But they do talk a bit funny, don't they?

Ryan Brate[1], Marieke van Erp[1], and Antal van den Bosch[2]

[1]KNAW Humanities Cluster, DHLab, Amsterdam, the Netherlands
[2]Utrecht University, Institute for Language Sciences, Utrecht, the Netherlands

1 Introduction

> Gimme dat ar axe en set right down and wait twel supper. You're des es white es a sheet dis minute [The Battle Ground by Ellen Anderson Gholson Glasgow]

On first read, above quotation demands a degree of deciphering from the reader, due to its blunt use of non–canonical spelling signalling a variation on pronunciation, or eye dialect Krapp (1926). The author is communicating something about the speaker to the reader. One can imagine authors wishing to *emphasise* a dialectal, regiolectal or socialectal variant; or perhaps a feature of speech specific to the individual. Regardless, a distinct speech pattern becomes a characteristic for interpretation by the reader, a marker for the *othering* of an individual or group with respect to population remainder. Indeed, the enforcement of linguistic norms within groups is a recognised consequence of tight-knit social networks Milroy and Milroy (1978). Variation in speech is also recognised in the field of sociolinguistics as a potential indicator of socioeconomic status Labov (1966), Trudgill (1974).

The potential for speech patterns as a social discriminator is especially relevant to the earlier quotation: the quotation is attributed in the original text to the term, *negro*. Given the implied historical connotations of this word as a vehicle for *othering*, we may reasonably ask, how is such *othering* reflected in the speech attributed to the term, and other such charged terms with serve to *other*? Motivated by the potential for readers to distinguish speakers based on their speech.

In this research, we are concerned with confirming whether there are statistically observable differences in the speech attributed to *various* charged terms, as compared to the general population. We do so in the context of the fiction and news domains. The former representing fictional characterisations and the latter representing purported truth. We ask the following research questions: *How strongly do quotations from charged terms differ statistically from the rest of the literature fiction domain?*; and *How strongly do quotations from charged terms differ statistical between fiction and news domains?*

2 Data

We use a fiction corpus from a total of 8,670 books from Project Gutenberg's American Literature collection, [1] yielding a total of 6,690,370 quotations from regex pattern matching. A limited news corpus is assembled manually from the Library of Congress digitized newspaper collection [2]. The first 100 *negro* quotations are collected from the search phrase, *negro said* for years, 1900–1963.

3 Analysis

We quantify how *surprising* the quotations are. We tokenize and estimate unigram and bigram token probabilities, with respect to the entire fiction sub-corpus. We include Laplace smoothing (to accommodate unseen news quotations). We calculate the negative log likelihood scores of each quotation, averaged over their respective lengths (AvNLL) as per Equation 1.

$$\text{AvNLL(quotation of length, N)} = \frac{-log(P(word_1) + \sum_{i=2}^{N} P(word_i|word_{i-1}))}{N} \quad (1)$$

Firstly, we consider the literature corpus only, contrasting AvNLL scores for selected terms against their respective remaining populations. We consider the reference terms, *gentleman*, *lady* and *drunk*: on the expectation that a *gentleman* and *lady* likely tend to the population norm, whereas a *drunk* may be expected to demonstrate lesser speech coherence. We consider 2 sets of known charged terms: those with racial connotations (*negro*, *colored*, *oriental*); in addition to *dwarf* and *cripple*.

Secondly, we consider the news corpus. Contrasting the AvNLL scores of *negro* quotations in the news corpus with: i) *negro* speaker quotations in the fiction corpus and ii) all quotations in the fiction corpus, removed of *negro* speakers.

The statistical difference, with respect to each aforementioned comparisons is estimated by Bayesian estimation Kruschke (2012). Specifically, we estimate the *magnitude* in difference between distribution means according to a Bayesian adaption of Cohen's d size affect factor metric (d) 2 – an estimate of difference in two distribution means as a fraction of pooled standard deviation.

$$d = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} \quad (2)$$

Each distribution of AvNLL scores is modelled by a separate Student's T distribution, adopting unopinionated priors as per equation set 3. The posterior probability distribution of credible values for each modelled distribution's mean, is updated based on observed AvNLL values via Markov Chain Monte Carlo, using PyMC Oriol et al. (2023). A posterior distribution of credible d values is then assembled according to Equation 2, for each distribution comparison.

$$\begin{aligned} AvNLL \sim StudentT(mu, nu, sigma) mu &\sim Normal(0, 5) \\ nu &\sim Exponential(1) \\ sigma &\sim Exponential(1) \end{aligned} \quad (3)$$

[1] `https://www.gutenberg.org/`, 05/02/2024

[2] `https://chroniclingamerica.loc.gov/newspapers/`

4 Results

Histograms of the literature corpus AvNLL comparisons are given in Figure 1. As expected, the general quotation population (light grey) is approximately Gaussian. As hypothesised, *gentleman* and *lady* closely approximate the overall population distribution, and similarly for *dwarf* and *cripple*. This is clearly contrastable to the distributions for *drunk*, *negro*, *colored*.

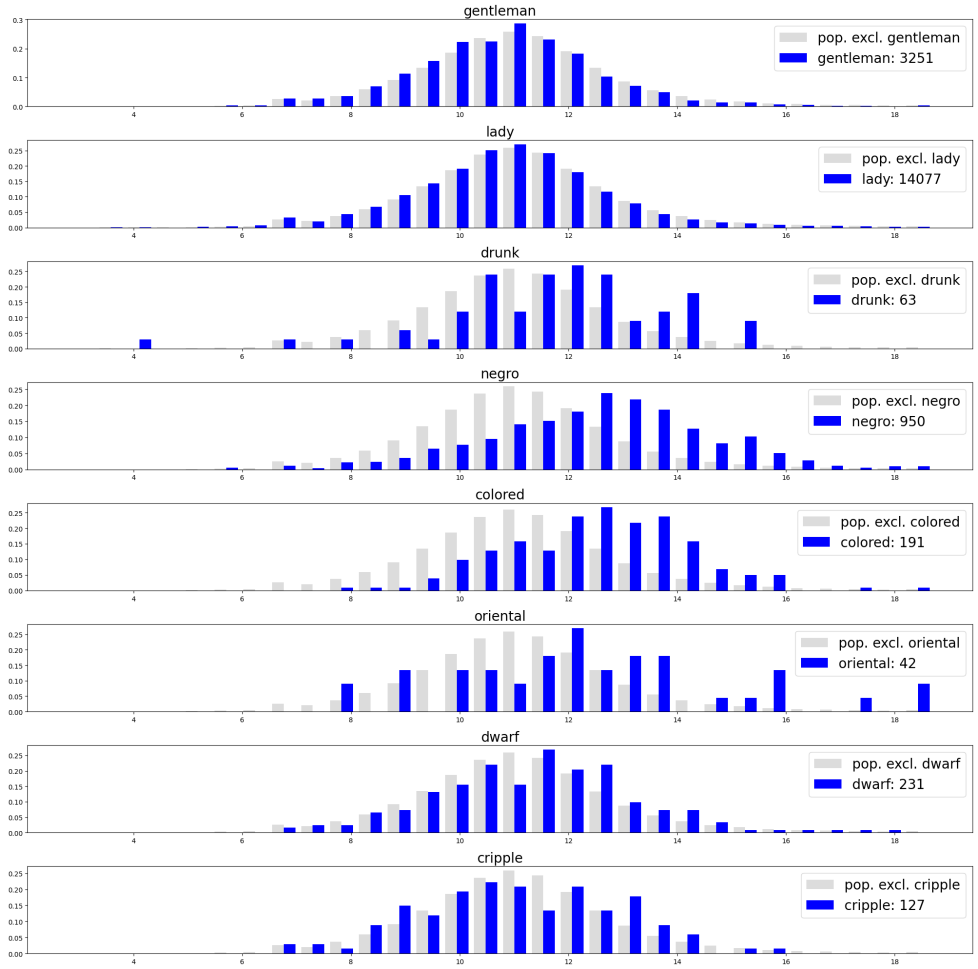

Figure 1: Distribution of Average Negative Log Likelihood (AvNLL) by term in the Literature Corpus. Each pair of bars on the distributions of each plot, correspond to the same AvNLL interval for pairwise comparison. Term quotation counts are given in the respective legends.

The Bayesian analysis reports Credible Intervals (CI) of d, which can be interpreted as the posterior probability of values of d predicated on the model and data. Informally, d values of 0.2, 0.5 and 0.8 are considered as small, medium and large differences between comparative distribution means Cohen (1988). Also, since d is a factored difference in means, a CI which does not include 0, represents a corresponding posterior probability of a non–zero difference in means.

In Table 1 we report the 99% CI for d with respect to the literature corpus comparisons. The analysis suggests *large* statistical differences for *negro* and *colored*, and *negligible to small* differences for *gentleman*, *lady*, *dwarf* and *cripple*.

In Table 2, we report 99% CI for d in respect of the constrast between *negro* quotations in the news corpus versus the literature corpus. The analysis suggests non–zero

differences between the *negro* quotations of the news corpus and the literature corpus, with less severe deviation in the news corpus from the general population. However, the analysis is not confident of the effect size.

5 Conclusion and Future Work

The results support our initial intuition: there are meaningful and measurable differences in the attributed speech of *some* charged terms in literature. The results are also suggestive of more extreme *characterisation* of *negro* speech in the literature domain than the news domain. For future work, we propose to consider: more fine–grained proxy metrics which target specific speech attributes; an expanded scope of news quotations; and the relative association between terms according to these metrics, indicative of possible reader inferences of relatedness. As the corpus of digitised historical texts grows, and with it the access to outdated views on society, identifying charged terms to contextualise them will become more important.

Term	**99% CI wrt., Cohen's d scores**
gentleman	-0.21 to -0.11
lady	-0.14 to -0.089
negro	+0.81 to +1.0
colored	+0.81 to +1.3
dwarf	-0.006 to +0.34
cripple	+0.004 to +0.39

Table 1: Reported Credibility Intervals of Cohen's d size effect factors for the difference in the posterior estimates of term AvNLL distribution means (term vs population removed of term). The models for the distributions of *population removed of term*, are estimated from random sub-sample of 100,0000 data points for computational efficiency. Note: *Oriental* is not considered owing to its clearly non–normal distribution in Figure 1.

Quotations to be compared with *negro* quotations from the news corpus	**99% CI wrt., Cohen's d scores**
literature fiction quotations removed of 'negro'	0.30 to 0.91
literature fiction quotations for 'negro'	-0.70 to -0.14

Table 2: Reported Credibility Intervals of Cohen's d size effect factors, for the difference in quotation scores between news quotations for *negro* speakers, against: a) fiction quotations removed of *negro* speakers; and b) fiction quotations consisting of only *negro* speakers

Reproducibility

All data and code, necessary to reproduce the results can be found on the GitHub repository.

Acknowledgements

This work was funded by NWO in the 'Culturally Aware AI' project.

References

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

George Philip Krapp. The psychology of dialect writing. *The Bookman*, 62:522, 1926.

John Kruschke. Bayesian estimation supersedes the t test. *Journal of experimental psychology. General*, 142, 07 2012. doi: 10.1037/a0029146.

W. Labov. *The Social Stratification of English in New York City*. Books on demand. Center for Applied Linguistics, 1966. ISBN 9780872811492. URL `https://books.google.nl/books?id=NbpZAAAAMAAJ`.

J Milroy and L Milroy. *Belfast: change and variation in an urban vernacular*. London : E. Arnold, 1978.

Abril-Pla Oriol, Andreani Virgile, Carroll Colin, Dong Larry, Fonnesbeck Christopher J., Kochurov Maxim, Kumar Ravin, Lao Jupeng, Luhmann Christian C., Martin Osvaldo A., Osthege Michael, Vieira Ricardo, Wiecki Thomas, and Zinkov Robert. Pymc: A modern and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9:e1516, 2023. doi: 10.7717/peerj-cs.1516.

Peter Trudgill. *The social differentiation of English in Norwich*, volume 13. CUP archive, 1974.