Estimating gender asymmetries in swearing

Bootstrapping versus log-likelihood ratios in corpus research

June 2019

AUTHOR	Raoul Buurke
COURSE	Methodology and Statistics for Linguistic Research
TEACHER	prof. dr. Martijn Wieling
TEACHER	dr. Antonio Toral Ruiz

1 Introduction

Swearing as expressive linguistic behavior has many facets. Swear words elude a common linguistic analysis, because they cover an interestingly heterogeneous group of words. They are semantically disparate (Napoli & Hoeksema, 2009), but are at the same time syntactically constrained, e.g. in *wh*-questions (Hoeksema & Napoli, 2008). Next to their troublesome linguistic status, it seems that there is an asymmetry in their use between males and females(McEnery, 2004). Based on an analysis of the British National Corpus (BNC, 2007) and intuitions on the basis of TV regulations on swearing (e.g. OFcom, 2016) McEnery claims that males more often use "stronger" swear words, such as *fucking*. Females tend to use "milder" words, such as *bloody*. Whether this pattern holds under a different kind of corpus analysis is interesting, because linguistic effects are not typically constrained by gender, and there is no apparent reason to assume it for expressive language. Next to this, if the same corpus data provides crucially different results under different approaches, there is probably one preferred method, which calls into question the regular practices in the research field.

A few words about methodology are in place here. McEnery computed log-likelihood (LL) values for the frequency data he collected, which is common in corpus linguistics since Dunning (1993). The test essentially compute a 2 × 2 contingency table and a p-value is computed from the χ^2 distribution with 1 degree of freedom. Dunning noted that word frequency distributions are notoriously skewed, because there are always many low frequency items that cause heavy tails. Relying on LL ratios indeed solves this problem, but these tests have been called into questions themselves. Lijffijt et al. (2016) compare different methods for comparing corpus frequency data, which also included LL ratios. They conclude that this test is anti-conservative, which means it inflates the Type I error rate. It moreover *still* relies on an incorrect assumption for frequency data, which is that the data are independently distributed across the text. This assumption is problematic, given that it essentially fails to take into account that texts have structure themselves.

Luckily, there are ways to tackle these problems. Specifically, *bootstrapping* is an intuitive approach to statistical inference that does not rely on parametric assumptions. Early efforts include those by Efron (1979), but approaches based on resampling (though without replacement) go back as far as Fisher (1937). The basic idea is to assume that the collected data sample reliably reflects the population distribution, as opposed to making specific and restrictive assumptions about the shape of the population. It can be used to estimate almost any statistic by simply recomputing it under (slightly) different configurations of the sample. This is typically done N times, where N is large (i.e. at least 1000). Through this Monte Carlo method a sampling distribution can be constructed, which provides inferential information. Several researchers have advocated the use of bootstrapping as a general approach to analysis in linguistics, e.g. Larson-Hall & Herrington, 2009. Larson-Hall and Herrington build on the indirect observation by Wilcox (2010) that typical sample sizes in linguistics are much too low to reliably control for Type I error rates. The possible usefulness of bootstrapping for corpus linguistics provides further support for these ideals.

Lijffijt et al. (2016) propose a bootstrapping approach in their comparison of tests, which relies on a different data structure compared to LL ratios. They opt for comparing *two vectors of normalized frequencies* as opposed to *two normalized frequencies*. Each vector value is the normalized frequency of the relevant words in a particular text. These vectors are bootstrapped and a test can be constructed from the resulting distributions (elaborated in the

Methods section). This approach allows for a more informative look than the typical one-on-one comparison of LL ratio tests. It is to be expected that if the gender effect on expressive language is truly valid the same results will be found under the bootstrapping approach. There is no particular reason to assume one outcome over another, so no predictions are set before conducting replicating McEnery's interesting work.

2 Methods

2.1 Materials

The data are the same as in McEnery (2004), i.e. the British National Corpus (BNC). The BNC is rich in metadata: there is information about the source and sex of the author of most text samples. This is crucial for the current analysis, because it allows separating the data by male and female authors. Assuming that the gender difference is reflected by this feature in the corpus no further preparation is required besides deciding what part of the corpus is used. McEnery himself does not state explicitly how he selected data for his analysis. For replication purposes only texts from *books* (as opposed to e.g. periodicals) are used. It can be argued that writers have a motive for writing books without offensive language in them, but it is reasonable to assume that this effect is not so large nowadays that it should be problematic, or that other text sources are significantly better in this regard. Moreover, the amount of data in the BNC that comes from books (>50M words) is much larger than that of the periodical genre (circa 28.5M words) or any other.

McEnery used a substantial list of swear words, but there is little reason to replicate the results for *all* items. The interaction effect of gender and swear word use should occur in any case, so the data will be checked for five specific swear words: *bloody*, *hell*, *ass*, *bastard*, and *fucking*. Their respective levels of offensiveness are reported in Table 1.

The rationale behind these particular swear words is as follows. *Bloody* and *hell*[1] are both considered "very mild" swear words, so it is to be expected that women either use them more or there is no significant difference between males and females. The opposite is assumed for *fucking*, which is chosen as opposed to *fuck* as its frequency is much higher (McEnery & Xiao, 2004). The "middle" cases of *ass* and *bastard* are primarily chosen for their place in the offensiveness hierarchy proposed by McEnery (as they were not in the original analysis), and their frequencies (see Table 1). This latter motive does not mean that the normalized frequencies are similar, but different instead. This is a useful distinction to make, as it will show whether either method is influenced significantly by sample size.

2.2 Measures

2.2.1 Log-likelihood ratio

Since Dunning (1993) there have been a few slightly different formulations of the original LL approach, but the one that is used is here is as noted in Rayson & Garside (2009). The

[1] These words can occur as a collocation *bloody hell*. The normalized frequencies per 1M words for *bloody*, *hell*, and *bloody hell* in the genre "books" are respectively 51.72, 59.06, and 1.85. The pointwise mutual information, the common formula for which is reported in i.a. Bouma (2009), is 6,4. This value is greater than 0, so there is evidence to assume that *bloody hell* is a collocation, but this is ignored as the occurrence is sufficiently low.

Table 1

Levels of offensiveness as reported in McEnery (2004) and normalized frequency per 1M words in the "books" genre of the BNC.

Swear word	Offensiveness	Normalized frequency
bloody	Very mild	51.72
hell	Very mild	59.06
ass	Mild	4.04
bastard	Moderate	18.75
fucking	Very strong	15.81

formula for the log-likelihood of any swear word α between the male corpus M and female corpus F is reported in (1) and (2). The critical LL value for a 0.05 significance level is 3.841, which is the critical value of a χ^2 distribution with 1 degree of freedom. This works as LL asymptotically follows this distribution. This test assumes that the occurrence of all words in the corpus can be modeled as a Bernoulli process (i.e. a series of random events with a binary outcome). The obvious consequence is that it assumes statistical independence of words.

$$LL = 2\Big(\big(f_M * \ln(\frac{f_M}{E_M})\big) + \big(f_F * \ln(\frac{f_F}{E_F})\big)\Big) \tag{1}$$

For which f = raw frequency of α, and E = expected frequency in corpus.

$$E_i = \frac{C_i(f_M + f_F)}{(C_M + C_F)} \tag{2}$$

For which $i = \{M, F\}$, f = raw frequency of α, and C = corpus size.

2.2.2 Bootstrap test

Lijffijt et al. (2016) devised an alternative test based on bootstrapping. For each text T within the BNC corpus it is noted whether or not the author of the text was either male or female, so the total corpus is split into a male M and female F corpus. Each vector value is the raw frequency of α in a text divided over the pooled corpus size, i.e. the summed number of tokens across all texts of M or F. An example of the two relevant vectors for the computation (further below) is given in Table 2. Note that these vectors are of length N, which need not be the same for both vectors, because there may be more female than male written texts. For the bootstrap the smallest N will be chosen as a resample size, i.e. if $|\vec{m}| > |\vec{f}|$, then all bootstrap samples are of size $|\vec{f}|$.

To compute a p-value for the difference between $\vec{m}$ and $\vec{f}$ Lijffijt and colleagues propose computing a mid-P test (based on Berry & Armitage, 1995; computed as in (3) and (4)). This value essentially computes a one-tailed p-value, and the formula in (5) is used to compute the corresponding two-tailed value.

$$p_{\text{one-sided}} = \frac{\sum_{i=1}^{N} H(f(\alpha, \vec{m}_i) - f(\alpha, \vec{f}_i))}{N} \tag{3}$$

Table 2
Vectors to be bootstrapped in the Lijffijt et al. (2016) procedure. The values are the normalized frequencies of swear word α in any male ($\vec{m}$) or female ($\vec{f}$) written text.

Swear word (α)	
Male author vector ($\vec{m}$)	Female author vector ($\vec{f}$)
$\vec{m}_{\alpha}^{1}$	$\vec{f}_{\alpha}^{1}$
$\vec{m}_{\alpha}^{2}$	$\vec{f}_{\alpha}^{2}$
...	...
$\vec{m}_{\alpha}^{\mathrm{N}}$	$\vec{f}_{\alpha}^{\mathrm{N}}$

For which H is a step function dependent on the frequency offset (see (4) below) and f is the mean of either $\vec{m}$ or $\vec{f}$of α. N is the number of bootstrap samples. It should always be set to at least 9999, as it is an empirical estimate and otherwise p-values of 0.001 can never be reached.

$$H(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0.5, & \text{if } x = 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{4}$$

For which x is the offset of the mean bootstrapped normalized frequency between $\vec{m}$ and $\vec{f}$. The function above essentially returns 1 if the bootstrapped mean of $\vec{m}$ is greater than the one of $\vec{f}$, while it returns 0 in the opposite case. It returns 0.5 if the bootstrapped means are equal. The resulting p-value is always in the range [0,1] after normalizing over N.

$$p_{\text{two-sided}} = \frac{1 + N * 2 * min(p_{\text{one-sided}}, 1 - p_{\text{one-sided}})}{1 + N} \tag{5}$$

The function above turns the one-tailed p-value into a two-tailed one, while smoothing to account for possible extremely small p-values (and since the one-sided p-value is an empirical estimate).

3 Results

The results are reported in Tables 3 and 4, with the summary of test conclusions in Table 5. The LL approach summarized in Table 3 mostly shows similarity with McEnery (2004), i.e. significant overuse by females of *bloody* and *hell*, although McEnery reports significant overuse by males for *fucking*. The current data actually show overuse by females, and the result is contrastingly not significant ($p > 0.05$), so it is likely that there is no difference between them. Interestingly, there is only one case across both procedures where the ratio indicates overuse by males in any case. This goes against the general conclusion of McEnery, because there is no asymmetry in swear word usage *across* levels of offensiveness. Instead, these results suggest that female writers use swear words more often than male authors in general.

Table 3

Summary of relative overuse by male/female per swear word in the pooled corpus, and the corresponding LL test results for the current analysis and McEnery (2004). Ratio values >1 indicate overuse by males, while values <1 indicate overuse by females.

Swear word	**Relative frequency (in pooled corpus)**	**LL**		**p-value**	
	Male/female ratio	McEnery (2004)	Current	McEnery (2004)	Current
bloody	0.616	314.150	11.289	<0.001	0.001
hell	0.212	15.690	209.808	<0.001	<0.001
ass	0.659	-	0.849	-	0.357
bastard	0.433	-	11.147	-	0.001
fucking	0.791	350.830	0.612	<0.001	0.434

Table 4

Summary of the bootstrap test results (with 10000 bootstrap samples). Ratio values >1 indicate overuse by males, while values <1 indicate overuse by females.

Swear word	**Mean bootstrapped relative frequency**	**p-value**
	Male/female ratio	
bloody	0.761	0.035
hell	0.193	0.001
ass	0.685	0.056
bastard	1.027	0.419
fucking	0.180	0.630

Table 5

Difference in test conclusions with a significance level of 0.05. P-value ratio values >1 indicate lower p-values of the LL test. Values <1 indicate lower p-values of the bootstrap test.

Swear word	**Difference in test conclusion**	**P-value ratio**
bloody	No	45.010
hell	No	4.628*10ˆ43
ass	No	0.158
bastard	Yes	498.150
fucking	No	1.452

The results from the bootstrap approach in Table 4 are, interestingly, also not too different from the LL approach in terms of conclusions (see Table 5) and male/female ratios. The ratios for *bloody*, *hell*, and *ass* are quite similar. For *fucking* the ratio shows a stronger preference by females, but the result is not significant (p=0.630). For *bastard* there is a different test conclusion: the LL approach results in a significant difference (p = 0.001), while this does not hold for the bootstrap test (p = 0.419). Note also that the bootstrapped direction is different: more use by males than females, although this is likely by chance given the small offset from 1.

Another interesting difference is highlighted by the *ass* results. This is the only case where the bootstrap test produces a lower p-value than the LL approach. In fact, it nearly reaches significance. Recall also from Table 1 that *ass* occurs the least in the BNC data. This can be taken as an indication that the difference between these methods is most clear when dealing with relatively few data. Taking into scope the other cases, it is also clear that the LL approach often produces much lower p-values, which is in accordance with the conclusion that it is anti-conservative (Lijffijt et al., 2016), although it is not a direct proof[2].

4 Discussion

A few interesting patterns are apparent when the results are pooled together. First and foremost: the interaction effect between offensiveness of swear words and gender reported by McEnery (2004) is not replicated. Especially the fact that *fucking* is both overused by females, and that this difference is not significant in either approach, serves as a clear indication of this conclusion. All evidence points towards concluding that female writers use more swear words in their texts. This is an interesting and somewhat unintuitive result given the fact that the same corpus is used. The most likely source of this contrast is the fact that for the current analysis a subcorpus instead of the full corpus was used. This either means that the effect is in fact not there, or that selecting another subcorpus might still reproduce the original results. Conjecturing what corpus this could be is left to future efforts, but the effect is possibly masked by the fact that these data come from books. It is fair to question to what degree the language use of writers reflects actual gendered language use in the spoken register, which is of course the most natural register. If unwilling to accept the "books" genre as an accurate reflection of these patterns, one can replicate the analysis with a spoken language corpus. This was not carried out in this paper, because (1) comparison with McEnery (2004) would have been impossible given its focus on the complete BNC, and (2) the spoken BNC has fewer metadata and data, so e.g. *ass* for the female data would consist of in total two cases in two texts.

Another interesting outcome is the apparent similarity in results despite clearly different approaches. There is, however, considerable evidence that log-likelihood ratios are anti-conservative estimates of frequency differences. The p-value ratios in Table 5 are quite extreme, especially for *hell*. Another argument in favor of the bootstrapping approach is its apparent lower sensitivity to small sample sizes. LL ratio tests essentially compute a 2×2 contingency table and then derive their p-values from the χ^2 distribution with 1 degree of freedom. One crucial assumption is that the expected values for each cell should not be too low, i.e. less than 5[3]. Levshina (2015), for example, notes that p-values are unreliable (often too high) in such cases. In conclusion: LL ratio tests are anti-conservative with large samples and too conservative with small samples. Bootstrapping has no specific requirements (except that the data are independently selected from the population), and it has been proven to work reliably with small sample sizes (e.g. Hinneburg et al., 2007; Chernick & LaBudde, 2014).

One possible shortcoming of the bootstrapping approach is that the required data structures can result in extremely small sample sizes. It is not unlikely that the frequency

[2] A Kolgomorov-Smirnov procedure to account for this is provided by Lijffijt and colleagues, but this is not carried out here for reasons of time and space.

[3] Computing these values for *ass* results in 9.9 and 10.1 for the male and female data respectively, which means the LL approach was still validly applied. It is unlikely that inflation or deflation effects have a clear cut-off point, though, so it remains problematic.

of swear words in any given text is quite low. For some texts, e.g. books with more personal and emotional dialogue, the frequency is high enough for bootstrapping. In others it might be that the frequency is as low as 1, which defeats the purpose of the bootstrapping approach: to account for textual structure. Naturally, there is no easy way around this. Selecting only texts with a certain frequency threshold equally defeats the purpose of the bootstrapping approach. The best solution is probably to carefully form assumptions about which type of genre best reflects a difference between male and female language use, e.g. informative books less so than novels, and consequently use data from such sources.

5 References

Berry, G., & Armitage, P. (1995). Mid-P confidence intervals: a brief review. *Journal of the Royal Statistical Society: Series D (The Statistician), 44*(4), 417-423.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 31-40.

British National Corpus, version 3. (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Chernick, M. R., & LaBudde, R. A. (2014). *An introduction to bootstrap methods with applications to R*. John Wiley & Sons.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics, 19*(1), 61-74.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist, 7*(1), 1-26. doi:10.1214/aos/1176344552

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Hoeksema, J., & Napoli, D. J. (2008). Just for the hell of it: A comparison of two taboo-term constructions. *Journal of Linguistics, 44*(2), 347-378.

Hinneburg, A., Mannila, H., Kaislaniemi, S., Nevalainen, T., & Raumolin-Brunberg, H. (2007). How to handle small samples: bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and linguistic computing, 22*(2), 137-150.

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*(3), 368-390.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Literary and Linguistic Computing, 31*(2), 374-397.

McEnery, T. (2004). *Swearing in English: Bad language, purity and power from 1586 to the present*. Routledge.

McEnery, T., & Xiao, Z. (2004). Swearing in modern British English: the case of fuck in the BNC. *Language and Literature, 13*(3), 235-268.

Napoli, D. J., & Hoeksema, J. (2009). The grammatical versatility of taboo terms. *Studies in Language*, *33*(3), 612-643.

OFcom (2016). Attitudes to potentially offensive language and gestures on TV and radio. Retrieved from https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/offensive-language-2016 . Published on 30 September 2016.

Wilcox, R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Springer.