ConConCor Dataset

Ryan Brate,[1] Andrei Nesterov,[2] Valentin Vogelmann,[1]
Laura Hollink,[2] and Marieke van Erp[1]

[1]KNAW Humanities Cluster
DHLab
Oudezijds Achterburgwal 185
1012 DK Amsterdam
The Netherlands
{ryan.brate,valentin.vogelmann,
marieke.van.erp}@dh.huc.knaw.nl

[2]CWI
Human-Centered Data Analytics
Science Park 123
1098 XG Amsterdam
The Netherlands
{a.nesterov,l.hollink}@cwi.nl

Motivation

1.1 For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
To enable exploration into the role context plays in modifying a word's perceived contentiousness and for the training of machine learning models for the detection of (Dutch language) contentious words in textual contexts.
1.2 Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The named researchers responsible for the creation of the dataset are part of the *Cultural AI*[1] lab: a pan-organisational lab, concerned with the application of culturally-aware AI in the humanities and cultural heritage domain.
1.3 Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
The dataset was created as a winning proposal in the EuropeanaTech Challenge for Europeana Artificial Intelligence and Machine Learning datasets.[2] The €2500 awarded was used to fund crowd-sourced annotations as well as rewards for expert annotators.
1.4 Any other comments?
Not applicable.

Composition

2.1 What do the instances that comprise the dataset represent (e.g.,documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description
The dataset is split into 4 sub-sets: • Extracts.csv: Dutch Language Europeana Newspaper collection[3] extracts. In the expert and crowdsourcing studies, each extract consisted of 5 sentences centred around a highlighted target word. Due to copyright, the extracts in the available dataset were shortened to max 140 characters (including whitespace). For each extract, we provide a link to Delpher with a newspaper article (scan and OCR) from which an extract was taken. The target words are reproduced in Appendix A. The majority of the target words are taken from the Words Matter publication,[4] "An Unfinished Guide to Word Choices in the Cultural Sector" by the Research Centre for Material Culture. Additional target words, from participant suggestions from Study 1, were incorporated as extra target words for

[1] https://www.cultural-ai.nl

[2] https://pro.europeana.eu/post/announcing-the-europeanatech-challenge-for-europeana-artificial-intelligence-and-machine-learning-datasets

[3] https://www.europeana.eu/en/collections/topic/18-newspapers

[4] https://www.materialculture.nl/en/publications/words-matter

the larger Study 2 (both studies are described in 3.1).

The extracts (containing the target words) are taken from OCR versions of the Europeana Newspaper collection, as provided by KB National Library of the Netherlands.[5]

- Annotations.csv: Anonymised participant multi-choice responses; in being asked to define whether the target word in the given textual context is *contentious* (to even the slightest degree), according to present-day sensibilities. Specifically, the multiple-choice options for each extract are as follows:
  - "Omstreden naar huidige maatstaven" ("Contentious according to current standards");
  - "Niet omstreden" ("Not contentious");
  - "Weet ik niet" ("I don't know");
  - "Onleesbare OCR" ("Illegible OCR")..

  Additionally, participants were asked to suggest other words within the context window that they consider contentious.

  Appendix C and D contain the annotation instructions and an example extract/ multiple-choice options, respectively.

- Demographics.csv: Anonymised participant demographic data (available for crowdsourcing participants from the Prolific[6] platform).

- Metadata.csv: metadata corresponding to the extracts in Extracts.csv. This metadata is extracted from the KB via the provided OAI-PMH protocol.

Note: the data splits are merely to reduce repetition in the data (and therefore stored size), and improve clarity of the data for inspection.

2.2 How many instances are there in total (of each type, if appropriate)?

The dataset contains 2,720 extracts, 2,395 of which have 7 or more separate annotations against them. Appendix E contains a breakdown of extract count by target word and decade.

More specifically, there are 21800 annotations spread over 2,720 newspaper article extracts. These were gathered in 2 separate studies. A HuC study consisting of 1000 annotations over 150 samples with an unknown number of separate participants. A study sourcing 417 participants via the Prolific platform of 20800 annotators over 2570 samples. The data from both of these complementary studies is presented together as a combined set. The entries are discernable, between both studies: with 'extract_id' values in Annotations.csv and Extracts.csv wrt., study 1 (described below), being preceded by a 'H'. E.g., extract H147.

Section 3.1 provides a description of the 2 complementary studies conducted.

2.3 Does the dataset contain all possible instances or is it a sample(not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable)

The extracts represent a sample only. The available sampling pool is restricted as follows:

- Of the entire Europeana Newspaper collection, extracts were limited to issues returned from a Europeana search[7] query of, *query="National Library of the Netherlands"&theme=newspaper*
- The newspaper article pool was further refined to that subset of articles defined as *type=artikel* via a KB JSRU[8] search query.
- The sample pool was limited to the 6 decades between 1890-01-01 and 1941-12-31

[5] https://www.kb.nl/en
[6] https://www.prolific.co
[7] https://pro.europeana.eu/page/search
[8] https://www.kb.nl/sites/default/files/docs/techniek-anp_0.pdf

- The sample pool was limited to the target words as defined in Appendix A. Specifically, as noted in appendix A, the unigram terms only.

Appendix B shows the available article pool according to the above criteria.

The dataset represents a stratified sample set over target word, decade, and newspaper issue distribution metadata, with respect to the available sample pool of Appendix B.

2.4 What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description

See 2.1

2.5 Is there a label or target associated with each instance? If so, please provide a description

As per 2.1, the extract label is one of four multiple-choice options. An example of extract and multiple-choice options is given in Appendix B

2.6 Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g.,because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text

The newspaper text extracts in the available dataset (Extracts.csv, column 'text') were shortened to max 140 characters including whitespace due to copyright.

2.7 Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit

Extracts and their corresponding annotations, in Extracts.csv and Annotations.csv are related by the 'extract_id' column values in both subsets.

The anonymised participants referenced in Annotations.csv can be related to the participant demographic data in Demographic.csv by the 'anonymised_participant_id' column values present in both subsets.

The extracts and corresponding metadata (of the newspaper issues from which the article extracts are drawn), can be related by the 'url' column values. I.e., the url to the OCR version of the article made available by the National Library of the Netherlands.

2.8 Are there recommended data splits (e.g., training, development/validation,testing)? If so, please provide a description of these splits, explaining the rationale behind them

No.

2.9 Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description

Fundamentally, the dataset is attempting to capture a consensus as to whether a target word is contentious in a given textual context. There are several potential sources of noise:

- Consensus is a hard thing to capture consistently with an eye to en-masse ML applications: i.e., each extract set is annotated by different groups of people. Therefore the annotation consensus for that set reflects very specifically the opinions of that group of annotators, which may differ significantly between annotator groups.

- Extracts were sampled by sampling multiple extracts stratified by target word, decade and newspaper distribution, such to capture variation. However, we are dependent on the output of this random stratified sampling wrt., the range and proportion of varying textual contexts captured in the dataset.

- Annotations are based on unaltered OCR text. The OCR quality can affect participant understanding of an extract;

- The participants from Study 2 are paid participants via the Prolific crowdsourcing platform - consequently their level of engagement with the task cannot be

guaranteed. Each participant in study 2, was asked to annotate the same 5 control questions (labelled 'C0-C4' in Extracts.csv and Annotations.csv) as a tool to assess participant quality. Demographic data also includes annotation duration to help screen data quality.

2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources:

A. are there guarantees that they will exist, and remain constant, over time;
B. are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created);
C. are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user?

Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data source does not rely on external sources. References to the source text and metadata from KB and the Europeana article item identifier are given to provide users with provenance information, but the resource can be used on its own.

2.11 Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description

No.

2.12 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The dataset is concerned with possible contentious words in context, based in-part on the list of 'sensitive' words from the Words Matter publication. Such a word list inherently contains words and contexts that a reader may find offensive or insulting.

2.13 Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. Annotator responses are captured in regards to their perception of potential contentious words in context, with anonymised demographic data made available for those participants consulted via the Prolific platform.

2.14 Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Subpopulations by age, gender, nationality, country of birth, country of residence and employment status are visible.

Refer to Demographics.csv for a breakdown of these subpopulations.

2.15 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No. All participant IDs referenced Annotations.csv and Demographics.csv are anonymised.

Whilst there is demographic data available (wrt., Study 2 participants via the prolific platform) against these anonymised IDs, it is extremely broad such that individual persons are not directly identifiable.

2.16 Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexualorientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such

as social security numbers; criminal history)? If so, please provide a description.
No sensitive data was collected from the annotators. The initial dataset is sensitive by its nature. It contains offensive and discriminative words and phrases. The annotation participants were warned about this.
2.17 Any other comments?
No

Collection Process

3.1 How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The dataset was assembled via 2 complementary studies using human annotators. In each study, sets of extracts were assembled according to specified ratios of *contentious*, *alternative* and *additional* words (Appendix A). Each set was assigned to multiple annotators, published via Google Forms[9] with responses subsequently collated over the multiple forms via Google Apps Script routines. <u>Study 1</u> The study consisted of 1,000 annotations, divided over 150 unique extracts, annotated by anonymous colleagues (volunteers) situated in the KNAW Humanities Cluster.[10] Each participant annotated (at least) one of 3 sets of 50 unique extracts as a Google form. In each set, there is a 50:50 split of extracts containing a *contentious* or *alternative* word of Appendix A. Two of three sets were annotated by 7 experts, and one by 6. The number of unique participants is unknown because the annotation process was anonymous, and an expert could annotate more than one (different) set. No control questions were used in this study. <u>Study 2</u> The study consisted of 20800 annotations, divided over 2,570 unique extracts crowdsourced from 416 participants via the Prolific[11] platform. Each participant annotated one of 57 sets of 50 extracts as a Google form. In each set, extracts containing *contentious*, *alternative* and *additional* words as defined in Appendix A, were randomly assembled from a sample pool in the amounts 20:20:5 respectively. The remaining 5 extracts for annotation in each form were control questions (labelled with extract_id c0-c1 in Extracts.csv and Annotations.csv). Except for the control questions, each extract was unique.
3.2 What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
Extract sets were assembled into separate Google Forms forms with responses subsequently collated over the multiple forms via Google Apps Script routines. In both studies, in checking as to whether the collated responses accurately capture the raw form response data, we checked by comparing random samples of particular forms against the collations for those forms. This is deemed sufficient as the same code was applied to each form iteratively with no variation between forms. For the Prolific participants in Study 2, participant id demographic data in Demographics.csv is linked to the annotations in Annotations.csv. A simple external web app was used to link participants transferred from the prolific platform to a particular form. In doing so, recording their unique prolific ID (anonymised in the dataset) against the extracts set they were directed to. Hence, we can be certain which participant is attributed to which set of

[9] https://www.google.com/forms/about/
[10] https://huc.knaw.nl/
[11] https://www.prolific.co/

annotations.

3.3 If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

1. An initial potential article pool from which to draw contentious extracts was defined as per 2.3. The available article counts by target word and decade are shown in Appendix B;

2. This pool was then stratified sampled as follows:

   200 articles, each containing a target word in the *Contentious words*, *Alternative words* and *Additional words* lists of Appendix A, across each decade 1890 upto and including 1941 was sampled. Ie., since there are 6 decades, 200x6 article samples are drawn for each target word.

   For each of these sampled articles, metadata was retrieved via KB JSRU API and OCR text version of each article was retrieved from http://resolver.kb.nl. This resulted in the following:

   - Approx 57K articles drawn from the *Contentious words* list;
   - Approx 12K articles drawn from the *Alternative words* list;
   - Approx 22K articles drawn from the *Additional words* list.

3. Potential extracts were identified via a "target word+decade+spatial" distribution combination. I.e., each sampled article had one of six spatial distributions associated with it: 'Landelijk', 'Nederlands-Indie / Indonesie', 'Verenigde Staten', 'Nederlandse Antillen', 'Suriname', 'Regionaal/lokaal'.'

   I.e., for each target word+decade+spatial distribution combination, 5-sentence extracts centred around a sentence containing the target word were identified.

4. Extracts were then sampled (separately) for Studies 1 and 2, from the extract set in step 3, as follows:

   <u>Study 2</u>

   See Appendix A for the *Contentious*, *Alternative* and *Additional* target words used in Study 2.

   Study 2 extracts were sampled to return *Contentious words, Alternative words, Additional words* and control extracts in the ratio 20:20:5:5, respectively.

   E.g., in the *Contentious words* list; each target word/decade combination was sampled twice with extracts weighted by P(extract)/Σ P(extract) for the matching word/decade extracts. These samples were then shuffled, and then cyclically going through the target words, a total of 1,200 extracts were taken. Estimates P(extract) were obtained from bigram frequencies in the entire initial OCR corpus (Step 2).

   This was similarly done for the Alternative words list, where n = 12, again retrieving 1,200 extracts; and for the Additional words list, were n = 1, and a total of 300 extracts were retrieved.

   Hence, the *Alternative words* are sampled at 6 times the rate of *Contentious words.*

3.4 Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

- Study 1 (of 150 extracts) involved anonymous volunteers from KNAW Humanities Cluster.

  Note: the participants in this study and their contributions are anonymous. To distinguish sets of completed annotations (which are necessarily for the same annotator) such sets are assigned an arbitrary participant id. However, it may be that, for example, "Unknown_3g", "Unknown_2a" are actually the same

annotator. I.e., that participants completed multiple annotation sets.

- Study 2 (2,570 extracts) involved paid participants sourced via the Prolific platform. These participants were paid on average £12.16/ hour.

3.5 Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.

Study 1 was commenced on 19th May, and left open for 3 weeks in total.

Study 2 was conducted in 3 rounds, each of 200 extracts using the Prolific platform:

- Round 1 conducted on the 9th June 2021
- Round 2 conducted on the 14th and 15th June 2021
- Round 3 conducted on 2nd to 6th August

3.6 Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

3.7 Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Yes, as per 2.14.

3.8 Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

All responses were obtained via Google Forms forms, hosted on team member Google Drive Accounts.

In all cases, participants were anonymously assigned a Google form to complete.

3.9 Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other informa- tion) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Appendices C and D contain the annotation instructions to participants and an example extract, respectively.

3.10 Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The study 1 participants were anonymous with no demographic data collected.

The study 2 participants do have anonymous demographic data as provided by the Prolific platform according to their GDPR policy.[12]

With regards to consent to incorporate their responses into the dataset, consent was deemed implicit for any participants of Study 1 or 2, who successfully completed and submitted their forms. With regards Study 2 on the Prolific platform where participants 'Returned'[13] their form, permission was considered revoked and not included in the final dataset.

3.11 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

No. No such mechanism was implemented.

3.12 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data

[12] https://researcher-help.prolific.co/hc/en-gb/articles/360009094594-Data-Protection-and-Privacy

[13] https://researcher-help.prolific.co/hc/en-gb/articles/360009259434-Returned-submission-status

protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Preprocessing/ cleaning/ labelling

4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
The extracts are taken directly from OCR text versions of the Europeana Collection newspaper issue articles. E.g, http://resolver.kb.nl/resolve?urn=ddd:010990094:mpeg21:a0074:ocr No further preprocessing was performed on the OCR text samples, except for converting target words to a unicode bold face and italicised version to highlight the target words to the participants. E.g., see the target word 'Aziatisch' in the Google Forms extract example in Appendix D.
4.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data
No
4.3 Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
No

4.4 Any other comments?
No

Uses

5.1 Has the dataset been used for any tasks already? If so, please provide a description.
Not at this time.
5.2 Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
Not at this time.
5.3 What (other) tasks could the dataset be used for?
The dataset was conceived as a means to train a machine learning algorithm to detect contentious language usage. As each term is sampled and annotated in multiple contexts, the dataset can also be used to investigate linguistic markers that may signal contentiousness. The data stems from historical newspapers and the seed terms were largely collected for use in the cultural sector.
5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
The Words Matter report states that it is an 'unfinished guide to word choice in the cultural sector', as such, the potentially contentious terms annotated in ConConCor is not an exhaustive or complete list. Furthermore, many of the seed terms stem from the colonial

domain, other types of biases or stereotypes are less well represented in this list. Another consideration is the fact that only contexts from newspapers are used, thus results and resulting analyses may not transfer to other text genres.
5.5 Are there tasks for which the dataset should not be used? If so, please provide a description.
We recommend that applications built on top of this dataset not be used autonomously but with a human-in-the-loop due to the sensitive and sometimes offensive nature of the material.
5.6 Any other comments?
No.

Distribution

6.1 Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
The dataset is publicly available to the research community.
6.2 How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is publicly available on the Cultural AI Lab GitHub[14]. The dataset does not have DOI at the moment.
6.3 When will the dataset be distributed?
30 June 2021 (v1) & 31 July 2021 (v2)
6.4 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
CC-By-SA? Discuss with Europeana & KB

[14] https://github.com/cultural-ai/ConConCor

6.5 Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
The data contains excerpts from the KB National Library of the Netherlands newspaper corpus that is also made available via Europeana. Content published more than 140 years ago is free of copyrights. Further information regarding data access rights is available via https://www.delpher.nl/ and http://www.europeana-newspapers.eu/
6.6 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
The dataset is freely available for research purposes.
6.7 Any other comments?
Not applicable.

Maintenance

7.1 Who is supporting/hosting/maintaining the dataset?
The dataset is publicly available on the Cultural AI Lab GitHub.
7.2 How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The researchers can be contacted via emails provided in the contact information in the header. It is also possible to communicate via Issues on the Cultural AI Lab GitHub.
7.3 Is there an erratum? If so, please provide a

link or other access point.
Not at the moment.
7.4 Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
The dataset may be updated or corrected by the Cultural AI Lab researchers, but not regularly. The changes can be seen on GitHub.
7.5 If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
Not applicable.
7.6 Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to user.
The versioning of the dataset is mainteded on GitHub.
7.7 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Check with Europeana
7.8 Any other comments?
No.

Appendix A: target words

The following *Contentious*, *Alternative* and *Additional* words were used as query words to query extracts from the Europeana Newspaper collection. We match them also to capture compound words. E.g., In querying extracts for 'knecht', we capture both extracts with 'knecht' as a target word, but also an extract with 'dienstknecht' as a compound target word. It is these target words (compound or otherwise) that participants were asked to annotate in context.

The following ***Contentious words***, ***Alternative words*** and ***Additional*** words lists, list both unigrams and longer n-gram phrases. We considered only the unigrams when sampling in both study 1 and study 2, with the expectation to consider longer ngrams in future work.

Study 1
In study 1, a random selection of target words: 50% drawn from the *Contentious words* list and 50% drawn from the *Alternative words* list was used as the basis of sampling extracts.

Those words in the extracts, extra to the target word, noted by participants as contentious

Study 2
In study 2, target words were drawn from the *Contentious words*, *Alternative words* and *Additional words* lists, below, in the proportions 40%:40%:10%. With the final 10% consisting of control questions. A decision was made to ignore the bold face words from the contentious and alternative word lists. In the case of 'trans', 'gay' and 'roma', they were discarded due to a extremely low return rate of the intended sense (0/100 samples for each). In the case of 'gay' and 'trans', this is perhaps to be expected since the Europeana Newspaper collection only extends to 1941.

Contentious words

The contentious word list, below, is taken from the Words Matter publication: i.e., a list of culturally *sensitive* words in particular use cases. Hence, this list represents words with a reasonable expectation

aboriginal, inuit, metis, afkomst, allochtoon, barbaar, barbaren, barbarlje, bediende, knecht, page, baboe, berber, blank, "mensen van kleur", "niet bevlekt", bosneger, boslandcreool, "derde wereld", "tweede wereld", "ontwikkelingsland", dwerg, lilliputter, eskimo, etniciteit, exotisch, gekleurd, "iemand van kleur", halfbloed, volbloed, mesties, handicap, hermafrodiet, homo, hottentot, inboorling, indiaan, indisch, indo, inheems, inlander, islamiet, jappenkampen, kaffer, kaukasisch, koelie, koppensneller, marron, medicijnman, mohemmedaan, mongool, moor, mulat, neger, ontdekken, orientaals, "politionele actie", primitief, pygmee, queer, ras, roots, slaaf, stam, traditioneel, westers, zigeuner, zwartje, **birma, batavia, zuid-rhodesia, gay, trans**

Alternative words

The Alternative word list, below, consists of those proposed by the Words Matter publication as more appropriate alternatives to specific *sensitive* words (context dependent).

"art nouveau", inheems, marron, "kleine mensen", dwerggroei, "van gemengde afkomst", "met een beperking", lesbisch, indo-europeaan, moslim, kwinti, traditionele, "tot slaaf gemaakt", achtergrond, gemengd, aziatisch, historisch, **roma, jakarta, zimbabwe**

Additional words

Additional potentially contentious words, informed by the participant suggestions of Study 1.

slaven, mengras, kokkie, zigeunermuziek, tovenaar, negerslaven, negercultuur, inlandsche, slavenschip, boesman, wild, onbeschaafd, aapman, misvormden, muzelman, afgod, negerrijk, "verre oosten"

Appendix B: Article pool available by target word and decade

The available pool as per the sampling restrictions described in 2.3.

Key

	Contentious target words
	Alternative target words
	Words included in both *contentious* and *alternative* word lists (depending on context)
	Additional target words

target word/ phrase	1890s	1900s	1910s	1920	1930s	1940s
aapman	0	2	1	6	290	56
aboriginal	4	2	4	16	6	0
achtergrond	9430	12532	14671	31904	54475	8336
afgod	1245	1176	1304	2415	2955	302
afkomst	6176	7278	8298	12524	15871	2624
allochtoon	0	0	1	0	0	0
aziatisch	949	1327	2088	2243	3486	464
baboe	1410	2068	1698	2459	3645	252
barbaar	470	598	686	765	717	126
barbaren	1131	1762	2419	2170	2731	538
barbarlje	0	0	3	0	2	1
bediende	15861	19127	16599	25085	27229	3640
berber	638	432	710	801	805	112
blank	5739	7679	9883	16938	18761	2684
boesman	55	79	39	594	282	54
boslandcreool	0	0	0	0	0	0
bosneger	0	3	0	12	31	1
dwerg	650	810	883	1905	2687	336
dwerggroei	0	6	12	22	31	8
eskimo	436	1094	886	2593	3464	401
etniciteit	0	1	1	5	2	0
exotisch	54	145	284	860	1265	134
gekleurd	5115	6768	6624	11893	14536	1793
gemengd	37197	55033	59230	97304	80513	12697
halfbloed	196	207	252	710	1003	165
handicap	2232	4376	3620	9680	18175	1943
hermafrodiet	0	1	0	0	0	0
historisch	7307	12650	16726	35825	43936	5976
homo	1219	1400	4524	3480	2371	264
hottentot	105	161	92	743	623	54
inboorling	1171	1674	1168	1856	2202	174
indiaan	1611	3635	1413	2373	2836	333
indisch	20442	22641	20052	35720	49807	5422
indo	4901	7056	11374	14613	27968	7516
indo-europeaan	177	154	146	467	1187	46
inheems	16	36	31	121	345	60
inlander	17080	20240	19458	21847	20544	1054
inlandsche	36452	45243	40184	67447	79533	3942
inuit	37	30	55	73	50	6
islamiet	55	59	106	256	324	23
jappenkampen	0	0	0	0	0	0
kaffer	1080	2232	612	966	955	92
kaukasisch	39	50	419	91	92	17
knecht	14391	18461	17662	25036	25445	3364
koelie	3994	7230	7376	10718	10007	419
kokkie	204	233	228	395	726	72
koppensneller	29	78	62	157	160	12
kwinti	0	0	2	5	4	0
lesbisch	0	1	0	6	3	0

lilliputter	42	34	80	125	259	68
marron	93	137	198	257	194	9
medicijnman	22	50	61	231	632	35
mengras	0	3	7	7	23	5
mesties	69	54	86	95	64	3
metis	58	90	172	87	49	3
misvormden	76	114	115	420	230	25
mohemmedaan	0	0	0	0	1	0
mongool	181	233	145	175	333	19
moor	8893	8502	60405	22766	10674	1517
moslim	124	303	294	490	516	77
mulat	356	295	865	687	551	46
muzelman	457	470	429	491	410	31
neger	3623	6342	5618	10455	12957	1166
negercultuur	0	0	3	27	54	2
negerrijk	31	3	1	11	18	3
negerslaven	122	112	108	241	458	50
onbeschaafd	682	810	673	1023	916	110
ontdekken	12418	16197	16660	32355	43932	6503
orientaals	0	1	0	3	14	0
page	2195	2287	2722	4410	4350	921
primitief	2090	2942	3914	7412	10159	1116
pygmee	10	21	14	17	82	15
queer	28	93	38	76	167	11
ras	20768	31922	32901	45171	59899	6648
roots	35	109	264	265	123	26
slaaf	2677	3670	3613	5360	5887	839
slaven	5463	6589	9054	11111	11298	1488
slavenschip	51	14	7	25	222	23
stam	12664	18698	17179	32187	40392	6223
tovenaar	11	13	96	90	323	190
traditioneel	526	812	1396	3237	4635	573
traditionele	32	39	54	203	1085	351
volbloed	2515	2848	3026	4426	5476	585
westers	406	1465	3092	2552	3873	401
wild	16237	23196	28468	39571	39632	5835
zigeuner	656	881	913	1864	4627	288
zigeunermuziek	20	29	24	103	2493	230
zwartje	105	174	267	475	500	78
"art nouveau"	9	44	7	26	15	0
"derde wereld"	2	6	10	44	38	3
"iemand van kleur"	0	0	0	0	0	0
"kleine mensen"	0	3	4	15	60	22
"mensen van kleur"	0	0	0	0	0	0
"met een beperking"	14	37	51	142	279	23
"niet bevlekt"	12	31	25	22	20	4
"ontwikkelingsland"	0	0	0	0	0	0
"politionele actie"	0	0	0	0	1	0
"tot slaaf gemaakt"	13	19	18	12	44	5
"tweede wereld"	8	8	10	69	112	13
"van gemengde afkomst"	2	3	1	6	6	0
"verre oosten"	2400	7440	4297	14521	48750	7947

Appendix C: Annotation task instruction

Dutch (original)

In deze annotatietaak vragen we u om te beoordelen of u de ***vetgedrukte term*** term omstreden vindt (contentious in het Engels) in de context van de tekst die eromheen staat. Voor het doel van deze taak beschouwen we gebruik een term als omstreden wanneer het volgens u mogelijk beledigend, denigrerend, kwetsend, of om een andere reden niet passend is naar huidige maatstaven. Om uw oordeel te vellen kunt u zich bijvoorbeeld afvragen of u de term in een soortgelijke zin zou gebruiken en/of u verrast zou zijn als de term op deze manier gebruikt zou worden door anderen. Sommige voorbeelden zijn duidelijk omstreden; voor andere is het moeilijker te beoordelen. Als u een term in de gegeven zin een klein beetje omstreden vindt, vragen we u om het hokje ‘omstreden’ aan te vinken. Als u een bepaald voorbeeld niet kunt beoordelen, kunt u het hokje ‘weet ik niet’ gebruiken.

Een zin kan meerdere omstreden termen bevatten. Beoordeel alleen de ***vetgedrukte term***. Andere (volgens u) omstreden termen kunt u kopiëren naar de kolom 'Andere omstreden termen in de context'.

De zinnen die u te zien krijgt komen uit het krantencorpus van de KB Nationale Bibliotheek. Deze kranten zijn gescand en automatisch computerleesbaar gemaakt via 'Optical Character Recognition' (OCR). Het kan voorkomen dat de originele bron te beschadigd was om kwalitatief goede OCR resultaten te verkrijgen, waardoor de tekst die u ziet vreemde karakters kan bevatten of zelfs onleesbaar kan zijn. Als er een paar kleine OCR foutjes in de tekst staan die uw evaluatie niet hinderen kunt u deze negeren, maar als u de term niet kunt beoordelen vanwege de slechte OCR, vinkt u dan het hokje 'Onleesbare OCR' aan.

In de laatste kolom van het formulier getiteld 'Notities' kunt u uitleg of andere suggesties over het voorbeeld aangeven. Dit is optioneel.

Aan het eind van het annotatieproces zouden we u graag willen vragen wat u van deze taak vond. Hartelijk dank voor uw medewerking en voor het nemen van de tijd om ons te helpen bij het meer cultuurbewust maken van AI systemen!

English (translated)

In this task, we would like to ask you to judge whether you find the ***bold-faced term*** contentious in the given context. For the purpose of this task, we consider 'contentious' terms that are insensitive, hurtful or in other ways inappropriate according to your present-day sensibilities. The question you can ask yourself to make a decision is whether you would use the highlighted term in polite conversation or similarly would be surprised or irritated to hear it used by others. Some examples are clearly contentious; and other examples can be more difficult to decide. If you find a term in the given sentence even a little bit contentious, we ask you to choose the 'contentious' option. If you cannot decide on a given example, please answer 'I don't know'.

A sentence may contain multiple contentious terms. Please provide your judgement only for the ***bold-faced term***. If you do notice any other terms that you judge to be contentious, please copy them into the column 'Other contentious words in the context'.

The sentences you see are derived from the newspaper archive of the KB National Library of the Netherlands. The newspapers were scanned and automatically processed using Optical Character Recognition (OCR). It can happen that the original source was too damaged for the OCR to work well, in which case the text you see may be garbled and illegible. Please try to ignore the minor OCR errors but if you absolutely cannot read the text, please choose 'Bad OCR' to indicate that you could not interpret the given example.

Finally, we added an optional column 'Notes' where you can share any thoughts on the specific term or its context.

At the end of the questionnaire, we would like to additionally ask you for feedback about this questionnaire. Thank you very much for taking the time to help us in our work towards culturally-aware AI!

Appendix D: Annotation task example extract as presented in a Google Forms form

Het heette toen en de rijk met provisie bedachte bankiers van West-Europa bazuinden 't uit in hunre prospectussen, dat Rusland een onmetelijk en rijk land was.

Wat is er al niet gefabuleerd over óe beroemde Russische zwarte aarde, die nooit gemest behoefde te worden. j Onmetelijk is Rusland ongetwijfeld, met zijn!

Aziatisch gebied is 't ± 700 X de oppervlakte j van ons eigen land.

Toch draagt de bodem niet meer dan 20 X zooveel menschen als de Nederlandsche.

De bevolking toch v?.n Rus-; land zal heden niet zooveel meer dan 130 èj 135 millioen zijn. *

- ◯ Omstreden naar huidige maatstaven
- ◯ Niet omstreden
- ◯ Weet ik niet
- ◯ Onleesbare OCR

Andere omstreden termen in de context (boven)

Jouw antwoord

Appendix E: Dataset extract count by target word and decade

Key

	Contentious target words
	Alternative target words
	Words included in both *contentious* and *alternative* word lists (depending on context)
	Additional target words

target	1890s	1900s	1910s	1920s	1930s	1940s	total
aapman	0	0	1	2	2	1	6
aboriginal	2	1	3	4	3	0	13
achtergrond	24	24	25	23	24	21	141
afgod	3	4	4	4	4	4	23
afkomst	3	3	4	4	4	5	23
allochtoon	0	0	1	0	0	0	1
aziatisch	25	24	25	23	23	21	141
baboe	3	5	4	3	5	4	24
barbaar	4	3	4	4	4	4	23
barbaren	3	4	4	5	4	4	24
barbarlje	0	0	2	0	0	1	3
batavia	1	1	0	0	0	0	2
bediende	4	3	5	4	4	4	24
berber	5	3	6	3	3	3	23
birma	0	0	0	0	1	0	1
blank	4	5	4	6	4	3	26
boesman	3	2	2	3	3	2	15
bosneger	0	3	0	3	4	1	11
dwerg	4	5	6	3	5	4	27
dwerggroei	0	5	11	15	14	5	50
eskimo	6	3	4	5	4	4	26
etniciteit	0	0	0	1	2	0	3
exotisch	4	4	4	4	4	4	24
gay	0	0	0	1	1	0	2
gekleurd	4	4	4	4	4	3	23
gemengd	24	23	20	22	27	23	139
halfbloed	4	4	3	4	3	4	22
handicap	3	3	4	5	4	4	23
historisch	23	25	22	23	24	21	138
homo	4	3	4	3	4	4	22
hottentot	4	4	5	3	4	4	24
inboorling	3	3	3	4	5	3	21
indiaan	3	3	4	4	4	5	23
indisch	4	5	5	3	3	4	24
indo	4	5	4	5	3	4	25
indo-europeaan	16	15	25	22	18	3	99
inheems	14	26	16	25	28	27	136
inlander	4	4	4	3	4	3	22
inlandsche	4	4	4	3	4	2	21
inuit	4	3	3	3	4	1	18
islamiet	3	3	4	4	4	1	19
jakarta	1	2	1	1	0	0	5
kaffer	5	4	4	7	3	3	26
kaukasisch	4	5	3	3	3	3	21
knecht	5	5	6	3	4	3	26
koelie	3	4	3	4	4	4	22
kokkie	2	3	4	4	3	2	18
koppensneller	3	4	4	4	4	3	22
kwinti	0	0	0	2	3	0	5
lesbisch	0	1	0	10	0	0	11
lilliputter	4	3	5	3	4	4	23
marron	29	28	27	28	28	9	149
medicijnman	2	4	4	5	4	4	23
mengras	0	2	2	2	2	2	10

mesties	5	5	4	4	3	1	22
metis	4	6	3	4	4	0	21
misvormden	4	3	3	3	4	2	19
mohemmed aan	0	0	0	0	1	0	1
mongool	4	4	4	4	4	4	24
moor	3	4	3	4	4	4	22
moslim	22	24	24	25	24	16	135
mulat	4	4	3	4	4	4	23
muzelman	4	4	4	4	4	2	22
neger	5	4	3	4	7	5	28
negercultuur	0	0	1	3	2	1	7
negerrijk	4	1	0	1	3	1	10
negerslaven	4	4	4	4	4	3	23
onbeschaafd	4	4	4	4	3	3	22
ontdekken	4	4	4	3	4	3	22
page	4	4	4	4	4	3	23
primitief	3	3	5	3	4	4	22
pygmee	2	3	6	3	4	3	21
queer	3	4	4	3	4	4	22
ras	4	3	4	4	3	4	22
roma	0	1	2	0	0	0	3
roots	4	3	5	4	3	3	22
slaaf	5	4	5	4	4	4	26
slaven	4	4	3	3	4	3	21
slavenschip	4	4	2	2	3	2	17
stam	4	4	3	4	4	4	23
tovenaar	3	2	2	2	4	2	15
traditioneel	4	4	3	3	5	3	22
traditionele	7	11	18	21	24	24	105
trans	0	0	0	1	0	1	2
volbloed	4	5	4	4	3	4	24
westers	4	4	4	3	3	4	22
wild	4	4	4	4	4	4	24
zigeuner	7	4	4	4	5	3	27
zigeunermuziek	2	2	2	2	2	2	12
zimbabwe	0	0	0	3	1	0	4
zwartje	5	3	4	4	4	4	24