Constructing ConConCor:
The Contentious Terms in Contexts Corpus

Project documentation version 0.9

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

Abstract: This document describes the process of creating the Contentious terms in Context Corpus (ConConCor). This project was carried out in the context of the EuropeanaTech Challenge for Europeana Artificial Intelligence and Machine Learning datasets. Due to the nature of the project, some examples used in this documentation may be shocking or offensive. They are provided only as an illustration or explanation of the resulting dataset and do not reflect the opinions of the project team or their organisations.

1. Introduction

Cultural heritage institutions recognise the problem of language use in their collections. The cultural objects in archives, libraries, and museums contain words and phrases that are inappropriate in modern society but were used broadly back in times. Such words can be offensive and discriminative. In our work, we use the term 'contentious' to refer to all (potentially) inappropriate or otherwise sensitive words. For example, words suggestive of some (implicit or explicit) bias towards or against something.

The National Archives of the Netherlands stated that they "explore the possibility of explaining language that was acceptable and common in the past and providing it with contemporary alternatives", meanwhile "keeping the original descriptions [with contentious words], because they give an idea of the time in which they were made or

included in the collection".[1] There is a page on the institution website where people can report "offensive language". The Amsterdam Museum published a statement in 2019 that the term "Golden Age" would not be used anymore in their exhibitions to refer to the Dutch 17th century. The museum sees giving up the term as "a step to enable other perspectives on that time."[2] In 2015, Rijksmuseum Amsterdam instituted a Terminology working group that critically assesses descriptions that use terms that were acceptable and in common usage decades ago but are now considered biased and contentious.[3] The Dutch National Museum of World Cultures is compiling a list of word choices for the cultural sector in the work-in-progress publication "Words Matter".[4]

These initiatives illustrate that the problem of contentious terms usage in cultural heritage institutions goes beyond replacing such words in object descriptions. Firstly, the process of manually detecting inappropriate words in collections is not scalable (for example, the digital collection of KB National Library of the Netherlands consists of more than 120 million pages). Secondly, there is no unified method to replace those words with alternatives. It requires knowledge of experts from different fields: collection curators, historians, anthropologists, linguists. Moreover, the replacement process may itself become a controversial issue. Thirdly, even if there is a suitable neutral synonym for an inappropriate word, should it be replaced in the entire collection or only in some cases? The context with contentious words may differ: in one case, it could refer to a group of people and carry offensive semantics, while in another, it is used in a descriptive context and carry additional historical meaning when the word used to represent a norm (such words can be 'historical objects' themselves).

The question “How to build AI systems that are ‘aware’ of the different cultural and speech contexts to tackle the problems of word usage in heritage collections?” arises. This question shaped our idea of collecting an annotated corpus of contentious contexts: to investigate the dependencies of words’ contentiousness on contexts and perspectives with statistical methods.

To facilitate the transparency of the corpus creation and possible use we adopted the idea and the layout of a datasheet from the "Datasheets for Datasets" paper recommended by

---

[1] Het Nationaal Archief.Taalgebruik in onze archieven. https://www.nationaalarchief.nl/taalgebruik-in-onze-archieven Last visited: 24/06/21

[2] Amsterdam Museum gebruikt term ‘Gouden Eeuw’ niet meer. https://www.amsterdammuseum.nl/nieuws/gouden_eeuw Last visited: 30/06/21

[3] https://www.rijksmuseum.nl/en/research/our-research/overarching/terminology Last visited: 30/06/21

[4] https://www.materialculture.nl/en/publications/words-matter Last visited: 30/06/21

Europeana.[5] The ConConCor datasheet is available on the Cultural AI Lab GitHub[6] and describes the motivation, composition, collection process, preprocessing, uses, distribution, and maintenance. While the datasheet is more technical and detailed, this documentation gives an overview of the main parts of the project and summarises the annotation results.

2. Data Selection

In this section, we describe our data selection process. We detail the selection of seed terms, the selection of contexts from the KB National Library of the Netherlands newspaper collection, and the corpus structure.

2.1. Glossary of contentious words and their alternatives

The importance of context in understanding contentiousness is exemplified in the publication "Words Matter" by the National Museum of World Cultures. It was published in English and Dutch versions in 2018 as work-in-progress. The publication has collated a list of sensitive terms, describing their historical usage and implications together with the context in which they may be used appropriately or inappropriately. Besides, the authors provided suggestions for contentious words: alternative terms or examples of how to use the terms appropriately.

"Words Matter" serves as an input for our task of constructing the corpus. It is the knowledge of the museum professionals and other experts that helps us to make a step towards analysing contentiousness with statistical methods.

The Dutch contentious terms and their suggested alternatives were put in a dictionary to query the KB newspaper archives available on Europeana. The dictionary included 84 terms. During the sampling, we considered only unigrams. Compound terms, for example, "kleine mensen" ('small people') were not included in the study.

We queried the suggested words along with contentious to capture different contexts and include a proportion of 'negative examples' (the words that are not considered contentious in certain contexts) in our dataset.

2.2. The corpus structure

The queries were run on OCR'd versions of the Europeana Newspaper collection, as provided by the KB National Library of the Netherlands.

---

[5] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2020). Datasheets for Datasets. ArXiv:1803.09010 [Cs]. http://arxiv.org/abs/1803.09010

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

The motivation for sampling historical newspapers is that they provide views on daily lives and customs and they are both producers and messengers in the public debate. As such, they provide unique insights into societies of the past. Furthermore, newspapers provide us with extensive texts containing contextual information, enabling further study into the characteristics of contentious words and phrases.

We limited our pool to text categorised as 'article', thus excluding other types of texts such as advertisements and family notices. We then only focused our sample on the 6 decades between 1890-01-01 and 1941-12-31, as this is the period available in the Europeana newspaper corpus. The dataset represents a stratified sample set over target word, decade, and newspaper issue distribution metadata. For the final set of extracts for annotation, we gave extracts sampling weights proportional to their actual probabilities, as estimated from the initial set of extracts via trigram frequencies, rather than sampling uniformly.

Two reasons motivated this decision:

> (1) The very high rate of OCR errors in the corpus. Uniform sampling would have implied a too high chance of providing annotators with entirely unintelligible sets of examples. At the same time, almost all extracts contained some errors, so we needed some way of scoring;
> (2) Albeit small in comparison, the set of contexts selected for annotation should still be representative of the larger corpus (the newspaper archive) in terms of common linguistic variables and semantic content.

Weighting extracts by their probability addresses both of these issues and does so in an indirect way, i.e. does not require laborious engineering of features and does not introduce any obvious unwanted sampling biases. Intuitively, extracts that rank high in probability tend to be of moderate lengths and contain words that are by and large 'actual' words (and not e.g. the outcome of OCR errors) and that are commonly used and known.

As confirmed by manual inspection, the extracts we obtained by this sampling strategy indeed were of much higher linguistic quality than those obtained by uniform sampling and thus more likely to make it easier for annotators to make good judgements.

The resulting corpus consists of 2,720 newspaper article extracts. This dataset is independent from the annotations and can be used as an input for other crowdsourcing and machine learning tasks.

The dataset available for annotations included extracts (samples), 3 or 5-sentence fragments from a newspaper article with a bolded target word.

3. Annotating contentiousness of terms in context

Deciding on whether a word is contentious or not may be a very subjective choice. Even if there are expert guidelines on the use of contentious words, such as the "Words Matter" glossary, the context in which a term is used may affect its meaning and whether or not the term is contentious. The goal and the idea of annotating the corpus are not to categorise words as "contentious/not contentious", but to collect people's judgements on which words are contentious in which contexts. In this case, peoples' opinions can be seen as 'scores of contentiousness: if more or fewer people agree on word usage.

The annotation process included 3 stages: pilot annotation, expert annotation, and crowdsourced annotation on the "Prolific" platform. All stages required the participation of Dutch speakers.

The pilot stage was intended for testing the annotation layout, the instructions clarity, the number of sentences provided as context, the survey questions, and the difficulty of the task in general. The Dutch-speaking members of the Cultural AI Lab were asked to test the annotation process and give their feedback anonymously using Google Sheets. Six volunteers contributed to the pilot stage, each annotating the same 40 samples where either a context of 3 or 5 sentences surrounding the term were given. An individual annotation sheet had a table layout with 4 options to choose for every sample (*'Omstreden'* ('Contentious), *'Niet omstreden'* ('Not contentious'), *'Weet ik niet'*, ('I don't know'), and *'Onleesbare OCR', (Illegible OCR')*), 2 open fields (*'Andere omstreden termen in de context' (Other contentious terms in the context')* and *'Notities' ('Notes')*), and the instructions in the header. The rows were the samples with the highlighted words, the tick boxes for every option, and 2 empty cells for the open questions. An example annotation sheet is given in Appendix A. The obligatory part of the annotation was to select one of the 4 options for every sample. Finding other contentious terms in the given sample, leaving notes, and answering 4 additional open questions at the end of the task were optional. The open questions for the pilot study are provided in Appendix C.

Based on the received feedback and the answers to the open questions in the pilot study, the following decisions were made regarding the next, experts' annotation stage:

- The annotation layout was built in Google Forms as a questionnaire instead of the table layout in Google Sheets to make the data collection and analysis faster as the number of participants would increase;
- The context window of 5 sentences per sample was found optimal;
- The number of samples per annotator was increased to 50;
- The option 'Omstreden' ('Contentious') was changed to 'Omstreden naar huidige maatstaven' ('Contentious according to current standards') to clarify that annotators should judge contentiousness of the word's use in context from today's perspective;
- The annotation instruction was edited to clarify 2 points: (1) that annotators while judging contentiousness should take into account not only a bolded word but also the context surrounding it, and (2) if a word seems even slightly contentious to an

annotator, they should choose the option 'Omstreden naar huidige maatstaven' ('Contentious according to current standards');
- The non-required field for every sample 'Notities' ('Notes') was removed as there was an open question at the end of the annotation, where participants could leave their comments;
- Another open question was added at the end of the annotation asking how much time it took to complete the annotation.

The goals of the expert annotation stage were to collect opinions from humanities scholars in the fields of history, linguistics, ethnology and literature studies' on contentiousness, their suggestions regarding both other contentious terms in samples and the task in general, and to test the annotation process design.

The experts participated in the study voluntarily and anonymously. They received a non-personal invitation email and were asked to choose an available annotation form in a Google Sheet. The annotation form example is in Appendix B. After the completion, they would mark the selected form as 'completed'. No demographic information was collected from the participants. Although there was no monetary reward, we gave the volunteers a chance to win one of 10 Cultural AI mugs. If a participant wanted to win a mug, they could leave their email address at the end of the questionnaire. As 7 participants left their emails, we rewarded the mugs to them. The project budget was partially spent on this reward. We did not link the participants' emails to the annotations. The personal information was used only for sending out the rewards with the participants' consent. The results of the expert annotation phase are provided in Section 4.1.

The main expected result of the expert annotation phase was the number of chosen options ("Omstreden naar huidige maatstaven" ('Contentious according to current standards'), "Niet omstreden" ('Not contentious'), "Weet ik niet" ('I don't know'), "Onleesbare OCR" ('Illegible OCR')) per every sample. Besides that, we collected suggestions from annotators to indicate other contentious words in a sample (as we targeted only one word per sample). We used the suggested words by experts to expand our query glossary and include new samples for the crowdsourcing phase.

Before putting the annotation task on a crowdsourcing platform, additional changes were made to the instruction based on the experts' feedback. It was emphasised that annotators should be guided by their opinion when judging contentiousness. The instructions for crowdsource workers are provided in Appendix D.

The selection of samples was also changed in the crowdsourcing study. The historical toponyms (for example, "Jakarta" or "Bombay") were excluded from the study as they were difficult for experts and caused high disagreement. Furthermore, their acceptable modern alternatives can be linked easily through available historical gazetteers such as those provided by the World Historical Gazetteer project.[7] New samples were generated based on the suggestions of the experts. They indicated at least 17 other contentious

[7] http://whgazetteer.org/ Last visited: 30-06-2021

words. In addition, we used 5 control samples for crowdsource workers selected from the expert annotations that yielded a 100% agreement. The control samples serve as both an attention marker and comparative samples i.e. to check whether the crowdsourcing annotators agree with the unanimous decision of the experts. The additional samples are given in Appendix A of the Datasheet. Thus, one set of samples for crowdsourcing consisted of 50 samples, including 5 control samples and 5 samples with the additional terms. The resulting questionnaire included 50 samples as required questions, an open text field to indicate other contentious words in every sample, and 4 open non-required questions at the end (the same as in the pilot study among Cultural AI team members).

The Prolific platform was used to distribute the tasks to the crowdsource workers. We did not use any pre-screening questions but stated the requirement of Dutch proficiency only. The participants completed the task in the same layout in Google Forms as the experts. A web app was developed to refer the Prolific workers to the forms.

All the participants received compensation for their work based on the rate of £12.23 per hour. These compensations were the main expenses of the project budget. Although there was a possibility to return or reject submissions based on several factors, for example, low completion time (the estimated time of the task was 30 min), we decided to reward all the participants and keep all the responses in the dataset.

The crowdsourcing was not anonymous as "Prolific" assigns unique IDs to their users and collects demographic data. The participants' IDs were anonymised in the dataset. The demographic data is stored separately in the 'Demographics.csv' file and was not considered in the analysis of the results at this stage.

The results of the crowdsourcing are described below in Section 4.2. The technicalities of the annotation process are put in the datasheet (see "Collection Process").

4. Results

4.1. Expert annotations

The expert annotation step included 20 anonymous participants from the KNAW Humanities Cluster.[8] The KNAW Humanities cluster is an alliance of three humanities research institutes covering (social) history, language, and culture. The annotators were divided in 3 groups, each group annotating a unique batch of 50 samples. Thus, the first batch was annotated by a group of 6 experts, the second and the third by 7 experts. In total, there are 1,000 annotated samples divided over 150 unique samples.

[8] https://huc.knaw.nl/

Krippendorff's alpha[9] was calculated for every group to measure the annotators agreement. A higher alpha means a higher agreement between annotators in a group. In Table 1 below, the alpha is calculated only for 2 options *"Omstreden naar huidige maatstaven" ('Contentious according to current standards')* and *"Niet omstreden" ('Not contentious')*. The numbers are fairly low which is as expected given the subjectivity and complexity of the task.

Table 1. Krippendorff's alpha per group in the experts' annotation study (only for options *"Omstreden naar huidige maatstaven"* and *"Niet omstreden"*)

Group	Number of participants	α
1	6	0.456
2	7	0.65
3	7	0.502

Although Krippendorff's alpha indicates the agreement between annotators, the percentage agreement between annotators per sample is more relevant in our case. While the alpha in our study can be very low in general because annotators might have different views (or different 'scales') on which words are contentious, the percentage agreement shows which words in which contexts are contentious or not contentious according to how many people. The percentage agreement allows us to see both the samples with the majority votes (if the majority of annotators selected one option) and the samples on which annotators disagreed.

In Table 2, the number of samples is grouped by the percentage agreement between annotators and by the majority vote option. For example, all of the annotators (100%) in every group (6 annotators in the first group and 7 in the second and the third) agreed that a target word was contentious in 7 samples. Or in the second row: 83–86% (5 out of 6 or 6 out of 7 in absolute values) annotators agreed that target words were not contentious in 21 samples. In 11 cases, there was no majority for any of the options. This happened, for example, in the sample with the ID 'H3' ("marron") ('maroon'), with 2 annotators marking it as contentious, 2 as non-contentious, and another 2 could not make a decision and selected the option "I don't know".

[9] Krippendorff, K. (2018). Content Analysis: An Introduction to Its Methodology, fourth edition, chapter 12. SAGE, Los Angeles.

Table 2. Number of samples by percentage agreement and the majority votes between the expert annotators

Percentage agreement	Number of samples with majority votes per option				Number of samples without majority	Total number of samples
	Contentious	Not contentious	I don't know	Bad OCR		
100%	7	41	0	0	0	48
83–86%	12	21	0	0	0	33
67–71%	9	19	0	4	0	32
43–57%	9	9	5	3	7	33
29–33%	0	0	0	0	4	4
Total	37	90	5	7	11	150

The breakdown of the target words grouped by options and percentage agreement is presented in Appendix E. For example, such words as "eskimo" ('eskimo'), "neger" ('negro'), "kaffer" ('kaffir'), "slaaf" ('slave'), and "dwerg" ('dwarf') in particular samples have 100% agreement between experts who marked them as contentious, while "gemengd" ('mixed'), "achtergrond" ('background'), and "historisch"('historical') are non-contentious according to all the experts. We selected 5 samples with these perfect-agreement words as control samples for the crowdsourcing participants.

Some of the words were found both contentious and non-contentious in different samples, which illustrates our hypothesis about the influence of context on a word's contentiousness. One of such cases is the word "primitief" ('primitive'): in sample 'H60' 6 out of 7 experts agreed that this word was contentious, but in sample 'H94' also 6 out of 7 experts agreed on the option 'Not contentious'. If we look at the texts of these samples, in 'H60' 'primitief' refers to people (*"Het volkskarakter is primitief"*) ('The race's character is primitive'), while 'primitief' in 'H94' is used to characterise conditions (*"De verzorging was primitief en de reizigers hadden vooral gebrek aan water"*) ('The care was primitive and the travellers mostly needed water').

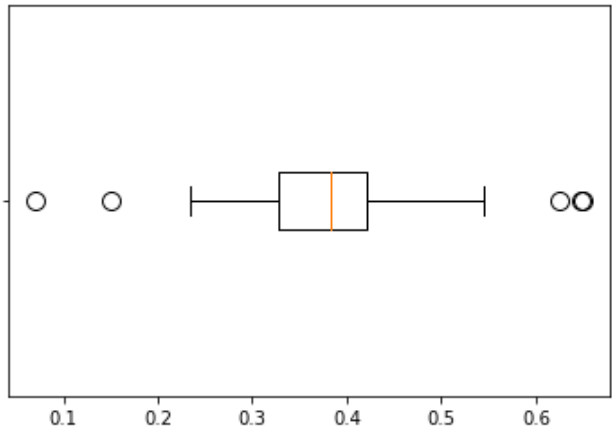
The most questionable word for experts was "marron" ('maroon'). In 3 samples the majority of annotators choose "I don't know", and some other samples with this word were annotated as non-contentious.

4.2. Crowdsourcing results

399 Annotators took part in the crowdsourcing study. They were divided into 57 groups, each annotating a unique batch of 50 samples. Together, the crowdworkers annotated 19,950 samples, and 2,570 of them are unique.

The annotators' agreement was measured using Krippendorff's alpha as in the expert's annotations. In Figure 1, the box plot shows the distribution of the alpha values across the crowdsourcing groups. There are 5 outliers. The highest agreement was reached in 3 groups with alpha values 0.65 (4 participants), 0.648 (6 participants), and 0.625 (7 participants). 2 groups have the lowest agreement with alpha 0.07 (6 participants) and 0.151 (7 participants). Most of the groups have alpha between 0.32 and 0.42, which is lower than the agreement in the expert groups.

Figure 1. The distribution of Krippendorff's alpha of crowdsourcing annotators.



The low agreement in the crowdsourcing groups highlights the problem of the subjective nature of the dataset and the task. However, a closer inspection of the (dis)agreement between annotators may provide information about the level of contentiousness of certain words in different contexts.

Table 3 below summarises the number of samples for which the majority of annotators selected one of the four options, as well as the number of samples for which no majority was found on any of the options (ties), grouped by percentage agreement. The table excludes 150 experts' annotations (as they were analysed separately), 5 control samples, and 45 samples with only 1 annotator. 45,7% of samples were annotated with the high percentage agreement (80–100%), 8,2% of the dataset constitute samples with no majority vote, and there are 4,9% of samples with OCR errors, because of which the annotators could not make a decision.

Table 3. Number of samples by percentage agreement and majority vote of the Prolific annotators

Percentage agreement	Number of samples with majority votes per option				Number of samples without majority	Total number of samples
	Contentious	Not contentious	I don't know	Bad OCR		
100%	52	519	1	4	0	576
80–92%	52	503	2	18	0	575
60–78%	126	486	48	55	0	715
50–58%	93	193	43	34	82	445
30–44%	19	28	23	13	125	208
Total	342	1729	117	124	207	2519

Appendix F lists top target words grouped by options and percentage agreement in crowdsourcing annotations. The top samples with perfect agreement on contentiousness contain the following words: "neger"('negro') (15 samples), "negerslaven" ('negro slaves') (12), "bosneger" (literally 'forest negro', term used to denote descendants of runaway slaves in Surinam) (5), "negerrijk"('negro kingdom') (5), "negercultuur" ('negro culture') (3), that share the same root, and "zwartje" ('blacky') (4). 100% agreement on non-contentiousness was reached in samples mentioning "achtergrond" ('background') (79), "historisch" ('historical') (79), "traditionele" ('traditional') (53), "aziatisch" ('asian') (49), and "gemengd" ('mixed') (41). This partially corresponds to the experts' annotations.

The control samples were annotated by every crowdsource worker (399 annotations). Table 4 shows percentage agreement and the distribution of options in absolute values in these samples. There is no perfect agreement in any of the 5 samples. The majority of the crowdsource workers agreed with the experts that the word in sample 'c1' was contentious, and words in samples 'c2' and 'c4' were not contentious. Samples 'c0' and 'c3' have low percentage agreement, so the opinion about contentiousness in these contexts were divided in crowdsourcing, while 100% experts marked 'c0' and 'c3' (samples 'H56' and 'H119' respectively in expert annotations) as contentious.

Table 4. Percentage agreement between crowdsourcing annotators in control samples

Sample ID	Target word	Percentage agreement	Contentious	Not contentious	I don't know	Bad OCR
c0	"kaffer"	41%	134	91	163	11
c1	"neger"	93%	371	21	6	1
c2	"achtergrond"	95%	13	378	7	1
c3	"dwerg"	55%	221	136	39	3
c4	"gemengd"	95%	6	380	11	2

Both experts and crowdsource workers found it difficult to make a decision about contentiousness with samples containing the word "marron" ('maroon') and "metis" ('metis'). Other difficult cases for crowdsourcing were samples with words "mesties" ('mestizo'), "baboe"(a nanny or a servant), "koelie"('coolie'), "muzelman"(archaic word for 'Muslim'). These samples have a majority agreement on the option "I don't know". The samples with low percentage agreement (23-33%) also include words "koppensneller" ('headhunting'), "hottentot" (refers to Khoekhoen Peoples), "dwerggroei"('dwarfism'), "inlander"('native'), and "indo-europeaan"('Indo-European').

Same as in experts' annotations, there are target words that appear to be contentious in one context and non-contentious in another. There are multiple cases in the dataset, here we give 2 examples:

> (1) The word "barbaren" ('barbarians') in sample '619' is contentious with 100% agreement (7 annotators). The same target word in sample '271' is non-contentious according to all 8 annotators. In the first case, the word is used to describe people as uncivilised ([619] "Als nu een rijke Chinees, dien het "noodlot" onder de barbaren (niet-Chineezen) gebracht heeft, naar een gemalin van zijn eigen stam verlangt <...>") ("If a rich Chinaman, who brought 'fate' to the barbarians (non-Chinamen) desires a consort of his own tribe <..>") . However, in the second sample the term is used in a more metaphorical sense, referring to the concept of an uncivilised enemy ([271] "Wij zullen ons wreken niet met de wapenen der barbaren, maar met het geestes zwaard onzer propaganda. Iedere man en Iedere vrouw, wien het ernst is met het streven der sociaal-democratie, moet tot die wraak bijdragen.") ("We shall take revenge, not with the barbarians' weapons, but with the spiritual sword of our propaganda. Every man and every woman, who is serious about striving for the social-democratic cause, must contribute to this revenge").
>
> (2) Another example shows the usage of words with the same root in contentious and non-contentious samples. The word "negerslaven" ('negro slaves') in the sample '2' is contentious with 86% agreement. The word "slavenhandel"('slave

trade') in sample '2379' is non-contentious with 83% agreement. While the first sample refers to people ("Deze vrouw bond den strijd aan tegen het zwartste onrecht, dat Amerika ooit gekend heeft, tegen de verdrukking der negerslaven.") (“This woman battled against the blackest injustice America has ever known, she battled against the oppression of negro slaves”), the second mentions "the trade of enslaved people" ("Wat zoude Washington in Lincoln's tijd van den slavenhandel gedacht hebben?")(“What would Washington have thought of the slave trade”).

In future work, we will analyse the contexts further to gain a better understanding of the nature of contentiousness. We expect this to aid machine learning algorithms trained to detect contentiousness as well as linguistics research.

Another aspect of the crowdsourcing process is participants' demographics. In this study, we did not set any demographic requirements except Dutch language fluency. The demographics data in our study shows that 63.9% of participants identified themselves as "male". And the majority of the participants were born in the Netherlands and Belgium. However, the diversity of participants in such subjective tasks as annotating contentiousness and historical colonial words, in particular, is significant, and it remains an open issue. How do we collect opinions of people with different backgrounds and levels of knowledge and enable more diverse perspectives? This question needs to be addressed when designing annotation studies in future work.

4.3. Project deliverables

The resulting dataset includes 4 files in CSV format:

- Annotations.csv contains anonymised participant IDs, extract IDs, responses, suggestions, and an indicator of whether an extract was used as a control sample;
- Demographics.csv is an export of demographic data of crowdsource workers from the Prolific crowdsourcing platform with anonymised IDs. This file also includes the time participants spent to complete their tasks;
- Extracts.csv has extract IDs, target words, the extracts' texts of 5 sentences, and links to XML files with the full text available via http://resolver.kb.nl;
- Metadata.csv has additional information about the sources of extracts such as publisher, date, place of distribution, language, and also links to the newspaper catalogue on Europeana.

Additionally, we provide Krippendorff's alpha scores, annotator IDs, and extract IDs per group in the file 'aplha_per_group.csv'. The percentage agreement and absolute values per sample are calculated in the file 'percentage_agreement.csv'.

The dataset will be made available on the Cultural AI Lab GitHub and through Europeana.

5. Use Cases & Outlook

As this project is undertaken in the context of the Cultural AI Lab, contentiousness was an evident choice as the core concept to structure this dataset around, since in our collaborations with cultural heritage institutions sensitivities around terminology comes up frequently. However, aside from a better understanding of the concept itself, contentiousness is also interesting for AI in a second way: as a complex and deeply cultural and social concept, it provides an interesting challenge in the pursuit of culturally aware and sensitive AI, both for developing new methods and testing existing ones.

Concretely, we aim to use ConConCor to gauge requirements for cultural AI with simple methods, i.e. given ConConCor, no heavy machinery is needed to investigate which properties, such as context-sensitivity and inclusion of perspective, cultural AI needs to have. Instead we can analyse the annotations and underlying texts with relatively simple statistical methods and go a long way. We foresee the following immediate avenues of research:

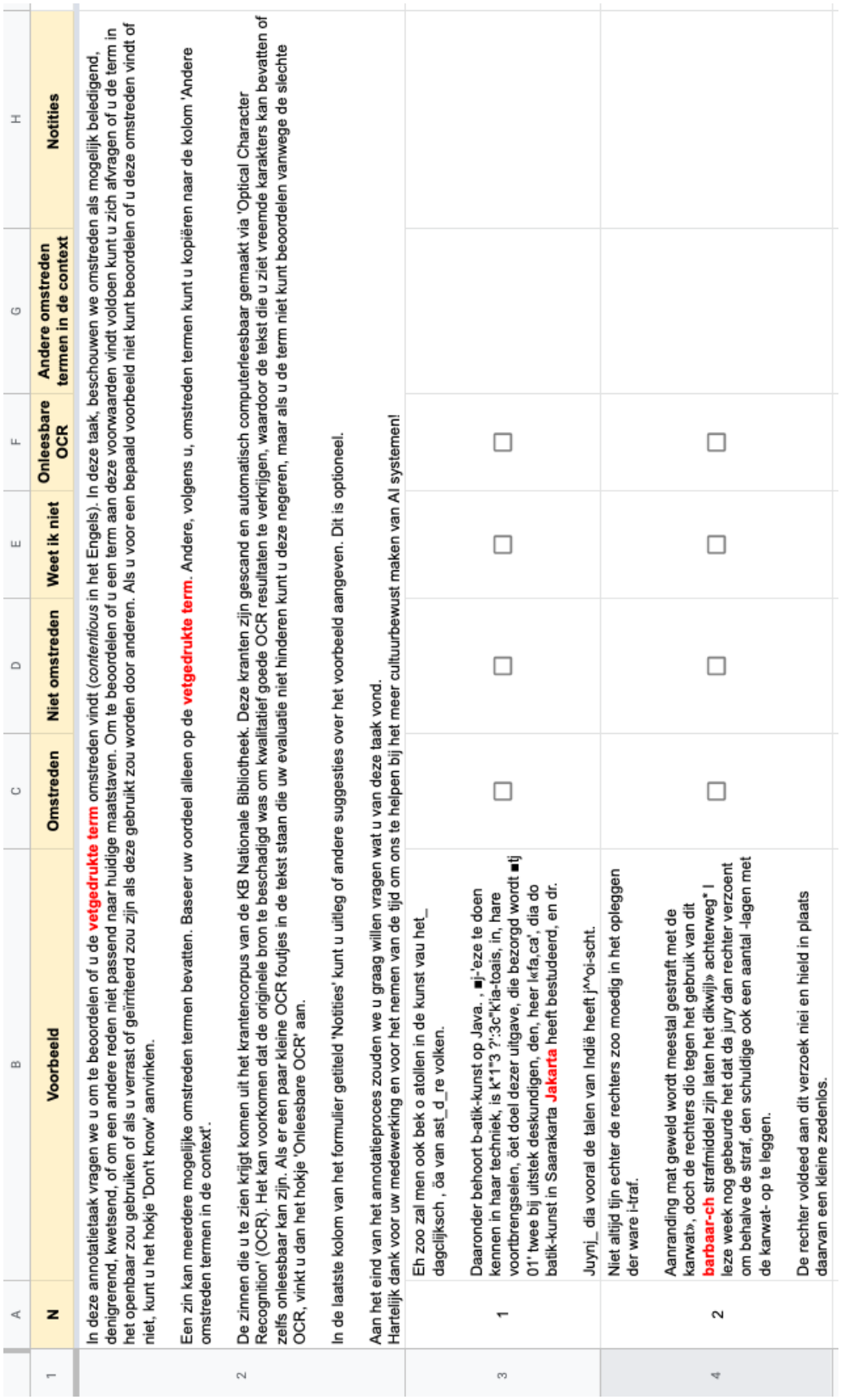
- Can we, and if so, to what extent, measure how contextual contentiousness is? To what extent and in what way does the context of a term influence how contentious it is perceived?
- If we can identify contentious contexts using ConConCor, can these contexts be generalised and used to detect additional contentious terms and/or contexts?
- Contentiousness is about perspective, as seen in the complex patterns of agreement above. Based on these patterns, and identity variables about the participants, can we recognise and analyse different perspectives?
- To what extent can ConConCor be used to investigate bias in other types of data such as other GLAM collections or contemporary newspapers?

In the Cultural AI Lab, we strive for creating traceable and interpretable systems. We will therefore first experiment with interpretable features and transparent machine learning models that allow us to make detailed conclusions about this intricate matter and does not unnecessarily obfuscate signals from an already complex dataset. Furthermore, we aim to develop these tools in close collaboration with GLAM professionals, such that the tools help them analyse their collections better and fit into their workflows. Our goal is for users who find ConConCor through Europeana to take a fresh look at their datasets or assumptions regarding bias and contentiousness in GLAM collections and other datasets.

Acknowledgments

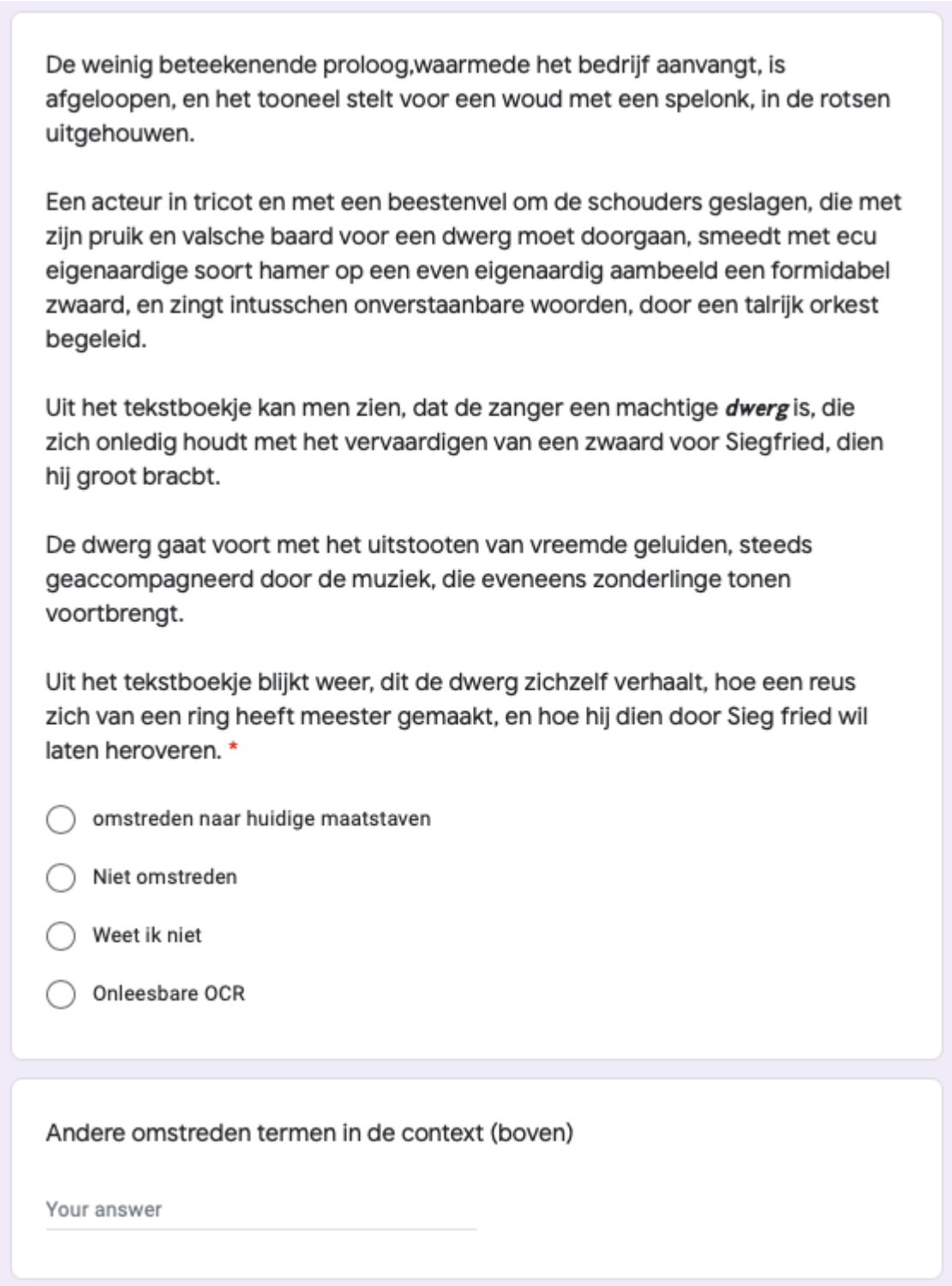
This work was funded by the EuropeanaTech Challenge for Europeana Artificial Intelligence and Machine Learning datasets. The authors would like to thank the Cultural AI Lab and KNAW HuC colleagues who provided feedback and performed annotations in the initial stages of the dataset creation and the anonymous Prolific annotators who so enthusiastically annotated the dataset. A special thanks to Mirjam Cuper of KB National Library of the Netherlands for providing guidance on KB and Europeana procedures.

Appendix A: Layout of the annotation sheet for the pilot study



	A	B	C	D	E	F	G	H
1	**N**	**Voorbeeld**	**Omstreden**	**Niet omstreden**	**Weet ik niet**	**Onleesbare OCR**	**Andere omstreden termen in de context**	**Notities**
	In deze annotatietaak vragen we u om te beoordelen of u de **vetgedrukte term** omstreden vindt (*contentious* in het Engels). In deze taak, beschouwen we omstreden als mogelijk beledigend, denigrerend, kwetsend, of om een andere reden niet passend naar huidige maatstaven. Om te beoordelen of u een term aan deze voorwaarden vindt voldoen kunt u zich afvragen of u de term in het openbaar zou gebruiken of als u verrast of geïrriteerd zou zijn als deze gebruikt zou worden door anderen. Als u voor een bepaald voorbeeld niet kunt beoordelen of u deze omstreden vindt of niet, kunt u het hokje 'Don't know' aanvinken.							
	Een zin kan meerdere mogelijke omstreden termen bevatten. Baseer uw oordeel alleen op de **vetgedrukte term**. Andere, volgens u, omstreden termen kunt u kopiëren naar de kolom 'Andere omstreden termen in de context'.							
2	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!							
3	1	Eh zoo zal men ook bek o atollen in de kunst vau het_ dagclijksch , öa van ast_d_re volken. Daaronder behoort b-atik-kunst op Java. , ■j-'eze te doen kennen in haar techniek, is k*1"3 ?':3c"k'ia-toais, in, hare voortbrengselen, öet doel dezer uitgave, die bezorgd wordt ■ij 01' twee bij uitstek deskundigen, den, heer k«fa,ca', dia do batik-kunst in Saarakarta **Jakarta** heeft bestudeerd, en dr. Juyrij_ dia vooral de talen van Indië heeft j^^oi-scht.	☐	☐	☐	☐		
4	2	Niet altijd tijn echter de rechters zoo moedig in het opleggen der ware i-traf. Aanranding mat geweld wordt meestal gestraft met de karwat», doch de rechters dio tegen het gebruik van dit **barbaar-ch** strafmiddel zijn laten het dikwijl» achterweg* I leze week nog gebeurde het dat da jury dan rechter verzoent om behalve de straf, den schuldige ook een aantal -lagen met de karwat- op te leggen. De rechter voldeed aan dit verzoek niei en hield in plaats daarvan een kleine zedenlos.	☐	☐	☐	☐		

Appendix B: Layout of the annotation form for the expert annotations



De weinig beteekenende proloog,waarmede het bedrijf aanvangt, is afgeloopen, en het tooneel stelt voor een woud met een spelonk, in de rotsen uitgehouwen.

Een acteur in tricot en met een beestenvel om de schouders geslagen, die met zijn pruik en valsche baard voor een dwerg moet doorgaan, smeedt met ecu eigenaardige soort hamer op een even eigenaardig aambeeld een formidabel zwaard, en zingt intusschen onverstaanbare woorden, door een talrijk orkest begeleid.

Uit het tekstboekje kan men zien, dat de zanger een machtige ***dwerg*** is, die zich onledig houdt met het vervaardigen van een zwaard voor Siegfried, dien hij groot bracbt.

De dwerg gaat voort met het uitstooten van vreemde geluiden, steeds geaccompagneerd door de muziek, die eveneens zonderlinge tonen voortbrengt.

Uit het tekstboekje blijkt weer, dit de dwerg zichzelf verhaalt, hoe een reus zich van een ring heeft meester gemaakt, en hoe hij dien door Sieg fried wil laten heroveren. *

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

Andere omstreden termen in de context (boven)

Your answer

Appendix C: Optional open questions for the pilot study

N	Dutch (original)	English translation
1	Was er genoeg context om te bepalen of u een term omstreden vond?	Was there enough context to make a decision about word contentiousness?
2	Vond u het moeilijk om te beoordelen of een term omstreden is?	Was it difficult to judge word contentiousness?
3	Vond u de instructies helder?	Was the instruction clear?
4	Heeft u nog andere opmerkingen of suggesties?	Do you have any other comments or suggestions?

Appendix D: Instructions for crowdsource workers

In deze annotatietaak vragen we u om te beoordelen of u de ***vetgedrukte 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!

Appendix E: Target words grouped by options and percentage agreement in experts' annotations

Percentage agreement	Contentious	
	Number of samples	Target words
100%	2	“eskimo”, “neger”
	1	“kaffer”, “slaaf”, “dwerg”
83–86%	4	“zigeuner”
	2	“neger”
	1	“slaaf”, “inheems”, “primitief”, “indiaan”, “eskimo”
67–71%	2	“inheems”
	1	“baboe”, “volbloed”, “blank”, “inboorling”, “kaffer”, “hottentot”, “dwerg”
43–57%	2	“pygmee”, “knecht”
	1	“dwerg”, “ras", “baboe”, “zwartje", "barbaren”
29–33%	0	–

Percentage agreement	I don’t know	
	Number of samples	Target words
100%	0	–
83–86%	0	–
67–71%	0	–
43–57%	3	"marron"
	1	“indo-europeaan”, “metis"
29–33%	0	–

Percentage agreement	Not contentious	
	Number of samples	Target words
100%	6	“gemengd”, “achtergrond"
	4	“historisch”,”aziatisch”, “zimbabwe”
	3	“moslim",
	2	"traditionele"
	1	“traditioneel”, “bediende", “jakarta”, ”handicap”, “ontdekken”, ”blank”,” roots”, “gekleurd", “gay", “exotisch", “indisch", "batavia"
83–86%	5	"moslim"
	2	"aziatisch"
	1	“primitief", “knecht", “dwerggroei", "berber", “trans", “afkomst", “jakarta", “historisch", “roma", “gemengd", “marron", “batavia", “inheems", "indo"
67–71%	3	"jakarta"
	2	“indisch", "gemengd"
	1	“kaukasisch" “stam", “aziatisch", “dwerggroei" “blank", “roma", “trans", “islamiet", “knecht", “indo-europeaan", “medicijnman", "historisch"
43–57%	2	“indo-europeaan”, "marron"
	1	“moslim", “birma", “mongool", “slaaf", "berber"
29–33%	0	–

Appendix F: Top target words grouped by options and percentage agreement in crowdsourcing annotations

Percentage agreement	Top target words by the number of samples they appear in		
	Contentious	**Not contentious**	**I don't know**
100%	"neger", "negerslaven", "bosneger", "negerrijk", "zwartje", "negercultuur"	"achtergrond", "historisch", "traditionele", "aziatisch", "gemengd", "inheems", "moslim", "wild", "ontdekken", "afkomst"	"marron"
83–86%	"lilliputter", "negerslaven", "neger", "negerrijk", "negercultuur", "zigeuner"	"aziatisch", "moslim", "achtergrond", "gemengd", "historisch", "traditionele", "inheems", "dwerggroei"	"marron"
67–71%	"halfbloed", "zigeuner", "lilliputter", "inboorling", "slaaf", "pygmee", "dwerg", "mulat", "misvormden"	"indo-europeaan", "inheems", "moslim", "gemengd", "aziatisch", "marron", "dwerggroei", "historisch"	"marron", "mesties", "baboe", "koelie", "metis"
43–57%	"indo-europeaan", "eskimo", "lilliputter", "kaffer", "zigeuner", "indo", "koelie", "hottentot"	"marron", "indo-europeaan", "inheems", "gemengd", "moslim", "koppensneller"	"marron", "mesties", "mulat", "baboe", "koelie", "metis", "muzelman"
29–33%	"koppensneller", "hottentot", "muzelman"	"homo", "dwerggroei", "inlander", "hottentot", "indo-europeaan"	"marron", "koelie", "kokkie", "baboe", "muzelman", "metis"