Swearing and abuse in modern British English

TONY McENERY,
PAUL BAKER and ANDREW HARDIE

1. Introduction

Recent work undertaken at UCREL in Lancaster has started to focus on the possible exploitation of corpora of spoken language, in particular for the examination of how swear words are used in modern British English (MBE). The impetus for this work has been partly motivated by the perception that little work has been undertaken on what, we are sure most people would agree, is an everyday part of language use. In addition to that, however, we also contest that to date such a study has been difficult to undertake – it is only now that corpora of spoken language, reflecting spontaneous language use by speakers of both genders, in a range of ages and social classes, has become available. Our work began by reviewing what analyses there have been of swearing to date (McEnery, Baker and Hardie 1999). On the basis of that review, we began a project to construct a problem oriented corpus, within which swear words[1] were annotated with an encoding scheme which encoded a set of analytical categories of swearing, plus information related to the target of the swear word and the producer of the swear word. This paper will carry the work of McEnery, Baker and Hardie (1999) forward, and examine the development of the original system of analysis, examining the way in which the original system underspecified the analysis of swearing, and introducing the final system of analysis used in our corpus. After this, we will progress to two analyses of swearing/abuse from the corpus. One analysis, of terms of abuse related to sexuality, examines how corpora may still be of use even where the weight of evidence coming from the

[1] Note that when we use the word swear word in this paper, we are using a broader sense of the word than is currently accepted. For us, swear word covers both words which would be generally viewed as being such, e.g. fuck, piss, shit as well as general terms of abuse, e.g. nigger, queer, paki. There is some warrant for doing this, as Hughes (1998: 7) discusses this broader meaning of the word. However, we are doing it here largely to avoid the repetitive use of the phrase 'swear words and terms of abuse' where 'swear word' will suffice.

corpus is slight. The second analysis, of the word *fuck*, shows how an extensive analysis may be made where a large number of examples are available.

2. The Lancaster Corpus of Abuse

The Lancaster Corpus of Abuse (LCA) has been under development at Lancaster University since January 1998. Supported by funds from the Faculty of Social Sciences at Lancaster, we have produced three versions of our corpus, each of which has successively embraced more words, and been provided with a more comprehensive and fine grained analysis.

The first version of the corpus, LCA 1.0, looked at a limited set of words[2] in the BNC spoken corpus, and encoded them according to a minimal set of annotations, derived largely from Hughes (1998: 31). The system of analysis will be returned to shortly. LCA 2.0, followed up on the work of the first version of the corpus by encoding a wider range of words, including terms of racial abuse, and words related to sexuality. Additionally, all of the variants of word forms already included within LCA 1.0 were included in 2.0. The focus was again upon spoken language, but this time spoken data from the Bank of English[3] was used to supplement the data available for certain words which were very infrequent, such as *puff*. At the time of writing, the third major version of the LCA is being produced, which encodes analyses from the written BNC for words present in LCA 2.0, with the aim of contrasting the use of swear words in written and spoken English.

The area of development of the LCA we would like to focus on here is the evolution of the annotations added to the swear words in the corpus. Each swear word is given a lengthy annotation which marks a series of linguistic and demographic features. LCA 1.0, as has been mentioned, used a system of classifying swear words derived from Hughes (*ibid.*). The basic categories of analysis are presented in Table 1.

The category information was coded with other data to create the full annotation of each swear word. Table 2 shows what information was encoded on each swear word.

[2] These words are were *prick, cunt, pillock, shit, moron, twat, arsehole, fart, idiot, prat, cow, bitch, swine, pig, bastard, bugger, sod, tit, turd, imbecile, cretin, sow, fucker.*

[3] The authors gratefully acknowledge the help of Geoff Barnbrook and Oliver Mason of Birmingham University in accessing the Bank of English.

Category	Annotation label	Description	Example from BNC
Personal	P	A second person insult of the form "You X" or similar	Yeah but Jonesy ain't here **you cunt** so it's only me.
Personal by reference	R	A third person insult of the form "The X" or similar	fucking serves **the cunt** right as well
Destinational	D	Swear word followed typically by *off*	Oh Jake **sod off**.
Cursing	C	An insult with a missing subject of the form "X you" or similar.	Can't even come and say hello, so **bugger him**!
General expletive of anger, frustration or annoyance	G	An imprecation with no particular target.	**Oh shit**, I haven't bought any scissors!
Explicit expletive of anger, frustration or annoyance	E	An imprecation with a specific target of the form "X it!" or similar	Liz got quite cross you know she's quite oh **bugger it**

Table 1. Categories of insult in LCA 1.0

Field	Feature marked	Possible values
1	Gender of speaker	M = male, F= female X = unknown
2	Social class of speaker	As per social class categories of BNC (see Aston and Burnard 1998)
3	Age of speaker	As per age categories of BNC (see Aston and Burnard 1998).
4	Category of insult	As per table 1 above
5	Gender of hearer	As per gender of speaker
6	Person of target	1 = first person, 2 = second person, 3 = third person, X = unknown
7	Metalinguistic usage	0 = no, 1 = yes
8	Animacy of target	+ = animate, - = non-animate, X = unknown
9	Gender of target	As per gender of speaker
10	Number of target	1 = singular, 2 = plural, X = unknown
11	Quotation	Q = quotation, N = non-quotation, X = unknown

Table 2. Categories of annotation

We soon found, however, that categories of insult only accounted for a sub set of the data. We added verbal use (**V**) to our analysis early to mark out the use of a swear word as a verb. Even so, we found that there was still a large rump of cases where our system could not classify a particular example of swearing. Consequently, for LCA 1.0 we developed a 'dustbin' category of **F** for words which defied classification. To give an idea of the coverage of the initial system of analysis, in LCA 1.0 34% of analyses were marked as F – in terms of coverage the original annotation scheme was leaving around a third of the data unaccounted for.

For LCA 2.0 we decided to revise the categorisation scheme. The obvious starting point was to examine our current coding, and then to examine those words categorised as F in LCA 1.0, and to try to extend our categorisation to cover them. Following is a discussion of how the mark-up scheme of LCA 1.0 changed as a result of that analysis.

2.1. Categorising swear words

Code	Description
A	Predicative negative adjective: "the film is shit"
B	Adverbial booster: "Fucking marvellous" "Fucking awful"
C	Cursing Expletive: "Fuck You!/Me!/Him!/It!" (retained LCA 1.0 category)
D	Destinational usage: "Fuck off!" "He fucked off" (retained LCA 1.0 category)
E	Emphatic adverb/adjective: "He fucking did it" "in the fucking car"
F	Figurative extension of literal meaning: "to fuck about"
G	General expletive "(Oh) Fuck!" (retained LCA 1.0 category)
I	Idiomatic 'set phrase': "fuck all" "give a fuck"
L	Literal usage denoting taboo referent: "We fucked"
M	Imagery based on literal meaning: "kick shit out of"
N	Premodifying negative adjective: "the fucking idiot"
O	'Pronominal' form with undefined referent: "got shit to do"
P	Personal insult referring to defined entity: "You fuck!" / "That fuck" (retained LCA 1.0 category)
R	'Reclaimed' usage – no negative intent
T	Religious oath used for emphasis: "by God"
X	Unclassifiable due to insufficient context

Table 3. LCA 2.0 analysis

On reviewing the original coding, we collapsed the original P and R categories together into P – the only distinction between them was of person (third versus first/second person). As this information was already encoded in the annotation, having separate categories based on person alone was wasteful. Similarly, the original C and E categories were collapsed into C alone for the same reason. Finally, the V category was dropped, as the part of speech information encoded on the word gave us examples of where the word acted as a verb.

We then moved on the classify the words which had received F as a coding in LCA 1.0. This led to eleven new categories. For convenience, the LCA 1.0 category of F was renamed X. The new categories are clearly shown in table three. Some of the new categories revealed interesting facts about swear words, for example:

- their use in idiomatic and figurative expressions is notable (as will be seen clearly in the later analysis of *fuck*).
- they have a role as providing positive and negative 'boosts' to the meaning of words, acting as gradable adverbs and emphatic adjectives/adverbs
- they can be used as referring expressions, e.g. taking the place of demonstratives.

The reclassification of the swear words in LCA 2.0 led to a greatly improved coverage of the classification – in LCA 2.0, the X category was only assigned to 101 words from a total of 8,011[4] analyses, a mere 1.3% of the cases considered.

It is clear that the current system of classification is quite comprehensive. Note that we are not claiming that the system is problem free – McEnery, Baker and Hardie (1999) discuss limitations of this corpus based approach with regard to LCA 1.0 which are just as relevant for later versions of the corpus. Also, the categorisation as it stands does not take intention into account – it may well be that some of the words in the corpus were actually aimed at their target to show affection and group solidarity. Not all uses of swearing are abusive. Our corpus does not encode such intentions, largely for want of information as our earlier paper outlines. Nonetheless, the data we have can be used to address a series of research questions as the following sections will show. In the next section we will examine how the corpus can be a useful starting point in the analysis of the use of a swear word, even where there are few examples available in the corpus. In the following section we will examine how the full annotation can be used where plentiful examples exist.

[4] This figure is for the words of the LCA 1.0 as analysed in the LCA 2.0, to make the comparison of coverage fair.

3. Using a corpus where few examples are available: sexuality abuse

Some of the word forms we wish to investigate have a relatively low frequency in the LCA; words related to sexual orientation are good examples of these cases. The possible reasons for this paucity of data are interesting to consider, and although small, the data may give suggestions which may be followed up by other forms of analysis.

For example, the aggregate number of examples of the words *gay, queer* and *poof/poofter* (when used to denote homosexuality) comes to only 39 cases. But even on this small scale the data is interesting. *Gay* occurs 24 times, and collocates with the following words: *is* (10), *he's* (9), *you're* (2), *Dad's* (1), *who's* (1). Its use is attributional in nearly all (21) cases, suggesting a strong colligation with an *X is gay* pattern, a notably 3rd person construct. The subject in this pattern is almost always male: *he's* (9), *he* (3), *chap* (1), *dad's* (1), *Mick* (1), *James* (1), *Male* (1), *Pat* (1), *Phil* (1), *sons* (1). Interestingly there are no first person attributions of being gay. Also, the examples appear to be simple reports of fact rather than an insult (though there is one case of insult, *Your Dad's gay!* which appears as part of an extended piece of verbal feuding).

Queer only has 7 examples,[5] 4 of which are clearly abusive, and three of which follow the attributional pattern of gay but with a twist – the attribution is negative i.e. *X is not queer*. Is it the case then, that we are abusive of that we claim we are not? Again speculation, but a possibly good starting point for further research using a different methodology, e.g. focus groups or questionnaires.

There are 6 examples of *poofter* and 2 of *poof*. These words always occur as singular common nouns and take the *P* (personal) form of abuse, but are not used in an attributive manner.

It should be noted that both the spoken corpus of the BNC and the Bank of English was gathered within what we may call a heterosexual (or at least nominally heterosexual) discourse community. The corpus was gathered from the general population, and therefore by weight of numbers alone overwhelmingly represents heterosexuals. It could be the case that the patterns of use of terms such as *queer, puff* and *gay* would shift if we gathered corpus data from within the gay community solely. Even the briefest of glances at the gay press shows that – within group – *queer* is a positive and comparatively frequent term in publications such as the *Gay Times*, *Pink Paper* and *Boyz*. Although it is a jump

[5] In the LCA there were only 3 examples of *queer* as abuse that were encoded with information as to gender, age and social class, but we were able to expand this number to 7 for the purposes of this exercise.

to assume that that the use of those words in the gay press would be reflected in the spoken language of gays, we want to emphasise here that corpus data – even in small quantities, can lead to the development of research questions which can be pursued by other means. The corpus can be the beginning of a journey of exploration, it need not always be its end point. Let us move, however, to the discussion of a word for which a substantial number of examples are available in the LCA 2.0 – *fuck*.

4. The "F" word

Some of the words we are looking at have a frequency high enough to allow us fully exploit the annotation on the corpus with some confidence. *Fuck*[6] is a good example, allowing us to carry out a user analysis and form corpus-based hypotheses. Table 4 shows word forms for *fuck*, with regard to the gender of the user. In all of the tables in this section, we have given the number of uses of "fuck" per 1000 utterances, rather than actual counts of words, as this should help to take into account the differences in size of the data when comparing across categories.

Form	Male	Female
Root	1.61	0.43
-ing	8.21	1.27
-ed	0.15	0.07
-er	0.06	0.01
-s	0.02	0.01
-ers	0.01	0
Total	10.06	1.79

Table 4. Male v. Female – word forms: occurrences per 1000 utterances

While quantity differs between gender, the ranking and proportions of the word forms remain fairly stable across gender. So while the use of this word may

[6] The reader should note that when we refer to *fuck* in this section, unless stated otherwise, we are considering the word and all of its morphological variants simultaneously, of which the total number of cases examined from the spoken section of the BNC was 1528.

differ quantitatively by gender, it does not differ qualitatively. Table 5 carries out the same analysis by social class:

Form	AB	C1	C2	DE
Root	1.5	0.12	0.84	1.03
-ing	2.9	0.67	5.58	5.51
-ed	0.29	0	0.07	0.06
-er	0.02	0	0.02	0.06
-s	0.04	0	0.02	0.008
-ers	0	0	0.02	0.008
Total	4.75	0.79	6.55	6.676

Table 5. Social class – word forms: occurrences per 1000 utterances

Here, as expected, the same pattern of word-form frequency is found, with *-ing* being the most common form, followed by the root. As regards social class, those from classes DE and C2 are the most frequent users, followed by AB. Interestingly, social class C1 has a much lower rate of *fuck* than the other groups. For age (see table 6 below) it is clear that the 16–25 group uses *fuck* most often, followed by the 26–35 group. It is difficult to determine whether the results show that people on the whole tend to use *fuck* more as young adults and increasingly less so as they age, or whether it is only the current crop of 16–25 year-olds who use *fuck* in this way. The observer effect of the tape-recorder may have influenced the results in that younger people (especially those under 15) might swear more (for the covert prestige of appearing more adult, perhaps) and ironically, older people (especially those over 36) may swear less. A further hypothesis is that persons aged between 25–45 are most likely to have their own children/teenagers around them, and thus swear less than those who are yet to have children.

Form	under 15	16–25	26–35	36–45	46–60	over 60
Root	2.16	2.53	1.04	0.05	0.73	0.02
-ing	3.55	13.30	7.68	0.65	2.69	0.13
-ed	0.23	0.45	0.12	0	0	0
-er	0.05	0.19	0.01	0	0.02	0
-s	0.08	0	0.01	0	0.02	0
-ers	0	0.03	0	0	0.02	0
Total	6.07	16.5	8.86	0.7	3.48	0.15

Table 6. Age – word forms: occurrences per 1000 utterances

We will avoid the presentation of results here where demographic factors are combined together, as one shortcoming of the LCA is that, for certain age/gender/social class combinations, there is no data at all.[7] This is true of the combinations *Male C2 Under 15* and *Male C1 15–24*, for example. We are currently examining a range of statistics which may allow us to make comparisons across such unbalanced data, but are also examining the option of further data collection to fill the gaps we have identified.

Examining the different uses of "fuck" and its related word forms (see table 7 below) it is clear that the most common use is as an emphatic adverbial or adjective – a category which was entirely absent from the category scheme used for LCA 1.0. Similarly, its usage in set phrases is quite high, another category not covered by the early scheme.

Code	Description	Male	Female	Total
G	General expletive "(Oh) Fuck!"	0.44	0.09	0.53
P	Personal insult referring to defined entity: "You fuck!" / "That fuck"	0.09	0.01	0.1
C	Cursing Expletive: "Fuck You!/Me!/Him!/It	0.27	0.10	0.37
D	Destinational usage: "Fuck off!" "He fucked off"	0.37	0.10	0.47
L	Literal usage denoting taboo referent: "We fucked"	0.07	0.04	0.011
B	Adverbial booster: "Fucking marvellous" "Fucking awful"	0.88	0.11	0.099
E	Emphatic adverb/adjective: "He fucking did it" "in the fucking car"	5.73	0.87	6.6
N	Premodifying negative adjective: "the fucking idiot"	1.33	0.28	1.61
A	Predicative negative adjective: "the film is shit" (*does not occur for* fuck)	0	0	0
O	'Pronominal' form with undefined referent: "got shit to do"	0.02	0	0.02
M	Imagery based on literal meaning: "kick shit out of" (*does not occur for* fuck)	0	0	0
F	Figurative extension of literal meaning: "to fuck about"	0.17	0.07	0.24
I	Idiomatic 'set phrase': "fuck all" "give a fuck"	0.42	0.09	0.51
R	'Reclaimed' usage – no negative intent (*does not occur for* fuck)	0	0	0
T	Religious oath used for emphasis: "by God" (*does not occur for* fuck)	0	0	0
X	Unclassifiable due to insufficient context	0.27	0.03	0.3
	Total	10.06	1.79	11.85

Table 7. Uses of *fuck* and related word forms per 1000 utterances

[7] When social class, age or gender are considered alone, the BNC is balanced, with fairly equal contributions made by males and females etc. However, when these categories are combined the resulting picture is an unbalanced one, with certain categories over-represented and others not represented at all.

An interesting difference occurs when we look at cases where *fuck* is used in cases of reported speech (table 8), the relative proportion of reported uses are much higher for females; roughly one in six as opposed to 1 in 36.

Quotation	Dataset	Male	Female
Yes	84	33	51
No	1418	1152	266
Unknown	26	25	1

Table 8. Swearing in reported speech (counts)

An examination of the targets of swearing (table 9) reveals that when females use *fuck* towards another person, they tend to aim it more towards other females, whereas males tend to aim it towards other males. Although, in most cases the target is not another person, when we do use *fuck* towards someone, we target our own gender most frequently.

	Speaker		
Target	All	Male	Female
Male	343	300	43
Female	113	40	73
0	777	615	162
Unknown	295	255	40

Table 9. Targets of swearing (counts)

Although with *fuck* we have much more data than we do for, say *gay*, the results in some ways are surprisingly similar. Although we have sufficient examples of *fuck* to examine each of the elements of its annotation separately, we again make interesting observations the explanations to which must lie beyond the corpus. Can the corpus explain why we target our own gender most when using the word *fuck*? Can the corpus explain why women appear to use *fuck* in reported speech more than men? The corpus is clearly of use in revealing facts about language use which may not be easily available to our intuition. However, in terms of explanation, the corpus as it stands appears to direct the path of future research rather than provide explanations. That is not to say corpora are useless for linguistic research. On the contrary, corpora such as the LCA can provide us with a clear view of a range of features of swearing, a very valuable result in itself. This can serve as a stimulus for hypothesis formation and establishes a set of empirically attested observations which any explanation of language use must account for. Yet as a sole path to the explanation of such language use – a self contained theoretical system – corpora may be judged and found wanting.

5. Conclusions

The significance of this study and its findings are threefold. First, in itself, it is another piece in the jigsaw of our understanding of contemporary spoken English. Second, and more importantly, it shows how the corpus methodology can be used to investigate swearing (and, by implication, other areas which have hitherto only been studied via intuition). Third, it exemplifies how investigating a single aspect of spoken language can cast light on variation in language across society. It would not be unproductive to consider the above findings in relation to what is known and/or theorised about the use of language by different social groups. Swearing is often a shibboleth (in the broader sense of the term), whose presence or absence marks group membership. Our knowledge of the social operations of language is founded on the study of such shibboleths.

This is not to say that a corpus-based investigation of, say, *fuck* can tell us everything we need to know. The obvious first step in a lengthier investigation would be to repeat the process with other swearwords. This would tell us first of all whether or not there is any such thing as a 'language of swearing' or whether the patterning of swearwords is largely random.[8] Particularly interesting would be a study of words across a range of 'taboo-ness' – for example, would *bloody*, which fulfils many of the functions of *fucking* but is generally recognised as milder, display the same demographic distribution as *fuck* or not? Would a more dissimilar word, such as *shit* or *bastard*?

A reliable study of the audience of swearing – i.e. in whose presence swearwords may/may not be spoken – would also be advantageous. We could thereby test the hypothesis that adults with children of their own are less likely to swear than any other group. It could be predicted that in the presence of children, swearing would be much less frequent. Such a study would also be interesting from the angle of group language – under what circumstances would members of a social group avoid swearing in front of someone from outside that group, or go out of their way to swear?[9] A study of this nature would necessarily be empirical, but would probably not be corpus-based.

---

[8] To repeat the study on *fuck* with a different, and wholly separate sample, would also be useful, to examine the reliability of this dataset.

[9] Something of this nature may have been triggered by the 'audience' of the tape recorders during the collection of the BNC and BOE data.

Finally, work in hand on the LCA 3.0 will extend our observations to written English. This will not only be for the sake of studying swearing but also as part of a broader examination of the differences between spoken and written English.

Our research is on-going, and future papers will address many of the research questions raised in this paper. For the moment, it is reasonable to conclude that the first two versions of the LCA have revealed much about language – and the use of corpora in studying language – which was hitherto unexpected.

References

Hughes, G. (1991). [2nd edition 1998]. *Swearing; a Social History of Foul Language, Oaths and Profanity in English* (second edition). London: Penguin.

McEnery, A., Baker, P. and A. Hardie (1999). "Assessing claims about language use with copus data: Swearing and abuse". *Corpora Galore. Papers from ICAME19–98.*