Your blog is (the) shit

A corpus linguistic approach to the identification of swearing in computer mediated communication

Ursula Lutzky and Andrew Kehoe
Vienna University of Economics & Business / Birmingham City University

The study of swearing has increased in the last decade, diversifying to include a wider range of data and methods of analysis. Nevertheless, certain types of data and specifically large corpora of computer mediated communication (CMC) have not been studied extensively. In this paper, we fill a gap in research by studying the use of swearwords in blog data, and illustrate ways of identifying swearing in a large corpus by taking context into account. This approach, based on the examination of shared and unique collocates of known expletives, facilitates the distinction of attestations of swearing from non-swearing in the case of polysemous lexemes, and the analysis of overlaps in usage and meaning of swearwords. This work therefore goes beyond basic sentiment analysis and offers new insights into the use of collocation for refining profanity filters, providing innovative perspectives on issues of growing importance as online interaction becomes more widespread.

Keywords: swearing, pragmatics, blogs, CMC, collocation

1. Introduction

Previous corpus linguistic analyses of swearing have tended to focus on transcribed speech, with several studies, from McEnery et al. (2000a) to Ljung (2009), using the spoken component of the British National Corpus (BNC). This is understandable, given that swearing is often seen as a feature of unplanned, spoken language – "an outlet for frustration and pent-up emotion and a means of releasing nervous energy after a sudden shock" (Crystal 1997: 61). Indeed, later work by McEnery & Xiao (2004) found expletives to be up to twenty times more frequent in the spoken component of the BNC than in the written component.

In this paper, we examine swearing in a written text format but one which has been shown to exhibit features traditionally associated with spoken discourse: the blog.

Our analysis is based on a 181 million word sub-corpus of our Birmingham Blog Corpus (http://www.webcorp.org.uk/blogs), which includes both blog posts and reader comments. The commenting feature on blog posts allows some degree of interaction between author and reader and, more frequently, between individual readers. This opens up new possibilities for pragmatic analysis in general and, in this study, for the analysis of swearing in particular. While there have been some corpus linguistic studies of impoliteness in Computer Mediated Communication, or CMC (see e.g. Angouri & Tseliga 2010, Haugh 2010, and Upadhyay 2010 in the special issue of the *Journal of Politeness Research* 6 or Hardaker 2010), most of them have focused on formats other than blogs and we intend to fill this research gap by focusing our analysis on this medium.

This study combines a corpus linguistic methodology with a pragmatic analysis. We demonstrate how a large corpus of blogs can be used in the study of swearing, which comprises a potentially infinite inventory of words and phrases and therefore cannot be searched for automatically. We do so by exploring the role context and collocation play in the detection of swearing and in gaining further insights about its use. Our approach is primarily descriptive but has the potential to suggest new solutions to practical, real-world issues encountered in the development and use of CMC platforms, including profanity filtering and the prevention of cyberbullying or "trolling" (e.g. Hardaker 2010). Much of the previous work in these areas has been based upon automatic sentiment analysis which, at its most basic level, involves labelling specific words as either "positive" or "negative" and then searching the corpus for particular concentrations of these words (see e.g. Mishne & Glance 2006). This is still an area, however, which is somewhat lacking in linguistic sophistication, and we present some of the limitations of automatic sentiment analysis in Section 4 before going on to demonstrate the role collocation can play in refining this approach when dealing with large data collections.

Our aims can therefore be summed up as follows: we aim to combine a pragmatic study with a corpus linguistic methodology, to study the use of "bad language" in blogs, an interactive and communicatively immediate text type, and to explore the role context and collocation can play in identifying and determining the use of swearing in online data. Our work thus offers new perspectives on crucial aspects of online interaction at a time when it is continuing to grow in importance and becoming more widespread.

2. Swearing

As McEnery (2006: 1) states, "[t]he use of bad language is a complex social phenomenon". This is reflected in the range of terms that are used to refer to bad language use. Beers Fägersten (2012: 3-5) gives an overview of such terms, including "bad words, curse words, cuss words, dirty words, four-letter words, expletives, epithets, obscenities, profanity, blasphemy, bawdy language, foul language, rude language, vulgar language, or taboo language" and points out that there is "a trend towards an interchangeability of terms". That is to say that these terms are used as synonyms or near-synonyms to denote bad language use. In addition to the variation in labels, there is disparity with regard to the constructions included in inventories of bad language use, which are more or less open-ended. This also relates to the fact that they are not restricted to individual words but may comprise more extensive constructions (see Andersson & Trudgill 1990: 58-59).

In this paper we use the term 'swearing' to refer to expressive uses of bad language. We follow Jay & Janschewitz (2008: 268) in defining swearing as "the use of taboo language with the purpose of expressing the speaker's emotional state and communicating that information to listeners". That is to say that we do not regard literal uses of taboo words as swearing (e.g. the word *shit* being used with reference to the excretory system).[1] On the contrary, we only regard the use of taboo words as swearing when they express emotions, which is the "main purpose of swearing" (Jay & Janschewitz 2008: 267, see also Ljung 2011). Depending on the context of use, swearing may convey negative feelings, for instance, of anger or frustration, or positive feelings of joy or excitement. Thus, the use of taboo words may act as a means of fostering group membership but it may also insult, offend or offer catharsis by allowing speakers to express pain (Mohr 2013: 13-14).

Jay & Janschewitz (2008: 269-270) distinguish between 'propositional' and 'nonpropositional' swearing. Propositional swearing is "consciously planned and intentional", whereas nonpropositional swearing is "unintentional, unplanned and uncontrollable" (Jay & Janschewitz 2008: 270). The latter category is not regarded as polite or impolite due to its automatic nature, triggered for example by sudden emotional outbursts such as surprise. Propositional swearing, on the other hand, can be polite, impolite or neither and it is this type of swearing that we expect to find predominantly in our data given that blogs are a written and therefore planned text type. In any case, the context in which swearing is produced is important and as Butler and Fitzgerald (2011:

527) note "[t]he use of such language demonstrates a speaker's understanding or treatment of an interaction as an informal and intimate one". It is the situational context and, as we will show below, the collocational environment that determine if a swearing expression carries a positive meaning and enhances social harmony (e.g. *fucking* as an emphatic intensifier in *fucking marvellous*) or has a negative connotation to the extent that it may be of a face threatening nature (e.g. *you fucking idiot*). Consequently, we regard swearwords as taboo expressions "which have the potential to be offensive" (Beers Fägersten 2012: 3) but which do not inherently carry impoliteness as a defining feature.

Previous research introduced categorisations of bad language based on different criteria. McEnery (2006: 30-33, see also McEnery et al. 2000a: 397), for instance, adopts six main headings: swearwords (e.g. *fuck*), animal terms of abuse (e.g. *cow*), sexist terms of abuse (e.g. *whore*), intellect-based terms of abuse (e.g. *idiot*), racist terms of abuse (e.g. *nigger*) and homophobic terms of abuse (e.g. *queer*). In addition to these broad categories, he adopts a more fine-grained categorisation according to the type of bad language use (a scheme originally developed for the Lancaster Corpus of Abuse, see McEnery et al. 2000a, 2000b), which to an extent relates to part of speech; examples include adverbial boosters (e.g. *fucking awful*), destinational usage (e.g. *fuck off!*), idiomatic set phrases (e.g. *give a fuck*), or religious oaths used for emphasis (e.g. *by God*). Ljung (2011: 29) introduces a classification of swearing based on the distinction between functions and themes, where the "functions are the uses that the swearing constructions are put to by the swearers, while the themes are the different taboo areas that the constructions draw on". The two main subgroups of Ljung's (2011) functions are stand-alones and slot fillers, depending on whether a swearword constitutes an utterance in its own right or forms part of a larger construction. These functions in turn comprise taboo words pertaining to one or several taboo themes, the five main ones of which are the religious/supernatural, scatological, sex organ, sexual activities, and mother (family) themes.

As stated above, the majority of studies on swearing in Present Day English are based on spoken English data, provided for example by the BNC (see e.g. Ljung 2009; McEnery et al. 2000a, 2000b; McEnery & Xiao 2004: Butler & Fitzgerald 2011). On the other hand, studies looking in particular at swearing in online data are more infrequent. An example is Thelwall (2008), who studies the social networking site MySpace for attestations of swearing. He bases his study on the demographic information provided on MySpace (the accuracy of which is questionable) to carry out a contrastive analysis of

US and UK data with the aim of uncovering similarities or differences in the distribution of swearing according to gender. His findings reveal that there "was no significant gender difference in the UK for strong swearing, especially for younger users (16-19)" (Thelwall 2008: 83), in contrast with the US where the incidence was higher for male users.

3. The Birmingham Blog Corpus and corpus pragmatic approach

Our study is based on the Birmingham Blog Corpus (BBC), a diachronically-structured collection covering the period 2000-2010 and totalling 630 million words. The corpus is searchable through the *WebCorp Linguist's Search Engine* (WebCorpLSE) software built by the Research and Development Unit for English Studies (RDUES) at http://www.webcorp.org.uk/blogs (Kehoe & Gee 2007). In this paper, we focus on a 181 million word sub-section of the corpus downloaded from the WordPress and Blogger hosting sites, which includes both blog posts and reader comments on these posts. This sub-section was built by downloading all posts from each of the blogs featured on the home pages of the hosting sites – known as "Blogs of Note" on Blogger and "Freshly Pressed" on WordPress – and then following links in these posts to other Blogger and WordPress blogs. The resulting corpus contains over 220,000 posts and over 2 million comments (see Kehoe & Gee 2012 for a fuller description of the corpus and the techniques used to create it).

Blogs have been defined as an online medium, rather than a genre (see e.g. boyd 2006, Herring et al. 2005, Kehoe & Gee 2012). They are an interactive medium in that they allow for "interaction-at-one-remove" (Nardi et al. 2004); that is to say that blogs are an asynchronous means of computer mediated communication that allows for interactivity on a potentially infinite time scale, with readers being able to comment on a post at any given point in time after its publication on the web. While blogs are a written type of text produced in the graphic code, they have been said to show features traditionally associated with spoken language and communicative immediacy (see Koch 1999). This communicative immediacy means that blogs often contain language innovations which may already have appeared in speech but have not yet found their way into more conventional written texts. For example, Renouf & Kehoe (2013: 182-183) find the new adjectival use of the word *genius* (as in *genius idea*) to be significantly more frequent in the BBC than in other written corpora. In doing so, they highlighted the fact that blogs are particularly suitable for diachronic analyses as they include more reliable

date information – for both posts and comments – than most other types of webpage (cf. Kehoe 2006 on the problem of extracting reliable dates from web texts more generally).

This study falls into the area of corpus pragmatics by combining the study of language use with a corpus linguistic methodology, a field of study that has gained increased attention in the last decade (see e.g. Aijmer & Rühlemann 2014, Jucker 2013, Romero-Trillo 2008). Corpus pragmatic studies are of an empirical nature and focus on types of data that represent naturally-occurring language use. While pragmatic analyses have traditionally been of a qualitative nature and focused on smaller data samples to illustrate certain phenomena, corpus linguistics typically builds on large data samples to ensure representativeness and is associated with quantitative analyses (see Archer et al. 2008: 614, Jucker et al. 2009: 3-4). Furthermore, corpus linguistics has been concerned mainly with studying specific linguistic forms (product, e.g. the form *sorry*), not least because forms can be searched for with linguistic software, whereas pragmatics has studied linguistic forms but also functions (process, e.g. the speech act of apology). One of the central questions that corpus-pragmaticists have therefore addressed in the last two decades is how the two can be combined.

Jucker (2013) mentions three main approaches to data analysis that have been taken in corpus pragmatics: form-to-function mapping, function-to-form mapping, and communicative expression or metadiscourse analysis. The present study straddles the first two approaches: it is interested in a specific linguistic function – swearing – and wants to gain further insight into the forms used to fulfil this function. On the other hand, it starts out from certain linguistic forms, as will be explained in more detail in the following section, to get a clearer idea of their functions in different collocational environments. By carrying out this analysis, we want to show how collocation can help differentiate between different functions of a form, so as to single out swearing uses and improve the precision of the search output. This is particularly relevant for corpora of a considerable size, such as the BBC, where it is not feasible to exclude unwanted hits manually.

4. Initial analysis

As outlined above, our aim was to extract examples of swearing from the BBC by adopting what could, in general terms, be classed as a lexical approach and, more specifically, as a collocational approach. This differentiates our work from studies that

have made use of automatic sentiment analysis which, while lexical in focus, does not typically take context of occurrence into account, except in very general terms.

One of the earliest studies of blog comments (Mishne & Glance 2006) makes extensive use of this technique, developing a 'disagreement lexicon' to detect disputes in comment threads. Mishne & Glance's (2006) study builds upon techniques developed by Nigam & Hurst (2004) in the market research field to analyse customers' opinions of products as expressed in online reviews: a field which has grown enormously in the decade since that paper was published. In their sentiment analysis of consumer product reviews, Nigam & Hurst (2004) class the words *blurry* and *crisp* as negative and positive respectively in the context of a digital camera review but they do not take immediate textual context into account. We see this as a significant limitation when analysing online data in general and blog data in particular, given that blogs have been shown to be at the forefront of linguistic innovation (cf. Renouf & Kehoe 2013). This is particularly true of the evaluative words, which form the backbone of automatic sentiment analysis. In addition to well-known examples such as *bad* and *wicked*, we find many other instances in the BBC where a word which may be classed as negative in conventional usage is actually used as a positive evaluator: *ill, sick, crazy, deadly, fierce, filthy, mean, ridiculous, rude,* etc. Consider Examples (1) to (4) below:

(1) He stops off Eastern Michigan University and hits em with an *ill* rhyme
(2) dude, get the GoChat app! its fuckin awesome, the layout is *sick* and it works perfectly
(3) Those cookies are so cool! They kind of look like spider webs - I'd like to try them in orange and black for Halloween. And that cake is *crazy*
(4) This shit is *ridiculous* [...] more please!

Examples (2) and (4) also contain what may, in a general sense, be classed as 'swearwords': *fuckin(g)* and *shit* respectively. However, in neither case is the swearword designed to convey a negative evaluation. The first is an example of what McEnery & Xiao (2004: 257) in their corpus linguistic analysis of *fuck* refer to as an 'emphatic intensifier' (*fucking marvellous* is their example). The use of *shit* in Example (4) is an instance of a less well studied phenomenon. One might assume that the word *shit*, when used as an evaluator, should always be labelled as negative. However, in our analysis of

blog data we find that this is not the case, as indicated by Examples (5) to (15) from the BBC:

(5) My job selling car insurance is *shit*. [negative]
(6) Your blog is *shit*. [negative]
(7) Your blog is the *shit*. I love it. [positive]
(8) That is the *shit* man, completely awesome. [positive]
(9) By the way, your outfit is the *shit* - fantastic. [positive]
(10) My Mom is the *shit* and here's why: [lists 6 reasons] I love my Mom. [positive]
(11) What's wrong with Pizza Hut? Some of their pizza is the *shit*! [positive]
(12) These bloggers really know how to shoot the *shit* as far as I'm concerned. [positive?]
(13) [...] on the internet where so many use their anonymity to harass, bully, and verbally kick the *shit* out of people. [negative?]
(14) I should really find myself a good therapist and get my *shit* together. [neutral]
(15) Drinking cocktails, flirting with handsome men and *shit*. [neutral]
(16) This is a dress you could wear all spring and summer, and probably in the fall too, with tights and *shit*. [neutral]

Examples (5) and (6) are indeed unambiguously negative, with the second potentially perceived as impolite. In Example (7), however, the addition of a single word – the definite article – reverses the meaning of the sentence: *the shit* is a positive evaluation, as reflected by the use of *love* in the sentence that follows, and given that the word *shit* here forms part of a compliment (see also Examples (8) and (9)), this example can be classed as a polite usage. Examples (8) to (11) are similarly positive and in most cases there is some other word in the same or adjacent sentence which indicates this: *awesome*, *fantastic*, *love* all reinforce the positive evaluation.[2] These are the kinds of 'contextual clue' (Renouf & Bauer 2001: 231) we are capturing when we carry out collocational analyses.

It is not the case, however, that all instances of the phrase *the shit* can automatically be classed as positive, as illustrated by Examples (12) and (13). These examples both include *the shit* as part of a longer idiomatic phrase, beginning with the words *shoot* and *kick* and meaning *chat*/*gossip* and *attack* respectively. Example (12) is vaguely positive and (13) is loosely negative but these examples illustrate the difficulties

faced by any automatic sentiment analysis system. Furthermore, Examples (14) to (16), like Example (4) in, illustrate a more general use of *shit* to mean "stuff", which is neither negative nor positive. This concept is referred to by Ljung (2011: 35) as replacive swearing, whereby a taboo word in its non-literal meaning replaces a non-taboo word.

It is clear, then, that automatic sentiment analysis based on the assignment of words to broad positive and negative categories is of limited benefit when dealing with textual data containing a large proportion of innovative language use. The examples given above illustrate that immediate context of use is vital in determining the specific meaning and illocutionary force of a word. In our study, therefore, context of use is of central importance. Our approach is to begin with a list of words that have been discussed with reference to swearing in other sources and then, through collocational analysis, examine potential overlaps and differences in the usage of these swearwords. This approach is based on the notion that the meaning of a word can be determined by its immediate textual environment or that "[y]ou shall know a word by the company it keeps" (Firth 1957: 11). This principle has been exploited in previous corpus-based research by the RDUES team on the automatic identification of synonyms and other semantic relations in the ACRONYM project (Renouf 1996). That project built a 'collocational profile' for each word (type) in a large newspaper corpus, examining every occurrence of the word and recording the other words occurring within a span of four positions to its left and right. Frequencies of co-occurrence were then recorded in a database, along with statistical measures of significance. By comparing the collocational profiles of words in the database, the ACRONYM software is able to extract pairs of words appearing in similar textual environments (or in similar company in Firthian terms). These word pairs are referred to as 'nyms' and, as Renouf (1996: 171) illustrates, include not only synonym pairs (e.g. *luxury*, *five-star*) but also antonym pairs (*luxury*, *no-frills*), and 'hyponym-like' relations (*luxury*, *lexus*).

Our approach in this paper is related to ACRONYM but with a slightly different emphasis. Instead of extracting semantically-related word pairs from the corpus based on their overlapping collocational environments, we start with a list of words we assume to be semantically-related – swearwords – and use collocational analysis to examine overlaps and differences in their usage and meaning in more detail.

In order to arrive at our initial list of potential swearwords, we consulted two different sources. Our first source was the various profanity filter wordlists designed for

use by developers of software tools and online communication platforms, e.g. bannedwordlist.com, noswearing.com. Indeed, there are some profanity filters available as plug-ins for the WordPress and Blogger blogging platforms, to allow blog authors to prevent swearing by readers in comments.[3] One of the most comprehensive wordlists we came across was the "list of 1,300+ English terms that could be found offensive" released by Carnegie Mellon University,[4] an extract of which is given in Table 1.

Table 1. Extract of profanity filter wordlist from Carnegie Mellon University

addict	*beast*	*bombs*	*christ*	*communist*	*criminal*
adult	*bible*	*bra*	*christian*	*conservative*	*criminals*
africa	*bigger*	*buried*	*church*	*conspiracy*	*dead*
asian	*black*	*burn*	*cigarette*	*corruption*	*death*
assassin	*blackout*	*cancer*	*cigs*	*crack*	*demon*
assassinate	*blow*	*catholic*	*cocky*	*crash*	*deposit*
assault	*bomb*	*cemetery*	*color*	*creamy*	*desire*
babe	*bombers*	*chin*	*colored*	*crime*	*destroy*
babies	*bombing*	*chinese*	*coloured*	*crimes*	*devil*

The limitations of such a list are clear from this short extract. No single word in the extract is particularly taboo in itself. It is only in very specific contexts, or in combination with other terms, that a word like *black* might become offensive. The context of a word must be considered carefully before it can be filtered out or allowed to remain. For this reason, we found the Carnegie Mellon list to be too comprehensive and did not make use of it in our study.

Our second source of swearwords in building our initial wordlist was the findings of previous linguistic studies on swearing, including Thelwall's (2008) study of MySpace, Hughes (1998), McEnery (2006), and Beers Fägersten (2012). Through these sources, we finally arrived at a list of 80 swearwords which we decided to use as the starting point for our study (see Appendix for full list).[5] The searches conducted for these words were case-insensitive and we accounted for differences in spelling, number and inflection by designing our search queries carefully in WebCorpLSE. In some cases, wildcard search was sufficient (e.g. *fuck**) but, in others, we had to develop more specific queries (as shown in the Appendix). We used the "refine query" option in WebCorpLSE to check for cases where wildcard search would be too "wild". For example, *arse** would have matched *arsenal*, *arsenic*, *Arsene*, *Arsenio*, etc. as well as the intended target words. In addition, this wildcard search would not have matched the US variant *ass* (and related

words). For these reasons, we used a more specific query to find examples of this swearword (see Appendix). The top 20 most frequent words in our initial list are given in Table 2. This table is ordered by overall frequency in the BBC sub-corpus. The specific frequencies of each word in posts and comments are also given, and these are discussed in more depth in Section 6. In what follows, we refer to each example by the headword given in the left column in the Appendix.

Table 2. The 20 most frequent (potential) swearwords in the BBC sub-corpus (frequencies per million words)

	Posts	**Comments**	**Overall**
god	692.14	608.62	652.34
jesus	262.53	141.63	204.93
hell	116.73	117.90	117.29
shit	115.68	114.92	115.32
christ	138.53	76.87	109.15
fuck*	112.03	101.42	106.98
damn	85.14	127.21	105.18
arse/arses/arsed/arsehole*/ass/asses/assed/asshat*/asshole*	83.54	106.86	94.65
crap*	59.73	85.85	72.18
suck/sucks/sucker/suckers	53.05	86.37	68.93
omg	13.31	101.22	55.19
gay	40.16	40.83	40.48
cow/cows	31.03	40.90	35.73
butt/butts/butthead/buttheads/butthole/buttholes	28.26	37.89	32.85
piss*	32.48	32.45	32.47
bitch/bitches/biatch/biatches	29.30	32.07	30.62
idiot/idiots	23.11	33.65	28.13
balls	29.50	24.76	27.24
screw/screws/screwed/screwing	25.37	26.00	25.67
pig/pigs	24.36	26.67	25.46

One thing that stands out in Table 2 is the proliferation of words with religious associations towards the top of the list: *god*, *jesus*, *hell*, *christ*; possibly also *damn* and *omg* (meaning *oh my god*; we return to *omg* in Section 6). The most frequent word in our list, *god*, is included as a swearword in several of the previous studies discussed above, and would fall into Ljung's (2011) 'religious' taboo theme. Clearly, though, it and the other religious words are highly context-dependent, as illustrated by Examples (17) to (25) from the BBC sub-corpus:

(17) I know that *Jesus Christ* was the Son of God, the Redeemer, and that He truly did atone for our sins and make forgiveness, hope, and happiness possible

(18) Holy shit and *jesus christ* in a rowboat, do these bastards think I am made of money?

(19) So Joel belongs to the greater Church of *Jesus Christ*. Well where the heck is that?

(20) I am no mathalete, or anything, but my calculations are indicating that purple flying pigs will ice skate on a lake of frozen gold in *hell* before I pay that much for this car.

(21) Get the *hell* back in your cage!

(22) Crap like this is exactly why you will burn in *hell*.

(23) I strongly believe in *God* and try to portray it in every form of art

(24) Who was that miserable woman? Thank *god* she's not around any more!

(25) You know what no matter what politics or religion you espouse, no matter what nation you're living in or what nation you're from, for one day a year at least out of life, people of good will ought to set aside a time to step back, open their eyes and say: '*God* damn it. this is one hell of a world'

Here we see three instances each of *jesus christ* (Examples (17) to-(19)), *hell* (Examples (20) to (22)) and *god* (Examples (23) to (25)). Within each group of three there are examples of the term in question being used in a religious context and as part of an instance of swearing. There is a little ambiguity in some of the examples – e.g. in Example (19) *Jesus Christ* is used in reference to a church, i.e. in a literal sense, but is followed by the mild swearword *heck* (not included in our list). However, in most cases there are clear indicators in the immediate context of use, e.g. *God*, *Redeemer*, *atone* and *sins* in Example (17); *shit* and *bastards* in Example (18). These are the indicators we are counting and summarising in the collocational analysis we carry out in the following section to distinguish literal from expressive uses.

5. Collocational analysis

The first step in our analysis was to produce a collocational profile for each of the potential swearwords in our initial list. WebCorpLSE was used to extract the top 100

collocates for each word at span 4 (i.e. four words to the left of the word and four words to the right).[6] The top 25 collocates of *bastard* are given in Table 3 as an example.

Table 3. Top 25 collocates of *bastard* (span 4)

Collocate	Collocate frequency	Co-occurrence frequency	z-score
poor	19,120	64	37.77
little	218,057	182	29.56
Jimmy	2,522	32	28.55
Mr	11,166	35	24.31
those	193,619	137	22.86
sick	14,335	35	22.37
fat	10,773	32	22.34
dirty	5,961	27	21.46
lazy	5,988	27	21.45
rich	11,374	29	19.79
fucking	6,997	23	17.54
greedy	1,093	19	17.33
evil	12,179	26	17.21
cheating	1,657	18	16.05
rat	1,603	17	15.13
selfish	2,698	17	14.57
sneaky	904	16	14.53
miserable	2,817	17	14.51
child	30,397	32	14.23
cheap	9,570	20	13.95
murdering	359	15	13.82
who	323,214	123	12.94
lucky	24,836	26	12.59
lying	5,278	16	12.51
universe	6,176	16	12.15

Table 3 is sorted by z-score, a measure of statistical significance which takes into account the frequency of the node (the swearword) and of each collocate in relation to corpus size. So, for example, although *poor* collocates with *bastard* less frequently than *little* does (64 times versus 182), it is given a higher z-score because *poor* is a much less frequent word than *little* (19,120 occurrences versus 218,057). Both of these words and many of the others in the top 25 are adjectives commonly associated with the noun *bastard*(*s*). Whilst the top collocate, *poor,* can be used in combination with *bastard* to express sympathy, the majority of adjectives in the list appear to be negative evaluators. The caveats we outlined in the previous section do of course apply although, interestingly, one of the words we discussed above, *sick*, appears to remain entirely negative when associated with *bastard*.

In all but six of the 35 span 4 co-occurrences, *sick* actually appears as an immediate left-hand collocate of *bastard* (i.e. at span 1). In four of the six exceptions, the two words are separated by an additional modifier: *twisted* twice, *fuckin'* and *degen* (i.e. degenerate) once each. Example (26) is typical of the majority:

(26) Unlike my parents, I live in almost daily fear that if I relax my vigilance just one second, some sick *bastard* will take my child from me.

Two collocates in Table 3 require further explanation: *Jimmy* and *Mr*. both refer to a blogger known as "Jimmy Bastard"' who was active around 2009-10. He ran his own popular blog hosted on Blogger[7] and was also a frequent commenter on blog posts written by other people on both Blogger and WordPress. It was in the latter context that he was sometimes referred to as "Mr Bastard" by his fellow commenters.

After similar collocational profiles had been built for each of the potential swearwords in our initial list, we took each word in turn and compared its top 100 collocates with the top 100 collocates of all the other words combined. Through this process we were able to uncover (i) the shared collocates of the swearwords, indicating overlapping usage, and (ii) the unique collocates of each, suggesting differences in usage and meaning.

5.1 Shared collocates

Figure 1 is a diagrammatic representation of shared collocates in our BBC sub-corpus. Each row represents one of the potential swearwords, each column represents a collocate, and shaded boxes indicate where swearwords share collocates. The number at the bottom of each column indicates how many of the swearwords share that collocate. For example, the first collocate, *a*, is shared by 54 of the swearwords (*bastard*, *bimbo*, *bitch*, *bloody*, etc.).[8] Figure 1 shows the top 33 shared collocates (all those collocates shared by 17 or more swearwords). We will make several initial observations about the shared collocate results below, as indicated by coloured highlighting in Figure 1.

Firstly, several of the shared collocates (columns) are themselves taboo words: *fucking*, *ass*, *shit*, *fuck*. The first of these – *fucking* – collocates with 31 of the 54 swearwords in our list. What this seems to reflect is that writers in our blog corpus frequently use multiple swearwords in sequence or in close proximity to each other. We

have already seen an example of this in Table 3 with *fucking* as a significant collocate of *bastard*.

Secondly, the taboo words in our list tend to be associated with males more so than with females. In Figure 1 we see *his* as a shared collocate of 27 words, together with *he* and *he's* (26 each), and *him* (17). Of course, we cannot tell from the shared collocate list alone whether the male is the speaker or the referent but this list does provide a useful signpost to guide us in our analysis.

We see evidence of fixed phrases where either *up* or *off* collocates with a swearword (compare the category "destinational usage" mentioned in Example (2), McEnery et al. 2000a, 2000b): *fuck off*, *fuck up*, *balls up*, *cock up*, etc. There are also several intensifiers in the shared collocate list, which contribute to the expression of emotions: *such* (*a*), *big*, *complete*, *total*. An extreme case is Example (27), which is taken from a blog post reviewing a book:

(27) It goes up and then down and there are times when you want to punch both of them – Cathy for being too gullible and Jewel for being such a complete and total *bitch*.

In the context of swearing, we have found that *little* (a shared collocate of 18 swearwords) can also function as an intensifier. We would argue that there is no difference in intensity between Examples (28) and (29):

(28) Will Folks claimed they had an extramarital affair and a pal of mine with ties to a newly elected state representative (his wife) claims she is a big *'skank'* who is 'sleeping with everybody'.

(29) Paula knows how I feel about that little *skank* and she would never, ever do that to me, right?

One final shared collocate worthy of attention is *like*, which is shared by 48 of the words in our list. Initially, we were rather puzzled by this until we looked more closely at the concordance examples. After analysing examples from the corpus in depth, we concluded

that the uses of *like* can be grouped into three main categories and that collocation with *like* can be regarded as a signal that a taboo word is used in an instance of swearing:

i. To act *like* a(n) idiot/moron/asshole/douche/etc.
ii. To feel/look/treat *like* shit/crap
iii. Quotative *like*: e.g. *I was like WTF* ["What The Fuck?"]

In Figure 1, we also see that there are some words (rows) with fewer shared collocates than others. The examples that stand out are *balls*, *bloody*, *christ*, *cow*, *dyke*, *god*, *hell*, *ho*, *homo*, *jesus*, *omg*, *pig*, *queer*, *swine*, and *tart*. Several of these have religious associations, and many of the others are highly polysemous so it is perhaps unsurprising to see that they demonstrate fewer overlaps in usage and meaning than core swearwords such as *fuck* and *shit*. However, it is useful to see this confirmed diagrammatically in Figure 1 and we analyse these words in more depth in Section 5.2.

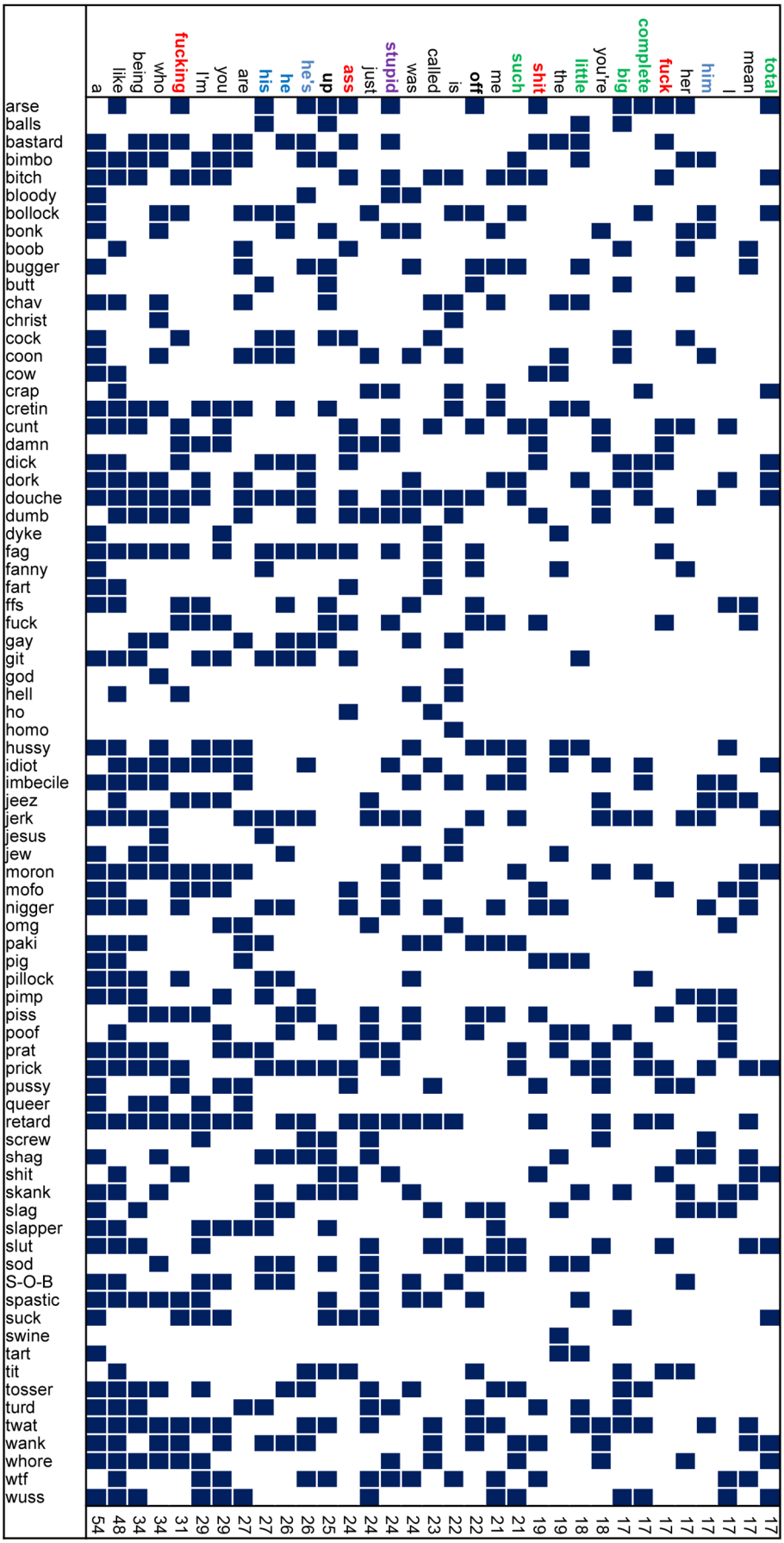

Figure 1. Shared collocates in the BBC sub-corpus

5.2 Unique collocates

Figure 2 shows the unique collocates of a selection of the words from Figure 1 with the fewest shared collocates. The words included in Figure 2 are those with the largest number of unique collocates (shown in parentheses next to the word at the top of each group). For example, *homo* has 72 unique collocates or, in other words, 72 of its top 100 significant collocates (72%) are not shared with any other taboo word.

In each group in Figure 2, the unique collocates are sorted by strength of collocation with the headword (with the top 20 unique collocates listed). The strongest unique collocate of *homo – sapiens* – along with the other unique collocates in that group offer a clear indication of the specific, non-swearing use of *homo* in our data. The same is true of *tart*, which has 66 unique collocates relating largely to baking, and *swine*, with 60 unique collocates relating in particular to swine flu. The headword *balls* has 56 unique collocates, relating to sports and other kinds of physical ball. Based on this collocational information, it would seem that these mild swearwords are rarely used for that purpose at all in our corpus.

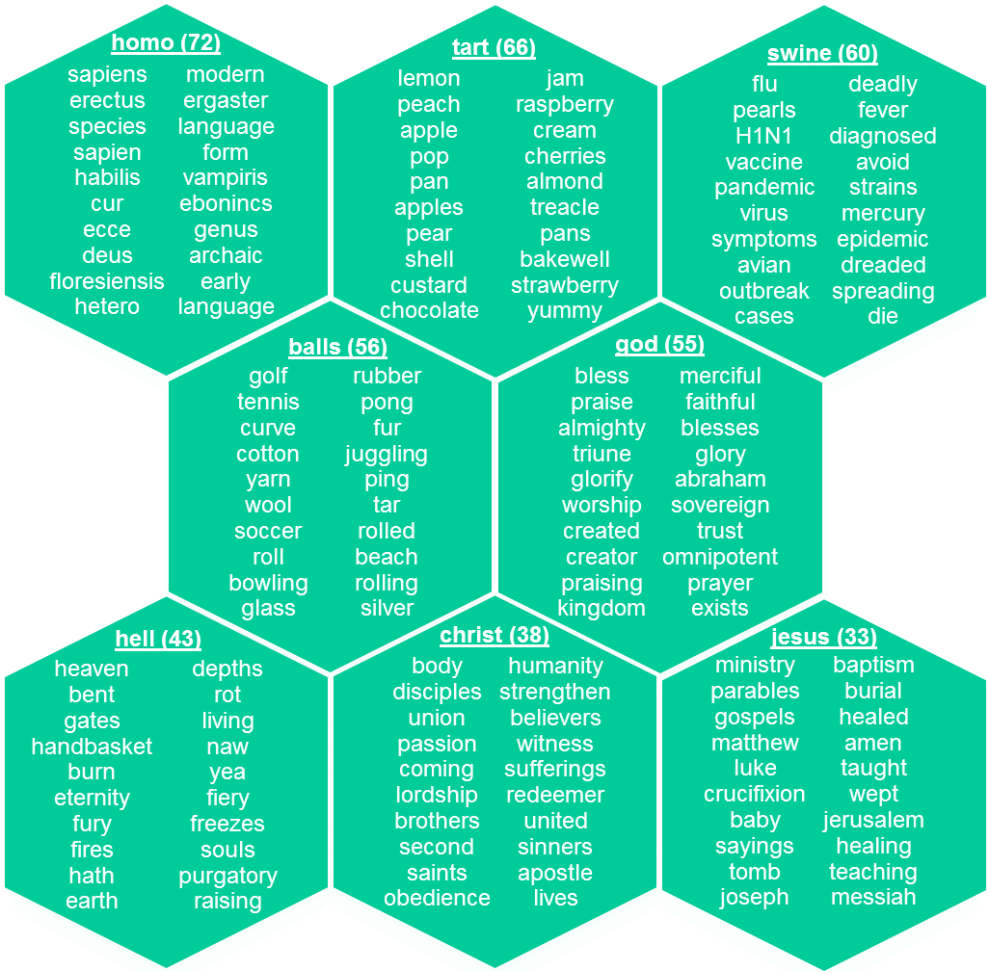

Figure 2. Unique collocates in the BBC sub-corpus

Turning to the religious words in Figure 2, *god* (55), *hell* (43), *christ* (38) and *jesus* (33) all have a large proportion of unique collocates, with *hell* particularly noteworthy for the number of fixed phrases evident in its unique collocate list: *hell bent*, *hell in a handbasket*, *burn in hell*, *hell hath no fury*, *hell on earth*, *depths of hell*, *rot in hell*, *a living hell*, *hell yea/naw* [yes/no], *when hell freezes over*, *hell raising/raising hell*, etc. It is also interesting to see that *christ* and *jesus*, though seemingly closely related, have many unique collocates when compared with each other.

It would seem, then, that the religious swearwords are used primarily for non-swearing purposes in our corpus. That is to say that words such as *god*, *hell*, *christ* and *jesus* are used mainly when discussing religion as a topic. In order to investigate this further, we examined the distribution of all swearwords between blog posts and reader comments, the results of which can be found in the following section.

6. Posts versus comments

Our initial assumption when comparing posts and comments was that topic-related words would be more likely to be found in the post whereas insults and swearing would be more likely in the more informal and conversational comments section.

Overall, we found the combined frequency of all potential swearwords in our list to be similar in posts and comments, with 2,290 per million words in the former and 2,325 per million words in the latter. However, it was when we used the log-likelihood statistic to compare the frequencies of individual words between posts and comments that we began to detect differences. This statistic is often used in corpus linguistic analyses to extract 'key words' from a corpus or sub-corpus by comparing it against a larger reference corpus. We have used log-likelihood analyses previously to extract topic-related words from individual blog posts and comment threads by comparing these against the whole BBC sub-corpus (Kehoe & Gee 2012). In this paper, our comparison is between posts and comments, and the first set of results is given in Table 4.

Table 4.: Key swearwords in posts

	Posts (pmw)*	**Comments (pmw)**	**Log-Likelihood**
jesus	262.53	141.63	3,307.76
christ	138.53	76.87	1,612.41
god	692.14	608.62	485.15
fuck	112.03	101.42	47.64

* pmw: per million words

As shown in Table 4, the word from our list which is most 'key' in posts when compared against comments is *jesus*, followed by *christ* and *god*. In order to explain why these three words are significantly more frequent in posts than in comments, we returned to the corpus and examined the blog posts in which these words are particularly frequent. An example with a high concentration of the words is a post entitled "Demythologizing the Divide between Barth and Bultmann" on an academic blog called *The Fire and the Rose*.[9] This post has only one comment, containing none of the words from our list. However, the post itself is over 7,000 words long and contains 218 instances of the word *god*, plus 18 instances of *jesus* and 13 of *christ*. Using the *WebCorp Live Wordlist Tool*[10] we see that *god* is the most frequent non-stopword in this text. The Wordlist tool also highlights other frequent words which indicate the topic of this text: *barth*, *bultmann*, *revelation*, *doctrine*, *trinity*, *theology*, *humanity*, *divine*, etc. That is to say that an examination of the wider context – going beyond the level of the collocational window to whole text level – tells us that the religious terms are unlikely to be used as swearwords in this case. We are still making use of automated corpus linguistic tools in our pragmatic analysis but we are doing so at a wider level.

If we now turn to the swearwords which are more 'key' in comments when compared against posts (Table 5), we see that one word in particular stands out: *omg*, which is found almost eight times more often in comments than in posts in our corpus and is given a high log-likelihood score as a result.[11]

Table 5. Key swearwords in comments

	Posts (pmw)*	Comments (pmw)	Log-Likelihood
omg	13.31	101.22	-7,014.75
damn	85.14	127.21	-762.01
suck	53.05	86.37	-730.30
crap	59.73	85.85	-427.68
arse	83.54	106.86	-259.80
boo	11.73	20.77	-231.61
idiot	23.11	33.65	-178.54
jeez	2.58	6.22	-141.73
butt	28.26	37.89	-127.70
cow	31.03	40.90	-123.28
fart	6.83	11.77	-120.37
wtf	10.44	15.74	-97.75
douche	6.65	10.80	-90.12
dumb	19.93	26.66	-88.41
ho	12.36	17.14	-70.44
moron	6.04	9.48	-69.52
bugger	5.27	8.39	-65.01
retard	5.10	7.57	-44.05
tit	6.49	8.98	-36.62
bastard	13.81	16.82	-26.84

* pmw: per million words

This word was of particular interest to us as it is an abbreviation of *oh my god* yet it behaves in the completely opposite way to *god*, which, as we saw in Table 4, is significantly more frequent in posts. For this reason, we wanted to analyse *omg* in more depth. We began our analysis of *omg* by examining various dictionary definitions of *Oh My God*, the phrase from which it derives. Several of these are given below:.

i. http://www.oed.com/
   The vocative, as *ah God*, *oh God*, *my God*, *good God*, etc., is used to express strong feeling or excitement.

ii. http://dictionary.cambridge.org/
   used to emphasize how surprised, angry, shocked, etc. you are:
   My God, what a mess!
   Oh my God, I've never seen anything like it!

iii. http://www.macmillandictionary.com/

God - interjection

used for expressing strong feelings such as anger, surprise, or worry. Some people consider this expression offensive.

God! Would you shut up for a minute?

My God, you scared me!

Oh my God, are you all right?

What is noticeable here is how the interjection is often associated with strong negative emotions such as anger, shock and worry. One dictionary – Macmillan – even goes as far as to say that "[s]ome people consider this expression offensive". Intuitively, we did not feel that the abbreviated form *omg* carries the same associations but we wanted to test this empirically using our corpus. To do so, we looked more closely at the span 4 collocates of *omg* in comments only (Table 6).

Table 6. Top 25 collocates of *omg* in comments (span 4)

Collocate	Co-occurrence Freq	Z-score
hilarious	148	62.43
love	950	41.17
funny	201	39.07
lol	234	38.67
cute	206	38.55
soooo	59	33.79
awesome	213	33.26
laughing	75	29.99
amazing	217	29.67
laughed	61	29.56
totally	149	28.99
hysterical	36	28.59
lmao	34	26.98
freaking	39	26.01
adorable	81	25.13
sooo	44	24.03
looks	205	23.66
sooooo	33	23.55
gorgeous	122	22.61
xo	71	21.36
xx	69	20.44
cutest	27	19.15
congrats	83	18.02

xoxo	49	16.78
hugs	89	16.74

It is clear from Table 6 that the strongest collocates of *omg* are all overwhelmingly positive, relating in particular to things that the writer finds funny or cute. We have noticed that *omg* frequently appears in comments on posts containing photographs and that it tends to be the first word in a sentence, as in Examples (30) to (33):

(30) *OMG* that costume is hilarious.
(31) *OMG*, I love your hair!
(32) *OMG* that backdrop is AMAZING!!!
(33) *OMG* it's official, you and your husband are the cutest couple ever!!!

These are very different from the contexts in which we find the word *god*, and there appears to be very little overlap in the usage of *god* and *omg* in our corpus. In fact, of their top 100 span 4 collocates, *omg* and *god* have only one in common – *believe* – and even this one shared collocate is used differently in each case. When *believe* collocates with *god* it is usually in the phrase *believe in god*, whereas the collocation with *omg* is usually a variant of *omg, I can't believe* [x]. This, in addition to our findings presented above, supports our conclusion that while *god* is mainly used in a literal sense in our data, *omg* almost exclusively appears in the context of expressing strong positive emotions.

7. Conclusion

This study approached the topic of swearing in blog posts and comments, with the intention of providing further insights into a topic that has not been investigated extensively in online data. To this end, we linked the study of the pragmatic phenomenon of swearing with a corpus linguistic methodology, showing in particular how a large corpus of written but communicatively immediate online language data can be used in the analysis of a potentially open-ended and expressive category of pragmatic markers.

In contrast to the majority of previous studies on swearing, we based our analysis on a broad definition of swearing, indicating that swearwords could potentially have positive, negative or neutral connotations. Thus, we regard swearing as the speaker's attempt to express their diverse emotions through the use of "bad" or taboo language and

to (potentially) convey these feelings to their interlocutor (see Jay & Janschewitz 2008, Ljung 2011).

We have demonstrated that in order to gain further understanding of a particular use of a swearword, it is essential to consider the context in which it appears. In this study, we therefore examined the company that swearwords keep by carrying out a focused analysis of their shared and unique collocates. This allowed us to obtain results in a semi-automatic manner, which is useful given that the size of the Birmingham Blog Corpus precludes a comprehensive manual analysis. Our study revealed that there is a group of "core" swearwords which are often used in close proximity to each other (e.g. *fuck*, *shit*, *crap*), together with a small set of peripheral swearwords which appear to be rarely used for that purpose at all in our data (e.g. *homo*, *tart*, *swine*). Between these two extremes there are many other words – several relating to religious themes (e.g. *jesus*, *christ*, *god*) – whose categorisation requires careful consideration of context. Further work is necessary to analyse all of these words in depth, but we believe that the innovative collocational approach we have presented in this paper, combining the study of unique and shared collocates, offers significant advantages in the pragmatic analysis of large corpora and that the information gained in this and future studies of swearing in online data will allow the refinement of profanity filters to sift out potentially offensive language use.

Notes

1. For a discussion of taboo and taboo words, see Ljung (2011: 5-8).

2. Example (11) is an exception in that there are no specific clues other than the juxtaposition of the two sentences, where the second is contrasted with the word *wrong* in the first.

3. The use of such plug-ins would, of course, prevent the occurrence of swearwords in a corpus built from WordPress and Blogger. However, the fact that we find many thousands of examples in comments indicates that such plug-ins are not widely used.

4. http://www.cs.cmu.edu/~biglou/resources/bad-words.txt

5. As this list is based on several studies on swearing, its composition reflects the diversity in approaches taken in these studies, to the extent of including forms such as *OMG*, which may traditionally not be regarded as examples of swearing. We further discuss the specific example of *OMG* in Section 6.

6. Although this window would miss some of the contextual clues illustrated in Examples (17) to (25) – e.g. *atone* and *sins* in Example (17) – we chose span 4 as it has been shown to offer meaningful results in previous RDUES projects. In a corpus as large as ours, there will be other examples where *atone* and *sins* do appear within four words of *jesus* and/or *christ.*

7. http://nevermindthebollix.blogspot.co.uk/

8. To some extent there are grammatical restrictions on collocation but this effect is reduced by considering span 4 rather than span 1. For example, *a* would be less likely to collocate with words beginning with a vowel at span 1, but this is not impossible at span 4.

9. http://fireandrose.blogspot.co.uk/2008/06/demythologizing-divide-between-barth.html

10. http://www.webcorp.org.uk/live/wdlist.jsp

11. The scores are negative in this table as we are using the opposite end of the scale used in the previous table (i.e. viewing the comparison in the other direction).

References

Aijmer, K., & Rühlemann, C. (Eds.) (2014) *Corpus Pragmatics. A Handbook.* Cambridge: Cambridge University Press.

Andersson, L., & Trudgill, P. (1990). *Bad Language*. Oxford: Basil Blackwell.

Angouri, J., & Tseliga, T. (2010). "you HAVE NO IDEA WHAT YOU ARE TALKING ABOUT!" From e-disagreement to e-impoliteness in two online fora. *Journal of Politeness Research, 6*(1), 57-82.

Archer, D., Culpeper, J., & Davies, M. (2008). Pragmatic annotation. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 613-641). Berlin: Mouton de Gruyter.

Beers Fägersten, K. (2012). *Who's Swearing Now? The Social Aspects of Conversational Swearing*. Newcastle upon Tyne: Cambridge Scholars Publishing.

boyd, d. (2006). A blogger's blog: Exploring the definition of a medium. *Reconstruction, 6*(4). Retrieved from http://www.danah.org/papers/ABloggersBlog.pdf (last accessed February 2016).

British National Corpus (BNC), XML Edition. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Butler, C. W., & Fitzgerald, R. (2011). "My f***ing personality": Swearing as slips and gaffes in live television broadcasts. *Text & Talk, 31*(5), 525-551.

Culpeper, J. (2011). *Impoliteness. Using Language to Cause Offence*. Cambridge: Cambridge University Press.

Crystal, D. (1997). *The Cambridge Encyclopedia of Language* (2nd ed.). Cambridge: Cambridge University Press.

Firth, J. R. (1956). *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research, 6*(2), 215-242.

Haugh, M. (2010). When is an email really offensive?: Argumentativity and variability in evaluations of impoliteness. *Journal of Politeness Research, 6*(1), 7-31.

Herring, S. C., Scheidt, L. A., Wright, E., & Bonus, S. (2005). Weblogs as a bridging genre. *Information Technology and People, 18*(2), 142-171.

Hughes, G. (1998). *Swearing: A Social History of Foul Language, Oaths and Profanity in English.* Oxford: Blackwell.

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research, 4*, 267-88.

Jucker, Andreas H., Schreier, D., & Hundt, M. (2009). Corpus linguistics, pragmatics and discourse. In A. H. Jucker, D. Schreier & M. Hundt (Eds.), *Corpora: Pragmatics and*

Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29) (pp. 3-9). Amsterdam: Rodopi.

Jucker, Andreas H. (2013). Corpus pragmatics. In J.-O. Östman & J. Verschueren (Eds.), *Handbook of Pragmatics* (pp. 1-18). Amsterdam: John Benjamins.

Kehoe, A. (2006). Diachronic linguistic analysis on the web using WebCorp. In A. Renouf & A. Kehoe (Eds.), *The Changing Face of Corpus Linguistics* (pp.297-307). Amsterdam: Rodopi.

Kehoe, A., & Gee, M. (2007). New corpora from the web: Making web text more "text-like"'. In P. Pahta, I. Taavitsainen, T. Nevalainen & J. Tyrkkö (Eds.), *Studies in Variation, Contacts and Change in English 2: Towards Multimedia in Corpus Studies*. VARIENG E-journal. Helsinki: University of Helsinki. Retrieved from http://www.helsinki.fi/varieng/journal/volumes/02/kehoe_gee (last accessed February 2016).

Kehoe, A., & Gee, M. (2012). Reader comments as an aboutness indicator in online texts: Introducing the Birmingham Blog Corpus. In S. Oksefjell Ebeling, J. Ebeling & H. Hasselgård (Eds.), *Studies in Variation, Contacts and Change in English 12: Aspects of Corpus Linguistics: Compilation, Annotation, Analysis. Proceedings of ICAME 32, VARIENG E-journal*. Helsinki: University of Helsinki. Retrieved from http://www.helsinki.fi/varieng/series/volumes/12/kehoe_gee/ (last accessed February 2016).

Koch, P. (1999). Court records and cartoons. Reflections of spontaneous dialogue in early Romance texts. In A. H. Jucker, G. Fritz & F. Lebsanft (Eds.), *Historical Dialogue Analysis* (pp. 399-429). Amsterdam: John Benjamins.

Ljung, M. (2009). The functions of expletive interjections in spoken English. In A. Renouf & A. Kehoe (Eds.), *Corpus Linguistics: Refinements & Reassessments* (pp. 155-171). Amsterdam: Rodopi.

Ljung, M. (2011). *Swearing. A Cross-cultural Linguistic Study*. Basingstoke: Palgrave Macmillan.

McEnery, A. (2006). *Swearing in English. Bad Language, Purity and Power from 1586 to the Present*. London: Routledge.

McEnery, A., Baker, J. P., & Hardie, A. (2000a). Assessing claims about language use with corpus data – swearing and abuse. In J. Kirk (Ed.), *Corpora Galore: Analyses and Techniques in Describing English* (pp. 45-55). Amsterdam: Rodopi.

McEnery, A., Baker, J. P., & Hardie, A. (2000b). Swearing and abuse in Modern British English. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC' 99: Practical Applications in Language Corpora* (pp. 37-48). Berlin: Peter Lang.

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

Mishne, G., & Glance, N. (2006). Leave a reply: An analysis of weblog comments. *Third Annual Workshop on the Weblogging Ecosystem* (WWW 2006).

Mohr, M. (2013). *Holy Shit. A Brief History of Swearing*. Oxford: Oxford University Press.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM, 47*(12), 41-46.

Nigam, K., & Hurst, M. (2004). Towards a robust metric of opinion. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*. Retrieved from http://www.kamalnigam.com/papers/metric-EAAT04.pdf (last accessed February 2016).

Renouf, A. (1996). The ACRONYM project: Discovering the textual thesaurus. In I. Lancashire, C. Meyer & C. Percy (Eds.), *Synchronic Corpus Linguistics: Papers from English Language Research on Computerized Corpora (ICAME 16)* (pp. 171-187). Amsterdam: Rodopi.

Renouf, A., & Bauer, L. (2001). Contextual clues to word-meaning. *International Journal of Corpus Linguistics, 5*(2), 231-258.

Renouf, A., & Kehoe, A. (2013). Filling the gaps: Using the WebCorp Linguist's Search Engine to supplement existing text resources. *International Journal of Corpus Linguistics, 18*(2), 167-198.

Romero-Trillo, J. (Ed.) (2008). *Pragmatics and Corpus Linguistics. A Mutualistic Entente*. Berlin: Mouton de Gruyter.

Thelwall, M. (2008). "Fk yea I swear": Cursing and gender in MySpace. *Corpora, 3*(1), 83-107.

Upadhyay, S. R. (2010). Identity and impoliteness in computer-mediated reader responses. *Journal of Politeness Research, 6*(1), 105-127.

Appendix. Full list of swearwords included in our study

Head	Pattern matched
arse	arse/arses/arsed/arsehole*/ass/asses/assed/asshat*/asshole*
balls	balls
bastard	bastard/bastards
bimbo	bimbo/bimbos
bitch	bitch/bitches/biatch/biatches
bloody	bloody
bollock	bollock*
bonk	bonk/bonks/bonking
boob	boob/boobs
bugger	bugger/buggers
butt	butt/butts/butthead/buttheads/butthole/buttholes
chav	chav/chavs
christ	christ
cock	cock/cocks
coon	coon/coons
cow	cow/cows
crap	crap*
cretin	cretin*
cunt	cunt*
damn	*damn*
dick	dick/dicks/dickhead/dickheads/dickwad
dyke	dike/dikes/dyke/dykes
dork	dork/dorks/dorky
douche	douche*
dumb	dumb/dumbass/dumbasses
fag	fag/fags/faggot/faggots
fanny	fannies/fanny
fart	fart/farts
fatass	fatass
ffs	ffs
fuck	fuck*
gay	gay
git	git/gits
god	god
hell	hell
ho	ho/hos/hoe/hoes
homo	homo/homos
hussy	hussies/hussy
idiot	idiot/idiots
imbecile	imbecile*
jeez	jeez
jerk	jerk*

jesus	jesus
jew	jew
moron	moron*
motherfucker	motherfuck*/mofo*
nigger	nigga/niggas/niggah/niggahs/niggaz/nigger/niggers/nigguh/nigguhs
omg	omg
paki	paki/pakis
pig	pig/pigs
pillock	pillock/pillocks
pimp	pimp/pimps
piss	piss*
poof	poof*
prat	prat/prats
prick	prick/pricks
pussy	pussies/pussy
queer	queer/queers
retard	retard/retards/retarded
screw	screw/screws/screwed/screwing
shag	shag/shags/shagged/shagging
shit	*shit*
skank	skank*
slag	slag/slags/slagged
slapper	slapper/slappers
slut	slut/sluts
sod	sod/sods
sonofabitch	sonofabitch/son-of-a-bitch
spastic	spastic/spastics
suck	suck/sucks/sucker/suckers
swine	swine/swines
tart	tart/tarts/tarty
tit	tit/tits/titties
tosser	tosser/tossers
turd	turd/turds
twat	twat*
wank	wank*
whore	whore
wtf	wtf
wuss	wuss*

Authors' addresses

Ursula Lutzky
Vienna University of Economics and Business
Welthandelsplatz 1
1020 Vienna
Austria
ursula.lutzky@wu.ac.at

Andrew Kehoe
School of English
Birmingham City University
Curzon Building
4 Curzon Street
Birmingham B4 7BD
United Kingdom
andrew.kehoe@bcu.ac.uk