Orsman, H. (ed.) (1997), *The Dictionary of New Zealand English.* Auckland: Oxford University Press.

Trudgill, P. and J. Hannah (1994), *International English: a Guide to the Varieties of Standard English.* Third Edition. London: Arnold.

Assessing Claims about Language Use with Corpus Data – Swearing and Abuse

Anthony McEnery, John Paul Baker and *Andrew Hardie*

University of Lancaster

1. Introduction

Swearing[1] is an everyday part of the language of many speakers of modern British English (MBE). However, while this may be true, detailed studies which have tried to outline the form and social function of swearing in MBE are very few and far between. In terms of a detailed, corpus-informed, account of swearing, there is a near complete lack of work. Currently at Lancaster the spoken sections of the British National Corpus (BNC) and the Bank of English (BOE) are being marked-up with an annotation scheme designed to provide such a description. The resulting corpus is called the Lancaster Corpus of Abuse.

This paper will describe a specific study undertaken with an early version of this corpus designed to test a series of claims about swearing made by Hughes (1998). In doing so, we will develop a critique both of the system of analysis outlined by Hughes and his findings. In conclusion we will outline future research developments for our corpus.

2. Studies of swearing

No detailed corpus informed description of swearing exists which attempts both to describe swearing and outline its function. Graves (1927) is an entertaining read but wildly out of date. Classic though work such as Lakoff (1975) may be, it is based purely on introspective data and is at the very least out of date. Her claims about swearing have never been satisfactorily tested via a large scale study, though small scale studies, e.g. by Hudson (1992) and de Klerk (1992) have led to doubt being cast on Lakoff's claims. The descriptive work of such writers as Partridge (1947), Sagarin (1962) and Montagu (1973), is firmly rooted in another age, and has little relevance to MBE. A chapter on swearing in Andersson and Trudgill (1992) is brief, hypothesis heavy and bereft of any contact with naturally occurring language data or language description. Butler (1997) is a text heavily engaged with theoretical accounts of hate speech in order, once again, to produce a largely data free argument which is poor descriptively. Some work in the field of education exists, but it is focused on classroom rather than linguistics issues, e.g. Fox (1990).

At Lancaster we are currently exploiting large corpus resources in order to study swearing on the basis of attested language use. Studies touching upon swearing to date have very largely been non-corpus informed (though Stenström 1991 and Rayson, Leech and Hodges 1997 are notable exceptions). The most

recent major work on swearing is that of Hughes (1991, 1998). Hughes draws most of his examples from literary texts, and the most recent edition (1998) contains some confusing and misleading statements about the use of corpus data for the study of swearing. In addition, Hughes makes a variety of claims about swearing in MBE based upon intuition that bear assessment by corpus analysis. Our aim in examining Hughes' work is not specifically to criticise Hughes. Indeed Hughes' work served as the impetus for the creation of our corpus. Rather we want to show how corpora and linguistic hypotheses can interact via the use of Hughes' claims about swearing.

In this chapter we will introduce a pilot scheme for the analysis of swearing in English, which is compatible with that presented by Hughes (1998). We have used this scheme to annotate swearing occurring within our corpus, as outlined in the next two sections.

After having outlined the annotation we have undertaken, we will move on in section 6 to a detailed examination of Hughes' claims in the light of available corpus evidence.

3. The Lancaster Corpus of Abuse

The Lancaster Corpus of Abuse (LCA) is a problem-oriented corpus based upon data extracted from the BNC and the BOE, containing examples of swearing from transcribed spoken language.

Swearing has been annotated within the LCA using a scheme developed at Lancaster to encode a range of information relevant to the linguistic study of such terms (see section 4).

In deciding which words to include within the corpus, we have partly been guided by claims within the literature, partly by our own intuition and partly by words we encountered within the corpus. To give examples of each of these, the claims we are examining by Hughes (1998) in this chapter revolve around a sub-set of swear words which we used to generate the first version of the LCA (1.0).[2] Yet we knew, on the basis of our own knowledge of swearing, that they did not represent a complete set, and hence we expanded the wordlist for the second version of the LCA (2.0) on the basis of our intuition. Beyond this, however, when we examined the corpus data we would, from time to time, come across new coinages - sometimes entirely new words/phrases from the perspective of the investigators (e.g. *batty man*[3]) and sometimes examples of word play (e.g. *Cuntona* occurs in the BNC as a pun on *Cantona*, the surname of a French footballer).

The words covered by LCA 2.0 are too many to list here, but can broadly be grouped under the following main headings – 'words traditionally viewed as swear words' (e.g. *fuck, piss, shit*), 'animal terms of abuse' (e.g. *pig, cow, bitch*), 'sexist terms of abuse' (e.g. *bitch, whore, slut*), 'intellect-based terms of abuse' (e.g. *idiot, prat, imbecile*), 'racist terms of abuse' (e.g. *paki, nigger, chink*), and 'homophobic terms of abuse (e.g. *queer, puff, lezza*)[4]. Obviously, there is an interplay between these broad categories - many animal terms of abuse are also sexist abuse forms for example. However, for the purposes of describing the contents of the corpus, this broad classification will suffice. There are categories absent from the corpus at present, and in time we hope to expand the corpus at least to cover, for example, terms of abuse based upon disability. We also want to expand the corpus to cover written as well as spoken language. Even so, as it stands the LCA allows us to focus exclusively upon the relevant extracts of the spoken sections of the BNC and BOE relevant to the study of swearing.

The version of the LCA used in this paper is version 1.0, and consists of data extracted only from the spoken section of the BNC. All told, there are 1,301 separate examples in LCA 1.0, focused on a small subset of swearing in MBE (see note two). The corpus also contains a pilot annotation scheme; in order to retrieve information from the LCA in a speedy and systematic way we needed to develop an annotation scheme geared towards the study of swearing. In the following section the development of this system for LCA 1.0 is outlined.

4. The development of a corpus annotation scheme to encode swearing

The start point for our scheme was a set of categories of insult outlined in Hughes (1998). Hughes (1998: 31) identifies eight categories of swearing, six of which we decided to use as an initial set for our analysis.[5] The advantage of doing this was that we could encode the data in a way which would allow it to be used directly to assess claims Hughes makes on the basis of his categorisation. The disadvantage, as will become apparent, is that the system is far from complete,[6] and as it stands the distinction between the categories is far from water-tight. Table 1 below outlines Hughes categories of swearing. Each category is exemplified from the spoken section of the BNC.

Each word in LCA 1.0 was marked-up by hand as belonging to one of these categories; during this process of hand-annotation we encoded further information.[7] As we were drawing data from the spoken section of the BNC, we could mark-up each word to denote the age, social class and gender of the speaker. When we moved on to use spoken data from the BOE, although we could not mark-up age and social class, we were still able to mark-up speaker gender. In addition to this ethnographic data for speaker, we also decided to mark-up at least the gender of the hearer. This proved difficult, as will be explained in the next section. What proved easier to encode and is certainly useful was the *gender* of the target of the swear word. Such information is often present because the target is addressed via a gender marked pronoun. Accordingly we undertook this analysis. The person and number of the target of the swearword was easily encoded also.

Three further forms of information were also included in the analysis as we worked through the data. Firstly, the animacy of the target was added to the system of analysis when a member of the team hypothesised a link between animacy and the use of certain terms of abuse.[8] Secondly, there are times in the corpus when swear words are discussed - there are metalinguistic discussions of swearing in the corpus, and at times the appearance of swear words depends upon this form of discussion. Hence we decided to mark any examples of swear words being discussed in this way. Finally, we also decided to encode whether or not the

Table 1: Hughes' categories of swearing

Category	Label	Description	Example
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 2: The annotation scheme

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, except X may denote a mixed sex target group also
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 hearer
10	Number of target	1 = singular, 2 = plural, X = unknown
11	Quotation	Q = quotation, N = non-quotation, X = unknown

swear word was part of reported speech rather than attributed directly to the speaker. Table 2 summarises our initial system of analysis. The following are examples of encodings:

*You rotten bastard*_fde0PX20+F1N !
*He's not funny at all, he's a bastard*_fde0RX30+M1Q

5. Problems encountered annotating the corpus

We discovered that the system of categories developed by Hughes becomes slightly redundant in the context of the full annotation scheme we developed. For example, *P* and *R* insults are only really differentiated on the grounds of the person of the target. As we were encoding this anyway, the distinction became useless and was abandoned in version 2.0 of the LCA.

Additionally, the system covered only a subset of the cases that we encountered. For example, there are cases in the corpus of people aiming abuse at themselves, as in the following examples from LCA 1.0 (all taken from the BNC):

I know I'm a real bitch saying all this.
(Female, social class D/E, aged 26-35)
Makes me look a right cunt dragging me out of the shop
(Male, social class C2, aged 26-35)
Erm cos I'm a right bastard really I think!
(Male, social class D/E, aged 36-45)

Neither the *P* nor *R* category of Hughes seemed to covered these.

Additionally, there were numerous cases of swear words being used to describe physical objects and acts - sometimes *shit* can refer to faeces, sometimes *fuck* can refer to the act of copulation. The other holes in the system of analysis are too numerous to mention - to exemplify this, of the first 1,301 examples encoded in LCA 1.0, we were left with a rump of 182 cases which defied analysis. Work on LCA 2.0 has focused on the elaboration and rationalisation of the categories of analysis to produce a set of categories which cover all of the cases encountered. However, for the purposes of the comparison of Hughes' claims against corpus data in section 6 we will be using LCA 1.0 as it matches his system of analysis.

Methodological problems also arose as we examined the data. The first, and possibly deepest problem, is the difficulty of encoding information related to hearer. The BNC and the BOE are encoded expressly from the speaker's point of view. Indeed, the authors cannot think of any corpus which is encoded from the hearer's perspective. The upshot of this is that it is almost impossible to extract reliable information relevant to the hearer of a swear word. Hence the question of how interaction with various types of hearer influences swearing is difficult to gauge[9]. Of the 1,301 examples cited previously, the gender of the hearer cannot

be determined in 1,061 cases. This is a problem, because examples are readily available in the corpus which can demonstrate the influence of a hearer on swearing.

Take the following example from the BNC, where the linguistic behaviour of *B* and *C* (males aged below 15) shifts significantly when *B*'s mother (*A*) leaves the room:

A: Alright then.
B: Okay.
A: Wha what paper do you want? The Sunday Times?
B: Okay.
A: Right.
B: See ya Mum.
A: I'll be back in a minute.
B: Alright.
A: Bye.
B: Now to, for some fucking dirty swear! Wo oh oh oh! You fucking bitch! You Irish bastard! Aidan and Mandy have it in bed! Wo oh! Bed squeaking! Ah ha, ah ha, ah ha, ah ah! Fucking slag! Dirty whore! Piss off you Irish slag.
C: Yeah, I'll fucking shag her! For a pint of fucking bitter! Ya pakis! And we hate Holland, the Dutch bastards! Ah ah!

Sadly, the current focus of spoken corpora on the speaker ensures that while we can discover individual examples which clearly show the importance of the hearer, it is impossible to examine the role of the hearer on a wide scale, as it is well nigh impossible to recover data from a corpus on a hearer oriented basis.

Another methodological problem facing those wishing to construct a corpus such as the LCA is the nature of swearing. As can be seen from the previous example, the behaviour is sensitive to being observed. Speakers who swear may suppress their swearing depending upon who is observing them. All of the donors of spoken material to the BNC and BOE knew their voices were being tape recorded; we can reasonably surmise that some swearing/abusive behaviour was repressed accordingly. Indeed, the corpus yields clear evidence that those being taped were aware that they were being recorded and were sensitive to having their swearing recorded, as the following examples show (the second example is particularly noteworthy, as the sequence is instigated by a discussion of the taping of the conversation for the BNC[10]):

(1) A: Turn it off!
B: Tu turn it off? Why? Are you frightened you will er swear?

(2) A: Oh shit I mustn't swear tonight.
B: No not allowed to swear tonight. Definitely not allowed to swear tonight so don't do it.

People are conscious of the taboo status of swearing language, hence we cannot believe that the swearing present in the BOE and BNC are representative of swearing in everyday English in terms of the quantity of usage. However, we see no reason to believe that the patterns of usage for individual words are affected by this observer effect.

6. Using the corpus to assess claims

In this section we will use the results from our preliminary construction of the LCA to test some claims made by Hughes regarding swearing.

6.1 The first claim: the distribution of swear words within categories

Hughes (1998:31) sets out which swear words he believes occur with respect to the categorisation of swearing he set up outlined in Table 1. Reproduced below is a summary of the claims made by Hughes for a range of words included within the LCA:

Table 3: Distribution of swearwords within categories (according to Hughes)

Word	Category Label					
	P	R	D	C	G	E
Cunt	X	X				
Shit	X	X			X	
Fart	X	X				
Bugger	X	X	X	X	X	X
Bastard	X	X				
Arsehole	X	X				

Table 4: Distribution of swearwords within categories (according to LCA)

Word	Category Label					
	P	R	D	C	G	E
Cunt	X	X				
Shit	X	X			X	
Fart	X	X				
Bugger	X	X	X	X	X	X
Bastard	X	X			X	
Arsehole		X			X	

When we check which of these words are associated with which functions in the LCA we find that, while substantially accurate, the picture painted by Hughes is not complete.

Both *Bastard!* And *Arseholes!* have a life as *Gs* within the LCA, and there are no attested examples of someone being called an *arsehole* as a *P*. The last example in particular is interesting as it relates to claims of falsifiability and corpus data. Just because *arsehole* does not occur as a *P*, we cannot conclude that it has lost that function. Indeed, if we did we think most native speakers would disagree with us. On the other hand, we are able to falsify Hughes' claim that *bastard* and *arseholes* cannot have the function *G* – they can. Even though we have only found 3 examples of *bastard* as *G*, and one of *arseholes* as *G*, this is sufficient to falsify the claim that neither of these are possible. While we may want to scrutinise these examples of *bastard* and *arseholes* further, as they stand they are good examples of how corpus data can be used to falsify claims about language usage.

6.2 The second claim: the acceptability of gendered targets for specific swearwords

Hughes (1998: 208) makes another range of claims, this time related to what the gender of the target of vehement personal abuse may be. The claims are summarised in Table 5.

Table 5: Distribution of swearwords between genders (according to Hughes)

Male targets only	*Prick, cunt, twat, pillock, tit, arsehole, shit, turd fart, idiot, imbecile, moron, cretin, prat, swine, pig*
Female targets only	*Cow, bitch, sow, fucker*

When we check this against the LCA, looking at *P*, *R*, *C* and *E* categorisations, the results are quite marked - the above claims are, by and large, false. Table 6 summarises the findings from the LCA, with words emboldened to represent an accurate prediction on behalf of Hughes:

Table 6: Distribution of swearwords among genders (according to LCA)

Male targets only	***Prick, cunt, pillock, shit, moron,***
Female targets only	
Targets may be of either sex	*Twat, arsehole, fart, idiot, prat, cow, bitch, swine, pig, bastard, bugger, sod*
No example of the word found as personal insult in the LCA	*Tit, turd, imbecile, cretin, sow, fucker*

These findings show that even terms which have been traditionally associated with sexist abuse (e.g. cow, bitch) can be applied to males. It seems that the most gender exclusive terms of abuse in Hughes' set are *prick, cunt, pillock, shit* and *moron*, all of which apply to males only in the LCA. Drawing

wider conclusions at present is difficult. It would be nice to look at Table 6 and conclude that terms of abuse that were once exclusively used to lambaste females can now be used of either sex. However, we cannot. Firstly we lack concrete evidence that they ever were used in that way. Secondly, there is still a definite preference for words such as *bitch* to have a female rather than a male target - *bitch* is used 6 times of a male and 37 times of a female in the LCA.[11] It is most likely that gradience applies to the gender specificity of terms of abuse.

A recheck of the corpus data from the first test reveals that gradience also applies to the categories outlined by Hughes. While it may be the case, that *shit*, for example, may be used within three of the categories presented, the examples of the use of *shit* in the LCA do not spread out evenly of those categories – words can have major and minor functions. Table 7 is an example of this: the relative distribution of the word *shit* among the categories of LCA 1.0.[12]

Table 7: Distribution of *shit* among categories

Word	**Category Label**					
	P	**R**	**D**	**C**	**G**	**E**
Shit	5	10	0	0	80	0

Hence, while it is true that *shit* can be a *P*, *R* or *G*, it is quite clear that the most important function that shit has in MBE is to act as a general expletive of anger/annoyance/frustration. Gradience can also apply to category membership.

7. Conclusion

Although the construction of the LCA is on-going at Lancaster, the results we gained from version 1.0 of the LCA were promising. Using the corpus has enabled us to assess claims about swearing made by other authors, and has allowed us to refine a category based typology of swearing. Importantly, it has pointed to gradience in the application of swear words to specific genders, with gender exclusive terms of abuse being the exception rather than the norm in the data examined so far. LCA 1.0 proved useful in outlining methodological and organisational problems with the construction of a corpus of swearing, as well as showing that even at a pilot stage such a corpus could be useful in the examination of linguistic hypotheses. LCA 2.0 will allow us to expand that study yet further, with a wider range of swearwords covered, and a reorganised annotation scheme. In the near future, we hope to be in a position to undertake a widescale corpus based description of swearing. For now, however, we are able to use the LCA to examine individual claims about swearing, and come up with answers that challenge our intuitions on this topic.

Notes

1. We are conflating what is traditionally regarded as swearing (vulgar and sacrilegious language, e.g. *fuck, damn, bloody*) together with abuse (terms deemed offensive but not vulgar or sacrilegious in MBE, e.g. *nigger, puff, chink*) under the term swearing in this chapter.

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. *Batty man/batty boy* is a Jamaican English insult meaning 'homosexual', as in the following example from the BNC (speaker is aged is 0 – 15, gender female, social class AB) *Kerry is a batty man! He fancies Michael!* It can also be found in the Corpus of Written British Creole *An' dat goes for any man from Grange an' the rest ah dem batty bwai deh!* Thanks to Dr. Mark Sebba for this example.

4. Though still under construction, LCA 2.0 contains 18,523 examples of such terms of abuse.

5. We excluded the category *verbal use* from our analysis, as it relates directly only to infinitival uses of the verb form in structures such as *to prat about.* The *adjectival usage*, such as *bloody* and *fucking*, was excluded as we were initially not interested in adjectival uses. These verbal form and adjectival forms were excluded from LCA 1.0. Both exist, in modified forms, in LCA 2.0.

6. Even the eight code system of Hughes does not account for the full range of swearing and abuse.

7. Each analysis is encoded by one analyst, and checked by a second. Cases of disagreement are arbitrated by a third analyst, with majority voting in force.

8. A claim not explored in this chapter.

9. To take this observation further and start to think of hearers in more detailed terms, after Goffman (1976), would most certainly be impossible. In Goffman's terms, we are looking for the gender of ratified participants who are the direct addressee of the speaker.

10. *A* and *B* are used to denote the interlocutors in each example - these are different from the previous example and from each other.

11. As an example of *bitch* being used of males, take the following example where the speaker is a male, age range 26-35, social class D/E: *I'll give you two thousand I'd be a bitch and sell them!*

12. For completeness, within the LCA 1.0, *shit* occurs as a verb 7 times, refers to excreta 8 times and is not used in an adjectival manner at all.

References

Andersson, L and P. Trudgill, (1992), *Bad Language*. London: Penguin.
Butler, J. (1997), *Excitable Speech*. London: Routledge.
Fox, M. (1990), *The Social Bases of Swearing*. M.Ed. Dissertation, Nottingham University.
Goffman, E. (1976), 'Replies and responses', *Language in Society*, 5: 257-313.
Graves, R. (1927), *Lars Porsena or the Future of Swearing*. London: Kegan Paul, Trench and Trubner.
Hudson, E. (1992), *Swearing: a Linguistic Analysis*, unpublished MA Dissertation, Birkbeck College, London.
Hughes, G. (1991 [Second Edition 1998]), *Swearing; a Social History of Foul Language, Oaths and Profanity in English*. London: Blackwell.
De Klerk, V. (1992), 'How taboo are taboo words for girls?', *Language in Society*, 21: 277-289.
Lakoff, R. (1975), *Language and Woman's Place*. New York: Harper and Row.
Montagu, A. (1973 [originally published 1967]), *The Anatomy of Swearing*. London and New York: Macmillan and Collier.
Partridge, E. (1947), *Usage and Abusage*. London: Hamish Hamilton.
Rayson, P, Leech, G and M. Hodges, (1997), 'Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus', *International Journal of Corpus Linguistics*, 2: 133-152.
Sagarin, E. (1962), *The Anatomy of Dirty Words*. New York: Lyle-Stewart.
Stenström, A.-B. (1991), 'Expletives in the London-Lund corpus', in: K. Aijmer and B. Altenberg (eds.), *English Corpus Linguistics*. London: Longman, 239-253.