# Bad language revisited: swearing in the Spoken BNC2014

Robbie Love (Lancaster University, UK)

## 1. Introduction

This paper reports on the use of McEnery's (2005) approach to analyzing swearing in spoken British English to investigate the use of bad language words (BLWs) in a sample of the Spoken British National Corpus 2014 (Spoken BNC2014S) (Love et al. 2017 fc), comparing this corpus with the original Spoken British National Corpus (Spoken BNC1994) (Leech 1993). The Spoken BNC2014 comprises transcripts of spontaneous, present day, informal conversations among speakers of British English recorded between 2012 and 2016.

## 2. McEnery's approach to swearing, and other typologies

McEnery (2005) analyses so-called *bad language words* (BLWs) in the Spoken BNC1994DS (demographically-sampled component). He uses the term *swearing* broadly to encompass a set of bad language words (BLWs), which includes literal and non-literal use of swear words (e.g. SHIT, FUCK) as well as other words which may be used offensively but which would not be considered *swear words* per se (e.g. PIG, TART). In addition to quantitative analysis of their distribution across the sociolinguistic categories of gender, age and socio-economic status, McEnery conducts qualitative analysis of each BLW using a bespoke bad language categorization scheme (originally developed for the Lancaster Corpus of Abuse – LCA, McEnery et al. 1999, 2000). He finds that "the use or lack of use of BLWs is a fault line along which age, sex and social class may be differentiated" (p. 50).

The LCA annotation scheme survives not without criticism. Ljung (2011: 12) defines bad language via a typology that has much in common with McEnery's scheme but, crucially, excludes literal uses of swear words: "taboo words with literal meaning cannot be regarded as swearing". Furthermore Ljung (p. 28) criticises several of the categories of the LCA scheme including *Idiomatic set phrase, Imagery based on literal meaning* and *Pronominal form with undefined referent,* suggesting that the scheme ought to be used with caution.

## 3. Method

The aim of this work is to replicate McEnery (2005) by analyzing a large set of BLWs in the Spoken BNC2014S and comparing their frequency, sociolinguistic distribution and use to that of the Spoken BNC1994DS, commenting on any changes in bad language over the last twenty years (for the sake of comparability with research on bad language in the BNC1994DS, I adopt McEnery's definition of BLWs as opposed to that of Ljung, 2011). To do this, studying only the set of 50 BLWs from McEnery (2005) would be insufficient; I had to take into account the possibility of new BLWs having emerged since the early 1990s. I therefore extended the original list of BLWs by adding those that were included by the UK's Office of Communications (Ofcom) in their guide to offensive language in broadcast media (Ipsos MORI 2016), as well as a set developed by Lutsky & Kehoe (2015). The use of these sources as bases for extending McEnery's original list resulted in a new set of 173 BLWs.

This study was thus conducted as per the following:

(1) Search in the Spoken BNC1994DS and Spoken BNC2014S for each BLW in turn;
(2) Observe BLWs which have a frequency of zero in both corpora and eliminate from further analysis;
(3) Analyse BLWs which have changed in relative frequency the most drastically between the two corpora;
(4) Demographic distribution: select some of the most commonly occurring BLWs in both corpora, and record their frequency per speaker metadata category according to gender, age and socio-economic status;
(5) Annotate them according to the LCA annotation scheme (McEnery 2005);
(6) Comment on possible language change in light of differences between the two corpora in the BLWs' speaker metadata distribution and categorization.

## 4. Data

Both corpora were accessed via Lancaster University's CQPweb server (Hardie 2012). The Spoken BNC1994DS contains 5,014,655 tokens across 153 texts, while the Spoken BNC2014S contains 4,789,185 tokens across 567 texts.

In terms of corpus comparability, it could be argued that since neither of the Spoken BNCs were sampled with the explicit aim of studying BLWs, it is difficult to claim that the sampling conditions allowed for a comparable amount of BLW use. However, it can firstly be assumed that the Spoken BNC1994 facilitated the natural occurrence of BLWs given its surreptitious approach to recording (Crowdy 1993: 260). Secondly, the aim of the Spoken BNC2014 team was to facilitate the recording of conversations in a way which minimized intrusiveness beyond what was required of modern ethics procedures (Love et al. 2017). Harry Strawson, a Spoken BNC2014 contributor who submitted over a dozen recordings, claimed that "it was surprising how quickly people seemed to forget they were being recorded" (Strawson 2017).

## 5. Initial results: wholesale frequency analysis, and the case of FUCK

32 BLWs, including BUKKAKE, FATASS and PUNANI, were found to have a frequency of zero in both corpora. Many of these are described by Ofcom (2016) as having "low recognition" among focus group participants, and several were labelled as having been identified by less than 40% of participants in an online survey of the words. Based on this it is perhaps unsurprising that they do not occur in the corpora.

141 remaining BLWs occur at least once in either of the corpora. Among these, some were found to have decreased in relative frequency significantly ($p<0.0001$) between the 1990s and 2010s, including SPASTIC, CUNT and BUGGER. Those which have increased significantly ($p<0.0001$) include RETARD, DYKE and SHIT, while there are some – including ARSE, BITCH and FUCK – which have very similar relative frequencies in both corpora and have therefore shown stability.

Initially, I then looked at one of these stable BLWs – FUCK – in more detail. By applying the relevant steps as outlined in the Method, I could assess how FUCK has changed in spoken British English in the last two decades, according to the sociolinguistic variables of gender, age and socio-economic status.

The results suggest that in present-day spoken British English:

- FUCK is now used equally as frequently by male and female speakers.
- The use of FUCK peaks among speakers in their twenties and decreases with age, apart from the 60-69 group which has a higher frequency than 50-59.

- The distribution of FUCK according to social class is similar to that of the Spoken BNC1994DS but only if the same classification scheme (Social Grade) is used. If a newer scheme is used (NS-SEC), then it is speakers in the middle of the scale that seem to use FUCK the most rather than those towards the bottom.

Wholesale frequency comparisons aside, the LCA annotation scheme revealed an interesting difference in the use of FUCK between the two corpora (Table 1):

**Table 1. Annotation of FUCK in the Spoken BNC1994DS and Spoken BNC2014S using the LCA annotation scheme: most populated categories.**

| Spoken British National Corpus (1990s) | | |
|---|---|---|
| **Rank** | **Code** | **Description** |
| 1 | E | Emphatic adverb/adjective: 'He fucking did it' 'in the fucking car' |
| 2 | N | Premodifying intensifying negative adjective: 'the fucking idiot' |
| 3 | G | General expletive '(Oh) Fuck!' |
| 4 | I | Idiomatic 'set phrase': 'fuck all' 'give a fuck' |
| 5 | D | Destinational usage: 'Fuck off!' 'He fucked off' |
| Spoken British National Corpus (2010s) | | |
| **Rank** | **Code** | **Description** |
| 1 | I | Idiomatic 'set phrase': 'fuck all' 'give a fuck' |
| 2 | G | General expletive '(Oh) Fuck!' |
| 3 | F | Figurative extension of literal meaning: 'to fuck about' |
| 4 | D | Destinational usage: 'Fuck off!' 'He fucked off' |
| 5 | A | Predicative negative adjective: 'Is it fucked?' |

There appears to have been a shift in the way in which FUCK is used in British English conversation. The direction of change appears to be towards further generalization (and perhaps weakening) of this BLW, with idioms such as *give a fuck*, *what the fuck* and *for fuck's sake* occurring very frequently and accounting for 31% of all instances. Furthermore there has been a rise in figurative extensions of the original meaning, including *fuck it up*, *fuck me off* and *fuck around* (14% of all instances). In turn, emphatic adverb/adjective forms of *fuck*, which did account for 55% of instances in the Spoken BNC1994DS, now only account for 2.7%. It seems then that FUCK has moved away from a modifying function which emphasizes other lexical words, and occurs much more idiomatically and figuratively than previously.

## 6. Conclusion

The design of the speaker metadata categories in the Spoken BNC2014 makes the new data comparable to the Spoken BNC1994 for the purposes of sociolinguistic analysis, and the case of FUCK as an example (along with others to be given in the presentation) suggests a clear change in use between the 1990s and 2010s, in addition to wholesale change and stasis of BLWs in the two corpora.

The compilation of the Spoken BNC2014 has facilitated large-scale diachronic analyses of spoken data on a scale which has until now not been possible. This study therefore exemplifies new opportunities – and challenges – in the sociolinguistic study of spoken data.

## Acknowledgements

The research presented in this paper was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1.

## References

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380-409.

Ipsos MORI. (2016). *Attitudes to potentially offensive language and gestures on TV and radio: Quick reference guide*. London: Ipsos MORI Social Research Institute. Retrieved 14 December 2016 from Ofcom: https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/offensive-language-2016

Leech, G. (1993). 100 million words of English. English Today, 9-15. doi:10.1017/S0266078400006854

Ljung, M. (2011). *Swearing: a cross-cultural linguistic study*. New York: Palgrave Macmillan.

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017 fc). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22:3.

Lutsky, U. & Kehoe, A. (2015). *Your blog is (the) shit:* A corpus linguistic approach to the identification of swearing in computer mediated communication. *International Journal of Corpus Linguistics*, 21:2, 165-191. doi:10.1075/ijcl.21.2.02lut

McEnery, T. (2005). *Swearing in English: Bad language, purity and power from 1586 to the present*. New York: Routledge.

McEnery, A., Baker, J.P. & Hardie, A. (1999) Assessing claims about language use with corpus data – swearing and abuse. In Kirk, J.M. (ed.) *Corpora Galore: Papers from ICAME 1998.* Amsterdam: Rodopi, 45-55.

McEnery, T. and Baker, P. & Hardie, A. (2000). *Swearing and abuse in modern British English.* In: PALC '99: Practical Applications in Language Corpora. Peter Lang, Frankfurt am Main, 37-48.

Strawson, H. (2017). Diary: The British National Corpus. *London Review of Books*, 39:6. Available at: https://www.lrb.co.uk/v39/n06/harry-strawson/diary.