# Contextual Profiling of Charged Terms in Historical Newspapers

**Ryan Brate** and **Marieke van Erp**
KNAW Humanities Cluster, DHLab
Oudezijds Achterburgwal 185
1012 DK Amsterdam, Netherlands
{ryan.brate,marieke.van.erp}
@dh.huc.knaw.nl

**Antal van den Bosch**
Utrecht University
Institute for Language Sciences Utrecht,
the Netherlands
a.p.j.vandenbosch@uu.nl

## Abstract

We extract nouns and corresponding co-occurrent targeted context features from a large corpus of Dutch language newspaper articles, from 1950s through the 1990s. Applying a well-established approach for scoring context feature and centre word associativity, we explore using the scores in the task of identifying key characteristics of known–charged terminology. Then use these features to draw parallels between known–charged and other terms. In the context of the very current decolonisation efforts amongst museum institutions, such approaches offer an opportunity to condense large quantities of data into the most-significant, salient information for digestion by heritage professionals. The methods were found to indeed yield insights into known and candidate charged terms.

**Disclaimer:** This paper contains derogatory words and phrases. 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 are presented in *"quotes, boldfaced and italicised"*.

## 1 Introduction

Museums of the World,[1] a database of cultural heritage institutions, records approximately 55,000 museums spread over 202 countries. The largest such collection, The Smithsonian Institution[2] alone holds in excess of 155M. Such collections enhance our collective understanding of our shared past, but in doing so, they give cultural heritage institutions powerful voices in the shaping of historical narratives in the public consciousness.

Many museum collections originate from the colonial period, with metadata and object portrayals stemming from the particular world of the time. There is now a growing movement of *decolonisation* in western museums aimed at the acknowledgement and accommodation of previously marginalised voices to combat biases propagated by the advancement of narrow viewpoints (Odumosu, 2020). Part of the decolonisation effort centres around greater sensitivity and reconsideration of the terminology and language used in item metadata. This is more complicated than wholesale removal of terminology from metadata and items from collections, even if such problematic terms are known. To handle the complexities properly, there needs to be greater contextual understanding of a term's implied characterisation in context. For instance, many terms nowadays considered problematic are ambiguous, also in their contentiousness: calling a plant *exotic* is different from calling a person the same. When and why terms are deemed problematic is complex, but the recognition of the social-cultural (contextual) aspects of terms provides a mechanism for some degree of understanding and comparison.

In this paper, we aim to explore the contextual profiles of a reference set of known charged collective nouns, reflective of some people group and identify the contextual features that distinguish them. Specifically, we consider four complementary context feature types: verbs for which the noun is the agent, verbs for which the noun is the patient, adjectives, and compound word modifiers as applied to the nouns. I.e., we are trying to capture the things done to them, the things they do and the attributes ascribed to them. In order to do so, we leverage the extensive digitised (and OCR'd) newspaper collection of the National Library of the Netherlands (KB), between the 1950s and 1990s, thereby capturing the period of European decolonisation to more recent post-colonial times. Such a

---

[1] https://www.degruyter.com/database/MOW/html
[2] https://www.si.edu/newsdesk/factsheets/smithsonian-collections

collection represents a valuable resource reflective of public discourse, attitudes and societal norms of the times.

In exploring context and its relevance to charged nouns, we make use of noun–context associativity measures. Specifically, we ask for each noun of a set of known charged nouns, *do contextual features exist, which for some noun–context feature associativity score threshold, are highly predictive of the noun?* Secondly, we seek to examine the parallels that can be drawn between known charged nouns: i.e., *are there context features which for some noun–context feature associativity threshold, recall multiple known charged nouns with a reasonable degree of precision with respect to our known charged noun set?* Finally, we examine those nouns, not part of our known charged noun set, which share similar context feature associations: asking, *can the context features of known charged nouns help identify other charged instances?*

## 2 Related Work

Our work is situated on the intersection of detecting and modeling bias and harmful language. Bias in large datasets and its effects on models learned on those datasets has gained more attention in recent years (cf. (Sap et al., 2020; Bender et al., 2021; Schick et al., 2021; Birhane et al., 2022)). Work done on the same corpus as ours is (Wevers, 2019), who aims to detect gender bias in Dutch newspapers. We focus on broader biases and harmful language, mostly coming from a colonial perspective. The GLAM community is very well aware of problematic artefacts of colonial history in datasets (cf. (Mohamed et al., 2020; Barabucci et al., 2020; Luthra et al., 2023)) but there has been less attention for this in the NLP community. In our prior work, we have started to investigate how certain terms are viewed by the general public via a crowdsourcing experiment (Brate et al., 2021). We found that context plays an important role in whether certain terms are deemed charged or not. In this paper, we extend this work by modelling contextual features of charged terms.

The detection of hate speech gained traction with the growing popularity of social media data and includes cyberbullying, insults, vulgar content and racist language (Schmidt and Wiegand, 2017). While the charged terminology we are investigating has overlaps with the dimension investigated in hate speech, colonially biased language tends to be somewhat more subtle than overt insults, although these do occur. It should also be noted that researching harmful stereotypes requires a balanced approach to not inadvertently incur more harm (Kirk et al., 2022).

Our approach to use adjectives and verbs directly associated with entities, as contextual features for distinguishing entities is inspired by (Bamman et al., 2013). They used a hierarchical Bayesian approach to group film-character types across film and film tropes, using the characterisation of characters in terms of *the things they do*, *the things done to them*, and *the way they are described* as features. However, whereas the soft-clustering iterative approach used by Bamman is based on broad feature commonality, and favours data-rich cluster types; we expect charged terms to yield often highly unique associations, not necessary given to easy feature clustering. Consequently, whilst inspired by this approach, we consider feature comparison based on a metric of noun-feature *keyness*, i.e., associativity score, based on the work of (Dunning, 1993).

## 3 Methodology

We use the raw data of the National Library of the Netherlands OCR'd newspaper dataset.[3] We split the data into discrete years to be analysed independently, as usages and characterisations of known-charged terms are subject to variation over time. We take sample years per decade, to be considered separately. The expectation is that one-year periods are too short to be regularly affected by confusing shifts in usage. We use the sampled data to create tables of associativity, or *keyness*, scores by collective noun and context features to answer our research questions.

The adjective–noun and verb–noun pairs are extracted by pattern matching against part of speech (POS) tagged dependency trees of the newspaper dataset. In the case of modifier–noun pairs, a corpus of modifiers and corresponding heads is bootstrapped from our set of known-charged words. Subsequently, the coincident collective noun–context feature pairs are assembled into separate frequency tables according to the context feature type (e.g., adjective) for each sample year. For the known-charged nouns, the frequencies for all plural forms of the noun are aggregated. The raw collective noun–context feature co-ocurrence fre-

---

[3] https://delpher.nl

quencies are then converted to some metric of *key-ness*, which is used as the basis for exploring the key features by collective noun, and for exploring the parallels between collective nouns.

### 3.1 Charged nouns

The terms in Table 1 are used as our reference set of known charged collective nouns. The basis of this list is the aforementioned Words Matter document (Modest and Lelijveld, 2018). We consider singular and plural forms.

### 3.2 Dataset

All available, publicly accessible OCR'd articles of the National Library of the Netherlands (KB) newspaper set, in each of the years as listed in Table 2, were taken in their entirety. The table also lists the number of approximate resulting extracted articles.

A dependency-parsed, POS-tagged version of this dataset was created via spaCy (Honnibal and Montani, 2017), with an intermediate step of rule-based tokenisation and sentence segmentation via regular expressions. To reduce the sentence complexity passed to spaCy, segmentation is additional performed on conjunctions, ":", "," and ";".

### 3.3 Building a corpus of the modifier–head components of compound nouns

As described in section 3.6, the keyness metric adopted in determining how key a some *particular context feature* is to some *particular noun* in question, is a function of the corpus-wide noun–context feature co-occurrence frequencies. Hence, a corpus of modifier–head instances is needed which consists of all modifiers coincident with known charged nouns, and all of the corresponding heads coincident with these modifiers. The spaCy dependency parse of the KB newspaper corpus provides a list of tagged instances of nouns. Using this list of nouns together with the charged noun set, we bootstrapped a corpus of modifier–head compound words.

Separately, for each of the years in Table 2, a corpus of modifier–head components of compound nouns was assembled. The result considers modifiers with or without terminating hyphens as being the same instance. The following approach was adopted to bootstrap the corpus from the known-charged nouns:

**Parameters**

| Category | Charged Nouns (*translation*) |
|---|---|
| race | **aboriginal**(s) (*aboriginal(s)*); **afkomst**(en) (*descent(s)*); **allochtoon**, allochtonen (*migrant(s)*); **Berber**(s) (*Berber(s)*); **blanke**(n) (*white person(s)*); **bosneger**(s) (*bush negro*); **creool**, creolen (*creole(s)*); **eskimo**('s) (*eskimo(s)*); **etniciteit**(en) (*ethnicity(-ies)*); **gekleurd**(en) (*colored(s)*); **halfbloed**(en) (*half-blood(s)*); **Hottentot**(**ten**) (*Khoikhoi people*); **immigrant**(en) (*immigrant(s)*); **inboorling**(en) (*primitive native(s)*); **indo**('s) (*Indo-European(s)*); **indiaan**, indianen (*Indian(s)*); **inheems**(en) (*indigenous*); **inlander**(s) (*native(s)*); **kaffer**(s) (*black African*); **Khoi** (*Khoisan people*); **kleurling**(en) (*colored(s)*); **koppensneller**(s) (*headhunter(s)*); **moor**, moren (*Muslim people of Arab and Amazingh descent*); **marron**(s) (*maroon*); **medicijnman**(nen) (*medicine man(men)*); **mesties** (*person of mixed-race background*; **migrant**(en) (*migrant(s)*); **mulat**(ten) (*mulatto(s)*); **neger**(s, in, innen) (*negro(s) (m/f)*); **njai** (*Indonesian mistress to coloniser*); **oorsprong**(en) (*descent(s)*; **primitief**, primitieven (*primitive(s)*); **Pygmee**(ën) (*Pygmy(Pygmees)*); **ras**(sen) (*race(s)*); **roots** (*roots*); **scalp**(en) (*scalp(s)*); **stam**(men) (*tribe(s)*); **stamhoofd**(en) (*tribal head(s)*); **wildeman**(nen) (*uncivilised man (men)*); **zigeuner**(s) (*gypsy (gypsies)*); |
| social | **baboe**(s) (*female servant(s)*); **barbaar**, barbaren (*barbarian(s)*); **bediende**(n) (*servant(s)*); **koeli**(es) (*contract worker(s)*); **piraat**, piraten (*pirate(s)*); **slaaf**, slaven (*slave(s)*); **slavenhandel**(s) (*slave trade*); |
| non-racial characteristics | **dwerg**(en) (*dwarf(dwarves)*); **hermafrodiet**(en) (*hermaphrodite(s)*); **mongool**, mongolen (*mongoloid(s)*); |
| sexual orientation | **homo**('s) (*gay person(s)*); **queer**(s) (*queer person(s)*); **trans** (*trans person(s)*); |
| place | **jappenkamp**(en) (*Japanese concentration camp(s)*); |
| religious | **islamiet**(en) (*muslim(s)*); **mohammedaan**, mohammedanen (*muslim(s)*); |

Table 1: Charged noun list. Word forms of each charged noun are aggregated and each aggregation is collected under its stemmed form (in bold).

| | sampled years (No. articles in millions [M]) | | |
|---|---|---|---|
| 1950s | 1951 (1.2M) | 1955 (1.4M) | 1959 (1.3M) |
| 1960s | 1961 (0.9M) | 1965 (0.9M) | 1969 (0.8M) |
| 1970s | 1971 (0.8M) | 1975 (0.7M) | 1979 (0.7M) |
| 1980s | 1981 (0.7M) | 1985 (0.7M) | 1989 (0.8M) |
| 1990s | 1991 (0.7M) | 1995 (0.3M) | |

Table 2: KB Newspaper Collection sampled years (taken in their entirety where publicly available), and corresponding number of articles rounded to the nearest 0.1M .

- The entire POS-tagged noun set from the spaCy-parsed dataset for each year, represents the *noun pool* from which to extract a corpus of modifier–head compound word pairs;

- The charged-words (including plural forms and variants) of Table 1 are used as seed heads:

**Steps**

- *Modifier extraction:* modifiers are harvested via trie-based character matching of the seed heads against the *noun pool*. Terminating hyphens are stripped from the modifiers. The output (modifiers) are filtered;

- *Head extraction:* heads are then harvested via trie-based matching of the previously harvested modifiers from the *entity pool*. Once again, hyphens are stripped and the output (heads) are filtered;

- *Final head–modifier extraction:* Repeating the *Modifier Extraction* step, a set of filtered set of head–modifier pairs is returned.

The filtering at each harvesting stage aims to improve the quality of the harvested heads and modifiers, by reducing the incidence of extracting false cases. Filtering consists of removing all heads or modifiers less than 3 characters in length or absent from the SoNaR-corpus [4] (ignoring case).

### 3.4 Building a corpus of noun–adjective pairs

Separately, for each of the years listed in Table 2, the corresponding spaCy dependency-parsed

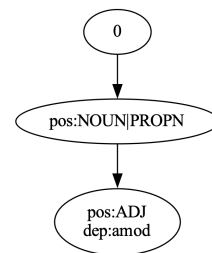[4] https://taalmaterialen.ivdnt.org/download/tstc-sonar-corpus/



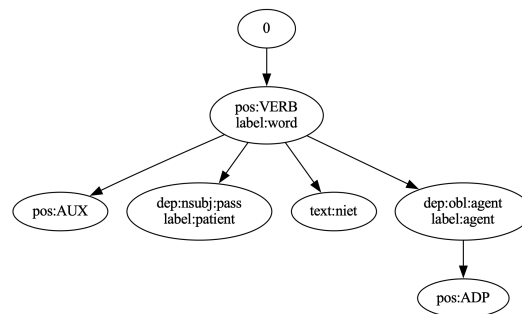Figure 1: Pattern A1 denoting the targeted adjective-noun relationship. '0' points to the root.



Figure 2: Pattern V1 denoting a targeted verb, auxiliary verb, agent, patient and preposition relationship. '0' points to the root. Negation is optionally matched. Pattern matching results in both verb, corresponding patient noun, corresponding agent noun. For example, Nu zullen de kleurlingen in de Senaat door één blanke senator worden vertegenwoordigd: yields vertegenwoordingen (verb), *senator* (agent noun) and *kleurlingen* (patient noun).

dataset is matched against the pattern tree shown in Figure 1. This pattern represents the simplest, most direct pattern for noun–adjective pair associations in the interest of high-accuracy results.

Noun and corresponding adjective pairs are returned. For the adjective, the lemma form is returned. For example, for the sentence fragment "Een op de vier vrouwelijke migranten werkt als ...", yields the noun-adjective (lemma) pair, *migranten–vrouwelijk* (*migrants–female*).

### 3.5 Building a corpus of noun–verb pairs

Separately, for each of the years in Table 2, the corresponding spaCy dependency-parsed dataset is subject to pattern matching against the pattern trees shown in Figures 2,3,4,5. The patterns are nested in their complexity, and hence patterns are grouped within tiers as shown in Figure 6. Each node in the dependency parse is compared against each pattern, capturing noun–verb pairs according to the highest-ranked matching pattern only.
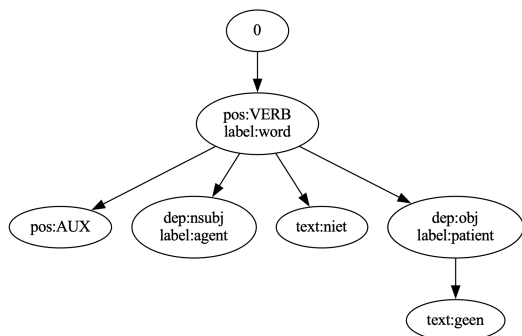
Figure 3: Pattern V2 denoting a targeted verb, auxiliary verb, agent, patient relationship. '0' points to the root. Negation is optionally matched. Resulting in verb–patient noun and verb–agent noun pairs. For example, de negers waren verdedigd door uit het Zuiden afkomstige blanke advocaten: yields *verdedigen* (verb), *advocaten* (agent) and textitnegers (patient).
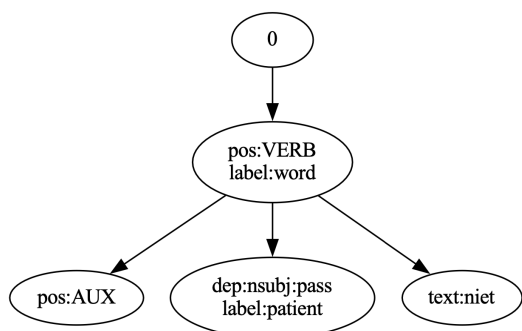


Figure 4: Pattern V3 denoting a targeted verb, auxiliary verb, patient relationship. '0' points to the root. Negation is optionally matched. Resulting in verb–patient noun pairs. For example, terwijl jaarlijks meer dan 150.000 immigranten worden toegelaten: yields *toelaten* (verb) and *immigranten* (patient noun).

## 3.6 Collective noun–context feature keyness scoring

The *keyness* scoring metric adopted in this paper, is the Log Likelihood Ratio (LLR) (Dunning, 1993). The resulting score is not based on normal approximations, and hence is applicable to low-frequency events commonly occurring in language and known generally as the Zipfian tail. The method can be thought of converting a frequency table, in our case of noun–context feature co-occurrences, to an equivalent table of scores reflective of the degree of association between the nouns and the context features. I.e., in our case, a high score reflects a context feature being particular important to the characterisation of noun.
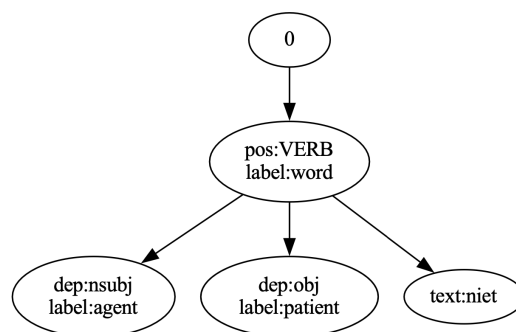
Effectively, we considered each noun and con-



Figure 5: Pattern V4 denoting a targeted verb, agent, patient relationship. '0' points to the root. Negation is optionally matched. Resulting in verb–patient noun and verb–agent noun pairs. E.g., waar de Berbers de Arabieren in aantal overtreffen: yields, *Berbers* (agent noun) and *overtreffen* (verb).

| count(context, noun) | count(context, noun') |
|---|---|
| count(context', noun) | count(context', noun') |

Table 3: Contingency table, forming the basis of the conversion of raw frequency table values of noun–context feature co-occurrence to LLR scores reflecting how key a context is to a noun.

text feature pair (cell) in the frequency table in turn, forming a contingency table as per table 3 for each.

The contingency table thus represents the binomial outcomes of the context occurring or not occurring with respect to two sub-corpora. The left-hand column of table 3 represents all instances for the context feature type and year, which is co-occurrent with the noun in question. The right-hand column of table 3, represents all instances for the context feature type and year, which is not co-occurrent with the noun in question.

To calculate LLR, two separate generative processes are considered for each sub-corpus. Firstly, that the two sub-corpora share a common binomial probability with respect to the occurrence of the context. Secondly, that the two corpora have different, distinct binomial probabilities with respect to the occurrence of the context in question. Maximum Likelihood Estimation (MLE) estimates of the binomial probabilities for both assumed generative processes are calculated.

The LLR value is then calculated via Equation 1, where Binom(x,y) denotes the binomial probability of the outcomes observed in sub-corpus x, assuming the parameters of the generative process, y, as previously described. A larger LLR value implies a greater co-location of the collective noun and context in question.
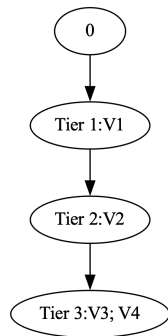
Figure 6: Pattern matching hierarchy: At each node in the spaCy dependency parse patterns are checked, moving through each tier until a pattern match is found and collecting all matches within that tier only.

| Pattern | Accuracy point estimate | Accuracy 95% Confidence Interval (Wilson) |
|---------|-------------------------|--------------------------------------------|
| Adjective Patterns | | |
| A1 | 97% (125) | 92 - 99 % |
| Verb Patterns | | |
| V1 | 95% (66) | 87 - 98 % |
| V2 | 91% (125) | 85 - 95 % |
| V3 | 94% (125) | 89 - 97 % |
| V4 | 88% (125) | 81 - 93 % |

Table 4: The results of manual evaluation of accuracy of the extracted noun–adj, and noun–verb pairs, due to combined dependency parse and pattern-matching errors. Results are rounded to 2 significant figures. The number of sample extracts for each pattern type are given in the brackets.

$$-2.log(\frac{Binom(1,1) \times Binom(2,1)}{Binom(1,2) \times Binom(2,2)}) \quad (1)$$

## 4 Evaluation and Results

The substantive output of the methodology of Section 3 are the tables of LLR associativity scores for each noun–context feature pair. These LLR scores are the basis for the evaluation methods in this section.

### 4.1 Pattern-matching accuracy

First, some evaluation of the accuracy of the noun–adjective and noun–verb pattern matching methodology is warranted. There are three main potential sources for error in the extracting pairs of adjectives or verbs and corresponding nouns as described in Section 3.4 and Section 3.5: OCR errors, dependency parse errors and pattern matching errors. OCR errors do not present a significant concern to this study, beyond their influence on dependency parsing performance. I.e., if a misrepresented word or artifact which otherwise looks like an adjective in terms of syntax, and is dependency parse tagged and pattern matched as such, then we simply end up with an extra nonsense context word.

Whether the pattern matching fails to correctly extract true noun–context word instances presents a greater concern. This was evaluated manually by sampling the noun–context word extracts via pattern matching of a random sample of 200 articles from the 1991 OCR set. The results are given in Table 4, and demonstrate a reasonably strong base accuracy with estimates ranging from 88% with the V4 pattern to 97% with the A1 pattern, support-

ing further conclusions derived from the noun and adjective or verb pair co-occurrence statistics.

### 4.2 Identifying high-association contexts for known-charged collective nouns

Our first research question, *do context features exist, which for some noun–context feature associativity score threshold, are highly predictive of the noun?* can be considered as a fundamental test of the base hypothesis that the methodology and dataset are sufficient to identify relevant and interesting high-association terms. It is fundamental that we can identify high-association contexts for known-charged collective nouns. We cannot draw effective parallels between terms with respect to their context features if they do not have sufficiently strong profiles.

For this research question, we adopt a high LLR threshold: For each year and for each collective noun in Table 1, we extract only those context features for which the collective noun is in the top 2 of LLR scores. For a selected number of known-charged collective nouns, the outcomes are given in Table 5. It should be reiterated here that the table is not a complete window into the all context features with a high degree association, merely those with an extremely high degree of association according to the LLR threshold. Clearly relevant, strong outcomes can be observed from this. I.e., in the case of the charged-noun, *"migrant"*, we see contextual features such as *aspirante* (aspirational), *tweede-generatie* (second-generation), *niet-geïntegreerd* (unintegrated). In the case of *"baboe"* (the general name given to nannies from Surinam), we see *zorgvol* (caring). In the more powerfully charged cases such as *"neger"*, we see a wealth of strong

| known-charged nouns | modifiers associated | adjectives associated | verbs associated for which the noun is the agent | verbs associated for which the noun is the patient |
|---|---|---|---|---|
| afkomst | bedoeïnen: 1991; bloedgroep: 1959; dansers: 1979; hugenoten: 1989; huurkamer: 1971 | | aanvuren: 1969; held: 1985; molesteren: 1995; verzekeren: 1995 | eemt: 1959; opsieren: 1971; raden: 1959; schreeuwen: 1991; traumatiseren: 1985; verloochenen: 1995, 1951, 1955, 1959, 1959, 1961, 1961, 1965, 1969, 1969, 1975, 1975, 1979, 1979, 1981, 1985, 1985, 1989, 1989, 1991, 1991, 1995; verraden:1955, 1959, 1985, 1989, 1991, 1995; zullen: 1971 |
| allochtoon | | laag-opgeleid: 1991; werkloos: 1989; werkwillig: 1991 | ongemerkt: 1991 | hulpbehoeven: 1989; instromen: 1989 |
| baboe | | soendanees: 1991; zorgvol: 1979 | inhalen: 1951 | |
| immigrant | commonwealth: 1971 | afriaans: 1991; arriveren: 1951; bengaals: 1981, 1985; blank: 1981; duits-joods: 1971; engtls: 1955; enjels: 1951; hds: 1959; illegaal: 1971, 1979, 1995; indies: 1975; latiinsamerikaans: 1991; millioen: 1955; mohammedaans: 1991; niet-blanke: 1965, 1981; niet-britse: 1955; niet-geïntegreerd: 1959; nietblank: 1965, 1985; nlet-blank: 1965; noordafrikaans: 1995; opper-egyptisch: 1981; ouz: 1959; portoricaans: 1959; roemeens-duits: 1989; russisch-joods: 1989; russischjoods: 1989; s'amitisch: 1995; salvadorlaans: 1969; sicillaans: 1961; sovjet-joods: 1991; steenrijk: 1969; steunen: 1951; urdu-talig: 1995; westindlsch: 1981; ülegal: 1981 | afpakken: 1989; binnensmokkelen: 1971; broeden: 1959; doodsteken: 1995; inpikken: 1989; klagen: 1951; meedragen: 1955; omsingelen: 1959; ontsluiten: 1981; overspoelen: 1989; terugbetalen: 1955; wegpikken: 1959; zjn: 1961 | aankomen: 1975; afbeulen: 1985; classificeren: 1979; huisvesten: 1991; stijven: 1985; toelaten: 1951; verkijken: 1959 |
| indiaan | amazone: 1969; apache: 1991; halfbloed: 1981; hopi: 1961, 1985; innu: 1989; miskito: 1985, 1989; navajo: 1991; noorda: 1959; oerwoud: 1981; platvoet: 1955, 1965; sioux: 1991; yanomami: 1991 | amazon: 1989; benedenlands: 1965; bonairiaans: 1989; een-ogig: 1965; eenogig: 1965; eht: 1961; grondloz: 1979; ploeteren: 1991; rillen: 1959 | aftroggelen: 1995; kapen: 1989; kauwen: 1991 | achterstellen: 1989; afhakken: 1951; afslachten: 1969; hakken: 1965; verontwaardigen: 1989 |
| islamiet | | dox: 1979; fundamentalisch: 1989; fundamentalistisch: 1981, 1985, 1989; imam: 1991; listisch: 1991; niet-chinees: 1989; radicaal: 1995; rechtgelovig: 1989; rechtzinnig: 1979; sjiietisch: 1979; sjiïtisch: 1979; sunnitisch: 1989; ugandees: 1989; weerspanning: 1959 | begraven: 1985; ijgen: 1989; vasten: 1979 | vluchtelingenkamp: 1985 |
| kaffer | zoeloe: 1961, 1959, 1951 | nagemaakte: 1961; roodgeverfd: 1961; tomm: 1951 | | |
| kleurling | élite: 1975 | biaziliaans: 1959; en'ander: 1955; fransi: 1955; inder: 1985; kaaplands: 1955; kaaps: 1951, 1955, 1969; kroesharig: 1959; opdringerig: 1981; tussen-d: 1981 | geincasseerd: 1955; herkrijgen: 1961; overlopen: 1991 | afbeelden: 1961; idealiseren: 1979; integreren: 1975; selecteren: 1979; tusaen: 1951; verwarren: 1955; volmaken: 1955 |
| koeli | raat: 1975; riksha: 1961; riksja: 1959, 1969 | doodarm: 1965; halfnaakt: 1965; rijkgekled: 1955 | fouilleren: 1951; splijten: 1951; voorttrekken: 1961 | |
| migrant | aspirante: 1951, 1955, 1959, 1989; heimwee: 1971; illen: 1981; irh: 1965; lrn: 1995; niete: 1951, 1955; plattelandse: 1955; proefe: 1959; spyt: 1961; tweede-generatie: 1995 | afro-caribisch: 1995; erkloz: 1991; haitiaans: 1979; marokkaans: 1989, 1991; niet-geïntegreerd: 1991; ondef: 1959; onvolwaardig: 1991; rokkaans: 1989; turks: 1989 | | |
| mohammedaan | sjia: 1965 | anti-eommunistisch: 1965; fans: 1955; inpopulair: 1971; kameroens: 1979; orthodox-radical: 1955; pro-frans: 1959, 1961; sjiitisch: 1975 | bestrijden: 1995 | |
| neger | bakongo: 1959; bantoe: 1955; benton: 1959; bos: 1955, 1975, 1989; congo: 1955; goudkust: 1955; grun: 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991; irun: 1951; mississippi: 1955; overleden: 1959; panko: 1951; soedan: 1961, 1965; soweto: 1979; Watts: 1965 | abject: 1951; afrikaans: 1955; afro-amerikaans: 1995; amerikaane: 1959; armlastig: 1965; bevoogden: 1965; blootvoet: 1955; diepbruin: 1965; eerbaar: 1995; golv: 1961; gracieus: 1985; ifrikaans: 1959; kiesgerechtigd: 1965; langbenig: 1955; lynchen: 1959; militant: 1971; miloz: 1965; negenenvijftig: 1981; noord-amerikaans: 1955; onschendbaar: 1961; oproerig: 1971; overigen: 1961; rfd: 1971; seigneurial: 1961; senegalees: 1951; sluip: 1955; stokoud: 1985; tiaans: 1979; west-afrikaans: 1955; wetsgetrouw: 1961; zelfbewust: 1969; zuidrhodesisch: 1965; zuidsoedanes: 1971; üntwikkel: 1965 | aandrukken: 1969; bijeenrapen: 1965; ebben: 1951; hf: 1961; inj: 1969; inladen: 1975; openscheuren: 1969; plunderen: 1969; straffen: 1965; toebedelen: 1955; t ransponeren: 1955; uitzingen: 1965 | aftuigen: 1965; besprenkelen: 1979; canoniseren: 1961; contra: 1959; doodschieten: 1965, 1981; executeren: 1951; gelijkberechtigd: 1965; inschepen: 1955; inschrijven: 1965; kamperen: 1951; lynchen: 1951, 1959; roven: 1979; slaven: 1951; terechtstellen: 1959; tiranniseren: 1961; toerekenen: 1959; uitgevlucht: 1981; uitmoorden: 1969; verafschuwen: 1969; verdrukken: 1965; vermengen: 1961; verschillen: 1955; voortrekken: 1969; weren: 1955 |

Table 5: Selected known-charged nouns of table 1, with together with (all) context features for which the noun-context LLR associativity score is in the top 2 for that context feature.

| Modifiers | known-charged nouns associated with the modifier according to the criteria of 4.3 (as head in the compound word) |
|---|---|
| *"nomaden"* | indiaan: 1965<br>stam: 1965, 1951, 1955, 1959, 1961, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995 |
| magazijn | bediende: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995 |
| *"indianen"* | stam: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995 |
| *"neger"* | slaaf: 1955, 1971, 1979, 1981, 1975, 1985, 1989, 1991, 1995<br>stam: 1955, 1971, 1979, 1981, 1951 |
| boom | stam: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995 |
| *pape* | ras: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991 |
| *grun* | neger: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991 |
| pomp | bediende: 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991, 1995 |
| joego | slaaf: 1951, 1955, 1959, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1991 |
| bantoe | bediende: 1961<br>neger: 1959, 1961, 1955<br>ras: 1959, 1951<br>stam: 1959, 1961, 1951, 1955, 1971, 1975 |
| *zeeg* | ras: 1951, 1955, 1959, 1961, 1965, 1969, 1975, 1981, 1985, 1989, 1991, 1995 |
| bosland | creool: 1989, 1955, 1959, 1965, 1969, 1975, 1979, 1991, 1995<br>indiaan: 1989<br>neger: 1989, 1955 |
| ether | piraat: 1959, 1961, 1969, 1971, 1975, 1979, 1981, 1985, 1989, 1995 |
| *mons* | trans: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1975, 1979 |
| hatte | ras: 1951, 1955, 1959, 1965, 1969, 1971, 1975, 1981, 1985 |
| *achte* | ras: 1959, 1965, 1969, 1971, 1975, 1979, 1981, 1985, 1991 |
| berber | stam: 1951, 1955, 1959, 1961, 1965, 1969, 1971, 1985, 1989 |
| zoeloe | kaffer: 1961, 1951<br>neger: 1961<br>stam: 1961, 1979, 1981, 1985<br>stamhoofd: 1961, 1989 |
| loket | bediende: 1959, 1961, 1965, 1971, 1975, 1979, 1981, 1985, 1995 |
| bacterie | stam: 1955, 1975, 1979, 1981, 1985, 1989, 1991, 1995 |
| dart | moor: 1951, 1955, 1959, 1961, 1965, 1979, 1989, 1991 |
| voortrekkers | stam: 1951, 1955, 1959, 1961, 1965, 1971, 1985 |
| bosneger | stam: 1989, 1959, 1975, 1979, 1991, 1995<br>stamhoofd: 1989 |
| papoea | stam: 1951, 1955, 1959, 1965, 1971, 1981, 1989 |
| bos | neger: 1951, 1955, 1969, 1975, 1979, 1989 |
| bedoeïenen | stam: 1959, 1961, 1971, 1985, 1991, 1995 |
| *ex-e* | migrant: 1955, 1961, 1965, 1985, 1989, 1991 |
| amazone | indiaan: 1961, 1969, 1975, 1979, 1991<br>stam: 1965 |
| bel | indo: 1979, 1981, 1985, 1989, 1991, 1995 |
| kantoor | bediende: 1951, 1955, 1959, 1961, 1965, 1969 |

Table 6: known-charged nouns (as compound word heads), and the common modifier they are associated with according to the criteria in 4.3. This table only lists instances of 6 or more associated noun and year instances. Modifiers that are themselves known-charged words are marked as such; italicized strings are decomposition errors.

context features, such as: tribal names, *lynchen* (to lynch), *militant* (militant), *kiesgerechtigd* (being eligible to vote), *executeren* (to execute) and *plunderen* (to plunder).

### 4.3 Identifying context features for multiple known-charged collective nouns

Our second research question, *are there context features which for some noun–context feature associativity threshold, recall multiple known charged nouns with a reasonable degree of precision with respect to our known charged noun set?* is concerned with whether the methodology is able to find common, meaningful associations that hold across known-charged words. To consider contextual feature overlap between known-charged collective nouns, we must adopt a less severe criterion allowing for overlap. For each context feature type (e.g., modifiers), for each year and for each context feature, the corresponding collective nouns are traversed, according to their descending LLR score, and every noun above a LLR threshold is acccepted. This results in a precision of 0.2, taking the Table 1 known-charged nouns as true positives.

Sample outcomes of this approach are given in Tables 6, 7, 8 and 9, corresponding to modifiers, verbs for which the noun is the patient, verbs for which the nouns are the agent and adjectives. Each table represents a selection of context features from a larger set, listing for those context features with the most year–context feature instances associated. The tables are otherwise in no way curated. Examination of the tables again shows some powerful associations between known-charged words over time frames. For example: in Table 7, *"gekleurd"* (coloured), *"immigrant"* (immigrant) and *"zigeuner"* (gypsy) peoples being subject to *deporteren* (to deport); in Table 8, *"immigrant"* (immigrant) peoples in 1975 and 1995 are associated with action of *overstromen* (to flood); and in Table 9: *"indiaan"* (indian) and near continuously over a large time window, *"stam"* (tribe) associated with *"primitief"* (primitive).

Table 6 shows that some of the modifiers that are discovered are known-charged words themselves. Table 6 also includes a number of modifiers that do not refer to a strongly related word, but are the result of an incorrect morphological decomposition; e.g. the charged word *"ras"* was mistakenly detected in words ending in the stem *as* (axis) or *gras* (grass), producing the incorrect assumed modifiers *zeeg* and *achte*. Either a lexical filter or a better morphological decomposition would allow filtering out these cases.

### 4.4 Discovering charged nouns from their common associations with known-charged nouns

Our final research question is *"Can the context features of known charged nouns, help identify other charged instances?"*. Considering each year, and

| Verbs for which the noun is the patient | known-charged nouns associated with the verbs according to the criteria of 4.3 (and the years they are associated) |
|---|---|
| NOTverloochenen | afkomst: 1951, 1955, 1959, 1961, 1965, 1969, 1975, 1979, 1981, 1985, 1989, 1991, 1995 |
| verloochenen | afkomst: 1969, 1989, 1995, 1951, 1959, 1961, 1975, 1979, 1985, 1991<br>oorsprong: 1969, 1995<br>roots: 1989 |
| verraden | afkomst: 1955, 1965, 1975, 1981, 1985, 1989, 1991, 1995 |
| lynchen | neger: 1951, 1959, 1975, 1995 |
| slaven | bediende: 1985<br>neger: 1951, 1971, 1975 |
| voortrekken | islamiet: 1969, 1955<br>neger: 1969, 1975 |
| terechtstellen | neger: 1951, 1959, 1965 |
| doodschieten | neger: 1959, 1965, 1971 |
| fokken | ras: 1959, 1961, 1975 |
| ronselen | inboorling: 1979<br>indiaan: 1979<br>koeli: 1951 |
| achterstellen | bosneger: 1989<br>indiaan: 1989<br>zigeuner: 1989 |
| uitroeien | indiaan: 1989<br>stam: 1989<br>zigeuner: 1959 |
| NOTverwarren | kleurling: 1955<br>primitief: 1975<br>ras: 1959 |
| deporteren | gekleurd: 1969<br>immigrant: 1985<br>zigeuner: 1979 |
| legaliseren | immigrant: 1955<br>piraat: 1981<br>zigeuner: 1981 |

Table 7: known-charged nouns, associated verbs and the years of association according to the criteria defined in 4.3, where the noun is the patient to the verb. The *NOT* prefix denotes negation of the verb. This table lists only those instances of 3 or more associated noun and year instances.

| Verbs for which the noun is the agent | known-charged nouns associated with the verbs according to the criteria of 4.3 (and the years they are associated) |
|---|---|
| enteren | piraat: 1955, 1959, 1989 |
| doodsteken | bediende: 1989<br>immigrant: 1995<br>wildeman: 1989 |
| NOTleven | blanke: 1971, 1991<br>piraat: 1969 |
| neerzetten | bediende: 1985<br>inboorling: 1951<br>zigeuner: 1979 |
| inbegrijpen | homo: 1961<br>koppensneller: 1961 |
| infecteren | neger: 1961<br>ras: 1991 |
| ongemerkt | allochtoon: 1991<br>neger: 1971 |
| serveren | bediende: 1955, 1959 |
| uitgooien | neger: 1991<br>piraat: 1961 |
| uitmoorden | blanke: 1969<br>indiaan: 1995 |
| NOTvergeten | immigrant: 1955<br>zigeuner: 1961 |
| herkrijgen | kleurling: 1961<br>slaaf: 1979 |
| uitzingen | neger: 1959, 1965 |
| aanbidden | blanke: 1959<br>slaaf: 1969 |
| verkrachten | piraat: 1989<br>wildeman: 1981 |
| stichten | immigrant: 1989<br>stam: 1989 |
| boren | piraat: 1979<br>stam: 1965 |
| zeulen | dwerg: 1955<br>inboorling: 1959 |
| kidnappen | indiaan: 1985<br>stam: 1985 |
| NOTdrinken | indiaan: 1985<br>mohammedaan: 1961 |
| binnensmokkelen | immigrant: 1971, 1981 |
| bejegenen | barbaar: 1961<br>kleurling: 1971 |
| overstromen | immigrant: 1975, 1995 |

Table 8: Known-charged nouns, associated verbs, and the years of association according to the criteria defined in 4.3, where the noun is the agent to the verb. The *NOT* prefix denotes negation of the verb. This table lists only those instances of 2 or more associated noun and year instances.

each context feature separately, and setting an LLR threshold with respect to the context feature as described in section 4.3, all corresponding nouns are extracted. Where a noun is coincident with a known-charged noun of Table 1, this pairwise association is recorded, together with the context feature and year responsible for the association.

The leftmost column of Table 10 provides clues for answering our third sub-question: whether we can automatically discover new candidate terms for our charged word list. The column in the table exhibits a small outtake of a list of 6,310 unique words that frequently occur in the same morpho-syntactic role as our charged words, along with their specific linguistic contexts. A manual inventory of this word list reveals a candidate set of about 10 new charged terms, including ***"joden"*** (*jews*), ***"indianen"*** (*indians*), ***"moslims"***) (*muslims*), and ***"slaviër"*** (*slav*). Other charged terms occurring in this list refer to nazism and radical movements such as ***"SS"*** and ***"RAF"***, and include formerly used terms for immigrant workers, such as ***"gastarbeider"*** (literally *guest worker*, *immigrant worker*). It takes manual inspection and expertise to extract

these term from this larger list of terms, of which the majority consists of general, uncharged, high-frequency words for family relations, demographic groups, locations, government, occupations, culture, religion, tradition, and arts — all to be expected, given that these are all hypernyms of our charged terms and occur in the same linguistic and semantic contexts.

## 5 Discussion and Conclusion

The paper posed three research questions which we can paraphrase as: do simple metrics of word associativity yield distinctive context profiles; can these context profiles be used to draw parallels between known-charged nouns; and finally, can we identify candidate charged nouns. Somewhat inherent to the complexity of the notion of a term being *charged* is that there exists no definitive gold standard dataset from which we are able to evaluate

| Adjectives | known-charged nouns associated with the adjectives according to 4.3 criteria (and the years they are associated) |
|---|---|
| indiaans | afkomst: 1969, 1979, 1989, 1955, 1975, 1985, 1991, 1961<br>halfbloed: 1959<br>medicijnman: 1979, 1989, 1991, 1995<br>ras: 1979<br>scalp: 1969<br>slaaf: 1969<br>stam: 1969, 1979, 1959, 1989, 1955, 1975, 1985, 1971<br>stamhoofd: 1969, 1959 |
| germaans | afkomst: 1959<br>barbaar: 1959<br>oorsprong: 1959, 1969<br>ras: 1959, 1969, 1951, 1961, 1965, 1975, 1979, 1985, 1989<br>stam: 1959, 1969, 1951, 1961, 1965, 1975, 1979, 1985, 1989, 1955, 1971, 1991, 1995 |
| arisch | afkomst: 1951, 1955<br>ras: 1959, 1965, 1969, 1975, 1979, 1981, 1985, 1989, 1991, 1995<br>stam: 1951 |
| hindostaans | Afkomst: 1965, 1969, 1951, 1955, 1959, 1961, 1975, 1979, 1985<br>immigrant: 1965, 1969<br>migrant: 1989 |
| primitief | indiaan: 1979<br>stam: 1979, 1951, 1959, 1965, 1969, 1971, 1975, 1981, 1985, 1989, 1995 |
| resistent | ras: 1985, 1989, 1955, 1959, 1961, 1965, 1971, 1991<br>stam: 1985, 1989 |
| nederig | afkomst: 1981, 1955, 1959, 1969, 1971, 1979, 1985, 1989, 1991<br>slaaf: 1981 |
| russisch-joods | afkomst: 1985, 1989, 1991, 1961, 1971<br>immigrant: 1985, 1989, 1991, 1979<br>oorsprong: 1985 |
| polair | oorsprong: 1951, 1959, 1961, 1965, 1969, 1971, 1975, 1979, 1981, 1985 |
| minderwaardig | ras: 1961, 1965, 1969, 1975, 1979, 1981, 1985, 1991, 1995 |
| indo-europees | afkomst: 1971, 1991, 1955, 1979, 1985<br>oorsprong: 1971<br>stam: 1991, 1961 |
| pools-joods | afkomst: 1991, 1951, 1979, 1981, 1985, 1989<br>immigrant: 1991, 1959 |
| niet-nederlands | afkomst: 1989, 1991, 1979, 1985, 1995<br>immigrant: 1991<br>oorsprong: 1989 |
| armeens | afkomst: 1965, 1969, 1975, 1981, 1985, 1995<br>immigrant: 1965 |
| subtropisch | oorsprong: 1959, 1961, 1965, 1969, 1971, 1975, 1981 |
| goddelijk | oorsprong: 1959, 1965, 1969, 1971, 1981, 1989, 1991 |
| duits-joods | afkomst: 1955, 1965, 1961, 1995<br>immigrant: 1955, 1965, 1971 |
| noordafrikaans | afkomst: 1991, 1995<br>immigrant: 1991, 1995, 1985, 1989<br>migrant: 1991 |
| oriëntaals | afkomst: 1971, 1981<br>immigrant: 1955, 1971<br>oorsprong: 1975<br>ras: 1955 |
| illegaal | immigrant: 1955, 1971, 1979, 1985, 1989, 1995 |
| negroïde | afkomst: 1989<br>ras: 1971, 1989, 1975, 1991<br>stam: 1971 |
| keltisch | oorsprong: 1955, 1959, 1965, 1969<br>ras: 1951<br>stam: 1951 |
| NOTnederlands | afkomst: 1989, 1991, 1979<br>oorsprong: 1989, 1991, 1961 |

Table 9: Known-charged nouns, associated adjectives and the years of association according to the criteria defined in 4.3. This table lists only those instances of 6 or more associated noun and year instances.

the methodology output on a purely numerical basis. Charged term detection, and an understanding of the manifest attributes that make terms charged, remains an open problem (and perhaps always will be). Consequently, any evaluation of methods used to answer the research question must inevitably rely on a degree of outside-of-data, human interpretation. On the basis of the observed associations and the links we can recognize, we contend that

the evaluation results are sufficiently strong to be able to answer all of the research questions in the affirmative. Additionally, the results in regards to supporting the methodology are supported by the fact that ultimately the basis of methods is simple, time-tested, and entirely open to inspection (being based on co-occurrence counts).

The underlying context in which the research questions were posed, was the application of digital humanities to help humanities scholars in exploring and charged language. The utility being the ability of condense many millions of narrative descriptions into a much smaller number of salient associations for human consideration. In this regard, the evaluation results tables in this document (and the complete versions, with english translations, available on the Github repository), can be viewed as reference set of associations. However, the results correspond to the specific (and arguably quite restrictive) LLR associativity score thresholds adopted for the purpose of method evaluation. It is envisaged that the methodology could be used on a more adhoc basis by humanities scholars in exploring context features and overlaps: where the outputs could be used as a both a reference with a probabilistic basis, but also as a pointer to consider axes of contentiousness at a high, human-expert level. For example, in the Words Matter publication in relation to **"stam"** it is noted that (translation): "The term tribe is often associated with a so-called not complex society with a simple political structure. although this fact in itself is not disputed, the term has the connotation of primitive". We see this precise association in our results: in Table 9, the adjective and known-charged term **"primitief"** is shown to be associated with **"stam"** in the newspaper articles consistently through the 1950s through the 1990s. In the case of **"mohammedaan"**, the Words Matter document details objections to the term on the basis of religious objects: but we also see context associations such as *orthodox-radical* (Table 5) which may or not provide further avenue for which contentiousness its contentiousness can be considered. In the case of **"neger"**, the Words Matter document notes the associations of the word with the sub-Saharan African peoples, but more problematically with racial stereotyping. Again, we see this as an output from the methodology in the table 5 profile of the term: *bakongo, bantoe, congo, goudkust, soedan; blootvoet, lynchen, militant*. Furthermore, the results of Table 5 allow us

| noun | known-charged noun | verbs associated for which the nouns are patients | verbs associated for which the nouns are agents | adjectives associated | modifiers associated |
|---|---|---|---|---|---|
| bevolking | afkomst | | | indiaans: 1969, 1979, 1989, 1955, 1975, 1985, 1991 / hindostaans: 1965 / creools: 1959 / hindoestaans: 1961 / papoeaas: 1961 / albanees: 1981, 1989 | |
| bevolking | bediende | opschrikken: 1971 | | niet-blanke: 1971 | bantoe: 1961 / heger: 1969 |
| bevolking | blanke | | | kiesgerechtigd: 1985 | |
| bevolking | creool | | | | bosland: 1989, 1975 |
| bevolking | halfbloed | | | indiaans: 1959 | |
| bevolking | immigrant | | | hindostaans: 1965 / niet-blanke: 1971, 1965, 1981 / nietblank: 1965, 1985 / straatarm: 1981 | |
| bevolking | indiaan | uitroeien: 1989 | | | bosland: 1989 |
| bevolking | inlander | ophitsen: 1959 | | | |
| bevolking | kleurling | | | kiesgerechtigd: 1985 | |
| bevolking | koppensneller | | | maleis: 1951 | |
| bevolking | medicijnman | | | indiaans: 1979, 1989, 1991, 1995 | |
| bevolking | migrant | | | nietblank: 1971 | |
| bevolking | neger | ophitsen: 1959 | | autochthon: 1961 / kiesgerechtigd: 1965 | bantoe: 1959, 1961 / bosland: 1989 |
| bevolking | ras | | | indiaans: 1979 / negroïde: 1971 | bantoe: 1959 |
| bevolking | scalp | | | indiaans: 1969 | |
| bevolking | slaaf | | | indiaans: 1969 | neger: 1955, 1971, 1979, 1975, 1985 |
| bevolking | stam | uitroeien: 1989 / geevacueerd: 1965 / uitmoorden: 1985 | | indiaans: 1969, 1979, 1959, 1989, 1955, 1975, 1985, 1971 / berbers: 1955 / negroïde: 1971 / inheems: 1965, 1981 | bantoe: 1959, 1961, 1975 / neger: 1955, 1971, 1979 / bosneger: 1989, 1991, 1995 / papoea: 1951, 1955, 1959, 1965, 1989 / nomaden: 1969 / eskimo: 1985 |
| bevolking | stamhoofd | | | indiaans: 1969, 1959 / berbers: 1955 | bosneger: 1989 |
| beweging | blanke | | | | mau-mau: 1959 |
| beweging | inboorling | | | oproerig: 1959 | |
| beweging | indiaan | | | opstandig: 1989 | |
| beweging | islamiet | | | fundamentalistisch: 1985, 1989, 1991 | |
| beweging | migrant | | | russisch-talig: 1989 | |
| beweging | neger | | | oproerig: 1959 / opstandig: 1971 | |
| beweging | oorsprong | | | vincentiaans: 1971 | |
| beweging | stam | | | opstandig: 1971, 1959 / oproerig: 1955 | zulu: 1991 / zoeloe: 1985 |
| beweging | stamhoofd | | | | zulu: 1991 |
| joden | afkomst | | | oriëntaals: 1971, 1981 / hongaars: 1969 | |
| joden | blanke | | | | nlet: 1961 |
| joden | gekleurd | deporteren: 1969 | | | |
| joden | immigrant | deporteren: 1985 | | oriëntaals: 1971 / ethiopisch: 1985, 1991 | |
| joden | islamiet | | | orthodox: 1981 | |
| joden | neger | lynchen: 1975 | | | |
| joden | oorsprong | | | oriëntaals: 1975 | |
| joden | stam | uitmoorden: 1985 | | | |
| joden | zigeuner | deporteren: 1979 | | staatloos: 1961 | |
| kwestie | afkomst | | | c'al: 1991 | |
| kwestie | blanke | | | rhodesisch: 1981 | |
| kwestie | kaffer | | | | zoeloe: 1961 |
| kwestie | neger | | | | zoeloe: 1961 / soedan: 1951 |
| kwestie | stam | | | | zoeloe: 1961 |
| kwestie | stamhoofd | | | | zoeloe: 1961 |
| communisten | barbaar | | | bloeddorstig: 1965 | |
| communisten | dwerg | | | bloeddorstig: 1965 | |
| communisten | indiaan | | uitmoorden: 1995 | | |
| communisten | islamiet | | | dox: 1979 | |

Table 10: Pairs of known-charged and other nouns as related by verbs, adjectives and modifiers, according to the associativity criteria defined in 4.4. This table represents only a demonstrative sample.

to extend the characterisation with detailed actions this collective noun term has been subjected to: *doodschieten* (shoot dead), *terechtstellen* (execute), *uitmoorden* (massacre) and *verdrukken* (oppress).

There is scope to further elaborate on, and strengthen the resulting context-feature profiles captured over a corpus. First and foremost, further work into the pattern matching routines such to expand the number of adjectives and verbs cap-

tured, whilst maintaining a high degree of accuracy. This is especially true of some of the most obvious and basic noun and verbs for which the noun is agent patterns, which we excluded from this study for yielding notably lower accuracy than other the patterns included in the study. However, there are other contexts that may be interesting and indicative of being charged: for instance context features which capture more information of the environs as

part of the narrative account of nouns (and known-charged nouns).

Lastly, whilst newspapers represent one particular narrative account type of people groups, other discourse types (such as literature) may yield rival or complementary accounts useful to humanities scholars.

## Acknowledgements

## CRediT Author Statement:

## References

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

Gioele Barabucci, Francesca Tomasi, and Fabio Vitali. 2020. Supporting complexity and conjectures in cultural heritage descriptions. In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, volume Vol-2810 of *1613-0073*, pages 104–115, Leiden, the Netherlands.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184.

Ryan Brate, Andrei Nesterov, Valentin Vogelmann, Jacco van Ossenbruggen, Laura Hollink, and Marieke van Erp. 2021. Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus. In *Proceedings of the 11th on Knowledge Capture Conference*, K-CAP '21, pages 17–24, Virtual Event, USA. Association for Computing Machinery.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. 2023. Unsilencing colonial archives via automated entity recognition. *Journal of Documentation*, ahead-of-print(ahead-of-print).

Wayne Modest and Robin Lelijveld. 2018. Words matter: an unfinished guide to word choices in the cultural sector. Technical report, The National Museum for World Cultures (Tropenmuseum, Afrikamuseum, Museum Volkenkunde, Wereldmuseum).

Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684.

Temi Odumosu. 2020. The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons. *Current Anthropology*, 61(S22):S289–S302.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Melvin Wevers. 2019. 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*, pages 92–97, Florence, Italy. Association for Computational Linguistics.

---

[5]https://credit.niso.org