotation analysis of these terms in Subsection 5.3, exploring how ideological differences between newspapers may have shaped the affective and judgmental aspects of colonial discourse. Together, these three analytical approaches provide both quantitative measures of semantic change and qualitative insights into the mechanisms driving linguistic transformation in colonial contexts. From the set of *Words Matter* terminology in Table 2, we selected some that underwent one of the selected changes, for which we conduct a more fine-grained analysis, combining the insights coming from the three different analyses.

[3] `https://spacy.io/models/nl`

Category	Terminology (translation)
Race	***blanke***(n) (eng., "white person"); ***bosneger***(s) (eng., "bush negro"); ***creool***, creolen (eng., "creole"); ***gekleurd***(en) (eng., "colored"); ***halfbloed***(en) (eng., "half-blood"); ***Hottentot***(ten) (eng., "Khoikhoi people"); ***inboorling***(en) (eng., "primitive native"); ***indisch***(e) (this word doesn't have a strict translation in English); ***indo***('s) (eng., "Indo-European"); ***indiaan***, indianen (eng., "Indian"); ***inheems***(en) (eng., "indigenous"); ***inlander***(s) (eng., "native"); ***kaffer***(s) (eng., "black African"); ***Khoi*** (eng., "Khoisan people"); ***kleurling***(en) (eng., "colored"); ***moor***, moren (eng., "Muslim person of Arab or Amazigh descent"); ***marron***(s) (eng., "maroon"); ***mesties*** (eng., "person of mixed-race background"); ***mulat***(ten) (eng., "mulatto"); ***neger***(s, in, innen) (eng., "negro (m/f)"); ***njai*** (eng., "Indonesian mistress to coloniser"); ***primitief***, primitieven (eng., "primitive"); ***wildeman***(nen) (eng., "uncivilized man").
Social	***barbaar***, barbaren (eng., "barbarian"); ***koeli***(es) (eng., "contract worker").

Table 2: Selection of *Words Matter* terminology by category (Race and Social). Word forms of each term are aggregated, and each aggregation is collected under its stemmed form (in bold italics).

Word	*De Telegraaf*		*Algemeen Handelsblad*	
	1900–1939	**1940–1959**	**1900–1939**	**1940–1959**
koeli	0.480	0.481	0.514	0.260
mestjes	0.704	0.585	0.487	0.640
moor	0.258	0.238	0.283	0.244
barbaar	0.545	0.585	0.556	0.581
neger	0.428	0.421	0.445	0.454
primitief	0.549	0.373	0.479	0.370
marron	0.309	0.509	0.439	0.665

Table 3: Cosine similarity of target words across different periods, with 1860–1899 as the reference period. In this reference period, the cosine similarity of each word is set to 1.

5.1 Cosine Similarity

As introduced in Subsection 4.1, the cosine similarity analysis reveals three distinct patterns of semantic change across the selected colonial terminology (see Table 3). In the case of *divergence*, we identified ***koeli*** and ***mesties*** as exemplar cases where similarity remains stable in one newspaper while showing changes in another, with *koeli* maintaining consistent scores across periods in *De Telegraaf* but dropping substantially in the third period in *Algemeen Handelsblad*, while *mesties* shows an inverse pattern with declining scores in *De Telegraaf* but growth in *Algemeen Handelsblad*'s final period. Such divergences suggest different editorial approaches or audience orientations toward colonial terminology. In contrast, ***neger***, ***barbaar***, and ***moor*** maintain relatively stable cosine similarity across both newspapers and all three periods, appearing to retain consistent semantic positioning despite the changing historical context and suggesting rather fixed usage patterns even as surrounding discourse evolved (*stability*). Finally, ***primitief*** and ***marron*** demonstrate parallel changes occurring in both newspapers, indicating broader shifts in semantic meaning that transcend individual newspaper contexts and likely reflect society-wide transformations in conceptual understanding rather than newspaper-specific editorial decisions (*parallel*).

5.2 Nearest Neighbours Analysis

In this and the following subsection, we focus on four words (*koeli*, *moor*, *neger*, *primitief*) that appear with sufficient frequency across all periods and newspapers to enable reliable analysis, see Figure 4. *Bar-*

Target Word	1860–1899	1900–1939	1940–1959
koeli	koelie, drager, chinese, indische, arbeid	djohan, oemar, ebak, daiar, huisbediende	kenyase, gratiebesluit, vergaderverbod, iokja, krijgsscholen
moor	arabier, mohammedaan, moren, noord-afrikaan, turk	aooll, omke, latne, rfnt, axy	denoue, othelio, hondenhuizen, lerscl, stnrimans
neger	inboorling, kaffer, hottentot, mulat, slavernij	kaffer, limbus, polkan, caid, ameer	kazemiera, moorc, koesewitski, strasberg, alcc
primitief	stam, inheems, natuurvolk, woestijnvolk, barbaars	primitieve, ingenieus, onpraktisch, ouderwets, rustiek	animaal, onsamenhangend, systeemloos, laboreerde, onvertroebeld

Table 4: Nearest neighbours in *De Telegraaf*. Each word is normalised, considering possible spelling variations to focus on semantic relationships rather than historical spelling conventions.

Target Word	1860–1899	1900–1939	1940–1959
koeli	bubber, otting, javanen, werkovereenkomst, immigranten	contractarbeiders, poenale, chinees, ordonnantie, mandoer	geisa, desaman, walglijk, kameeldrijvers, vteen
moor	borah, laerte, rodolpho, germont, cassio	ananias, andreus, saathoff, falconer, fricker	sohroder, chrlstlaan, tiaan, pieok, asselyn
neger	kleurling, kabyl, kokkin, mestietzen, negermeisje	kaffer, roodhuid, elroy, alarik, uncle	kleurling, alabama, negerjongen, chuster, folsom
primitief	primitieve, heffingspercentage, kohier, degressief, intomen	primitieve, suppletoir, kohier, straatgeld, ingericht	armetierig, primitieve, sensibel, seerend, aartsvaderlijk

Table 5: Nearest neighbours in *Algemeen Handelsblad* (spelling variations normalised).

baar, *marron*, and *mesties* were excluded from these detailed analyses as they lack adequate contextual data in the present corpus for meaningful neighbour extraction and adjective co-occurrence patterns.

The nearest neighbour analysis reveals that cosine similarity patterns alone provide insufficient evidence for understanding semantic change, as identical similarity scores can mask fundamentally different semantic trajectories between newspapers. ***Koeli*** shows contrasting patterns, with neighbours in *De Telegraaf* evolving from labour-focused terms to institutional contexts, suggesting gradual semantic broadening that supports stable similarity scores, while *Algemeen Handelsblad* demonstrates a dramatic change from labour contracts to administrative terms and finally discriminatory contexts such as *walglijk* (“disgusting”). This divergence is particularly noteworthy as the more liberal newspaper shifts toward derogatory content, possibly reflecting increasing social tensions around labour migration. Meanwhile, ***neger*** and ***moor*** exhibit stable similarity scores, hiding different underlying changes, with ***neger*** shifting from Dutch colonial racial discourse toward American racial contexts in both newspapers, while ***moor*** shows identical similarity scores despite completely different semantic contexts—ethnic-geographic terms in *De Telegraaf* versus literary references in *Algemeen Handelsblad*—revealing how quantitative stability can conceal qualitative divergence. In contrast, ***primitief*** demonstrates convergent

semantic evolution across both newspapers, transitioning from racialised colonial descriptors to general temporal adjectives, reflecting broader societal shifts away from explicit racial categorisation. These findings challenge assumptions about newspaper ideology, with the more liberal *Algemeen Handelsblad* sometimes showing shifts toward more derogatory content than the conservative *De Telegraaf*. The list of neighbours for each keyword across newspapers and periods is provided in Table 4 and Table 5.

5.3 Connotation Analysis

The connotation analysis validates and enriches the patterns identified through cosine similarity and nearest neighbour analysis. ***Koeli*** demonstrates subtle connotational shifts that support the nearest neighbour analysis findings, with adjectives in *De Telegraaf* remaining relatively stable across periods, dominated by regional descriptors like *chineesche* (Chinese) and *javaansche* (Javanese), though some negative undertones emerge in later periods, such as *flauwe* (weak/poor quality) and *matige* (moderate). While regional descriptors also dominate in *Algemeen Handelsblad*, the much higher frequency of *vrije* (free) compared to *De Telegraaf* aligns with the liberal newspaper's ideological orientation. ***Neger*** demonstrates consistent connotative stability across both newspapers, with persistent references to geographical origins (*amerikaansche/amerikaanse* [American], *afrikaansche/afrikaanse* [African]), demographic descriptors (*jonge* [young], *oude* [old]), and racial contrasts (*blanken* [whites], *zwart* [black]), though *Algemeen Handelsblad* shows slightly more positive connotations in the final period with terms like *vrije* (free) and *gelijkstelling* (equality), suggesting evolving social attitudes without fundamental semantic change. ***Moor*** exhibits similar connotative stability, with the consistent appearance of *Othello* across periods in both newspapers confirming the literary contextualization revealed in the neighbour analysis. Finally, ***primitief*** shows clear connotative evolution, transforming from racially-charged colonial descriptors to neutral temporal and aesthetic categories, with early periods showing minimal evaluative content but later periods shifting toward cultural-geographical references (*vlaamsche* [Flemish], *italiaanse* [Italian]) and aesthetic judgments (*moderne* [modern], *mooie* [beautiful]), supporting the parallel deracialisation identified in Subsection 5.2.

5.4 Discussion

The nearest neighbour analysis across all four words reveals that cosine similarity patterns alone provide insufficient evidence for understanding semantic change trajectories. ***Koeli*** exemplifies how identical similarity measures can conceal opposing semantic developments, with stable scores in *De Telegraaf* masking gradual institutional broadening while dramatic drops in *Algemeen Handelsblad* reflect deteriorating discourse toward discriminatory contexts. Words exhibiting apparent stability (***neger, moor***) often mask fundamental shifts in semantic positioning, while words showing dramatic similarity changes can reflect fragmentation (***primitief***) in their usages. Beyond these within-newspaper complexities, identical similarity patterns between newspapers can conceal entirely different semantic contexts, as demonstrated by ***moor***'s ethnic-geographic associations in one newspaper versus literary references in another. These findings underscore that nearest neighbour analysis is essential for distinguishing between generic statistical patterns and semantic transformations, revealing whether apparent stability represents true consistency or underlying repositioning within discourse. The connotation analysis further demonstrates that associated adjective shifts often occur more gradually than nearest neighbour changes might suggest, with newspapers maintaining relatively neutral descriptive language even as semantic contexts evolve. This reveals that ideological differences between newspapers may manifest more subtly in their linguistic choices than initially hypothesised, requiring multi-layered analysis to capture the full spectrum of semantic change.

6 Conclusions, Limitations, and Future Directions

This study has demonstrated how computational semantic change detection can reveal patterns of linguistic transformation in Dutch colonial terminology that reflect broader societal and political changes

between 1860 and 1960. By applying word embedding techniques to digitised newspapers from Delpher, we provided empirical evidence for how colonial categorisations evolved in public discourse during a critical period of colonial expansion, ethical imperialism, and eventual decolonisation. Our methodology combined distant reading approaches through cosine similarity analysis with closer examination of semantic neighbourhoods, revealing distinct patterns of semantic change across different types of colonial terminology. Across all examined words, we find that cosine similarity patterns alone provide insufficient evidence for understanding semantic change trajectories, as words exhibiting stability often mask shifts in semantic positioning, while words showing big similarity changes can reflect fragmentation or specialisation in their usages. Most importantly, identical similarity patterns between newspapers can conceal entirely different semantic contexts, demonstrating that nearest neighbour analysis is essential for distinguishing between generic statistical patterns and semantic transformations. These findings contribute to computational approaches for studying historical language change while offering quantitative insights that complement existing qualitative scholarship on Dutch colonial categorisations.

Several limitations should be acknowledged. First of all, our corpus may not represent a complete collection of articles from *De Telegraaf* and *Algemeen Handelsblad* during 1860-1960. Several factors can contribute to potential data gaps: not all historical newspapers have been digitised by the Koninklijke Bibliotheek (KB), some are only digitised for specific periods, and API access limitations may have affected data retrieval. Additionally, the temporal imbalance between newspapers—with *De Telegraaf* established only in 1893—creates unequal coverage. However, as shown in Table 1, the substantial token counts provide enough data to support the validity of our semantic change analysis despite these limitations. On average, *De Telegraaf* contains 163.6 million tokens per period, and the *Algemeen Handelsblad* 278.2 million. It should be noted that the middle period (1900–1939) spans 40 years compared to 30 years for the first period and 21 years for the last period, which accounts for the higher token counts in this period. Additionally, although the analysis focused on two major newspapers that have high circulation compared to others, the corpus selection might imply that semantic variations that occurred in other newspapers could have been missed. Another limitation is represented by the relatively small sample of terms that we selected from the *Words Matter* collection, which is enough as a proof of concept, but requires further analyses to assess the degree of generalisability of our findings. Finally, the digitisation of the Dutch historical newspaper corpus used in this study relies on Optical Character Recognition (OCR) technology, which may introduce noise that affects embedding quality and potentially impacts semantic change detection. While we applied text pre-processing procedures to remove obvious noise, systematic assessment of OCR quality and its impact on diachronic semantic change analysis remains a limitation of this study. We used 'nl_core_news_sm' from spaCy for POS tagging, in our pre-processing of data, which reports accuracy of 0.96 on modern Dutch text. However, its performance on historical Dutch may be lower, which potentially introduces systematic errors that could affect semantic analysis quality.

Future research could address these limitations through several methodological improvements. Incorporating newspapers from different regions, political orientations, and target audiences would enhance the representativeness of our findings. Expanding the lexical sample to include a broader range of colonial terms might increase generalisability and enable more comprehensive claims about colonial discourse evolution. The temporal scope could be extended beyond our current span to investigate how colonial terminology continued evolving in pre- and post-colonial contexts, contributing to contemporary discussions about decolonisation in cultural heritage institutions. Systematic OCR quality assessment and correction strategies should be implemented to reduce the potential impacts of digitisation artefacts. Further analyses could include linguistic investigations into the influence of evaluative content on semantic change, for instance, examining whether pejorative colonial terms shift toward neutral/positive meanings following the same patterns as neutral terms acquiring negative connotations. Given recent advances in language representation, we also plan to investigate the capabilities of Large Language Models (LLMs) in capturing the semantic changes of colonial terms, exploiting token-based embeddings and generative approaches.

Data Release

The code to train the embeddings, and perform the cosine similarity and connotation analyses is available at https://github.com/trifecta-project/Dutch_Colonial_Terms_in_Newspapers, together with the embedding models we trained.

Acknowledgments

Funded by the European Union under grant agreement 101088548 - TRIFECTA. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We thank our HuC colleagues Manjusha Kuruppath and Jelle van Lottum for their discussions and suggestions. We also thank Mirjam Cuper from the KB Data Services for her help with the API.

Author Contributions

Author contributions (by author initials) are listed according to the Contributor Roles Taxonomy (CRediT). Conceptualization: JZ; Data Curation: JZ, TP; Funding Acquisition: MvE; Investigation: JZ, TP; Methodology: JZ, TP, MvE; Project administration: MvE; Software: JZ, TP; Supervision: MvE, TP; Visualization: JZ, TP; Writing (original draft): JZ, TP; Writing (review and editing): JZ, TP, MvE

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Claude for formatting assistance, grammar, and spelling checks, as well as rephrasing sentences.

References

[1] Bell, Allan. *The Language of News Media*. Oxford, UK: Blackwell, 1991.

[2] Brate, Ryan, van Erp, Marieke, and Van den Bosch, Antal. "Contextual Profiling of Charged Terms in Historical Newspapers". In: *Proceedings of the 4th Conference on Language, Data and Knowledge*, ed. by Sara Carvalho, Anas Fahad Khan, Ana Ostroški Anić, Blerina Spahiu, Jorge Gracia, John P. McCrae, Dagmar Gromann, Barbara Heinisch, and Ana Salgado. Vienna, Austria: NOVA CLUNL, Portugal, Sept. 2023, pp. 97–108. URL: https://aclanthology.org/2023.ldk-1.9/.

[3] Doornbos, Julia Rosa, van Hoven, Bettina, and Groote, Peter D. "Negotiating claims of 'whiteness': Indo-European everyday experiences and 'mixedrace' identities in the Netherlands". In: *Social Identities 28*, no. 3 (2022), pp. 383–399. DOI: 10.1080/13504630.2022.2029739.

[4] EtymologieBank. "Koolie – Etymologiebank.nl". https://www.etymologiebank.nl/trefwoord/koelie. Accessed: 2025-07-11.

[5] Jones, Guno and Hart, Betty de. "(Not) Measuring Mixedness in the Netherlands". In: *The Palgrave International Handbook of Mixed Racial and Ethnic Classification*, ed. by Zarine L. Rocha and Peter J. Aspinall. Cham: Palgrave Macmillan, 2020, pp. 367–387. DOI: 10.1007/978-3-030-22874-3_20.

[6] Kuitenbrouwer, Vincent. "Propaganda that Dare not Speak its Name: International information services about the Dutch East Indies, 1919–1934". In: *Media History 20*, no. 3 (2014), pp. 239–253. DOI: 10.1080/13688804.2014.920204. URL: https://doi.org/10.1080/13688804.2014.920204.

[7] Meersbergen, G. van. *Ethnography and Encounter*. Leiden, The Netherlands: Brill, 2021. DOI: https://doi.org/10.1163/9789004471825.

[8] Menini, Stefano, Paccosi, Teresa, Tekiroğlu, Serra Sinem, and Tonelli, Sara. “Scent mining: Extracting olfactory events, smell sources and qualities”. In: *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 2023, pp. 135–140.

[9] Mikolov, Tomás, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. 2013.

[10] National Museum of World Cultures. “Words Matter: Publication on sensitive language in the museum sector”. Online publication. https://amsterdam.wereldmuseum.nl/en/about-wereldmuseum-amsterdam/research/words-matter-publication. 2025.

[11] Park, Jaihyun and Cordell, Ryan. “A Data-driven Investigation of Euphemistic Language: Comparing the usage of “slave” and “servant” in 19th century US newspapers”. In: *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, ed. by Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar. Albuquerque, USA: Association for Computational Linguistics, May 2025, pp. 350–364. ISBN: 979-8-89176-234-3. DOI: 10.18653/v1/2025.nlp4dh-1.31. URL: https://aclanthology.org/2025.nlp4dh-1.31/.

[12] Pedrazzini, Nilo and McGillivray, Barbara. “Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers”. In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, ed. by Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. Taipei, Taiwan: Association for Computational Linguistics, Nov. 2022, pp. 85–95. DOI: 10.18653/v1/2022.nlp4dh-1.12. URL: https://aclanthology.org/2022.nlp4dh-1.12/.

[13] Marlies Philippa, Frans Debrabandere, Arend Quak, Tanneke Schoonheim, and Nicoline van der Sijs, edited by. *Etymologisch Woordenboek van het Nederlands*. 4 vols. 4 volumes, 2003–2009. Amsterdam: Amsterdam University Press, 2003.

[14] Raben, Remco. “Colonial shorthand and historical knowledge: Segregation and localisation in a Dutch colonial society”. In: *Journal of Modern European History 18*, no. 2 (2020), pp. 177–193. DOI: 10.1177/1611894420910903.

[15] Willem van Schendel, edited by. *Embedding Agricultural Commodities: Using Historical Evidence, 1840s–1940s*. London: Routledge, 2017. ISBN: 9780815366843.

[16] Schönemann, Peter H. “A generalized solution of the orthogonal Procrustes problem”. In: *Psychometrika 31*, no. 1 (1966), pp. 1–10. DOI: 10.1007/BF02289451.

[17] Soni, Sandeep, Klein, Lauren F, and Eisenstein, Jacob. “Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers”. In: *Journal of Cultural Analytics 6*, no. 1 (2021).

[18] Tahmasebi, Nina. “A Study on Word2Vec on a Historical Swedish Newspaper Corpus”. In: *Proceedings of the 3rd Conference on Digital Humanities in the Nordic Countries*, ed. by Mila Oiva, Laura Saarenmaa, and Asko Nivala. Vol. 2084. Helsinki, Finland: CEUR Workshop Proceedings, 2018, pp. 387–397.

[19] van der Jagt, Johan Willem. *Imperialism and Morality: Indonesia, Suriname, the Caribbean, the Netherlands and the Colonial Administration of A.W.F. Idenburg, 1901-1935*. English. PhD-Thesis - Research and graduation internal. Vrije Universiteit Amsterdam, June 2021.

[20] Wevers, Melvin. “Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990”. In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, ed. by Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 92–97. DOI: 10.18653/v1/W19-4712. URL: https://aclanthology.org/W19-4712/.