Title: **Vulgarity in Online Discourse around the English-Speaking World**

Authors

Martin Schweinberger\*

The University of Queensland

Address: School of Languages and Cultures, Gordon Greenwood Building, Union Road,

The University of Queensland, St Lucia, QLD 4072, Australia

Phone: +61 7 3365-6374

Email: [m.schweinberger@uq.edu.au](mailto:m.schweinberger@uq.edu.au)

ORCID: 0000-0003-1923-9153

Kate Burridge

Monash University

Address: School of Languages, Literatures, Cultures and Linguistics, Menzies Building, 20 Chancellors Walk, Monash University, Clayton 3800, Australia

Phone:

Email: [kate.burridge@monash.edu](mailto:kate.burridge@monash.edu)

ORCID: 0000-0002-2349-5160

\*corresponding author

**Abstract**

This paper takes a corpus-based approach to study vulgar language in online communication across 20 English-speaking regions based on the Global Web-Based English Corpus (GloWbE). The identification of vulgar items combines word lists used in profanity detection with regular expressions to identify a wide range of vulgar elements including spelling variants. The results show a notable trend for inner circle L1-varieties to exhibit higher rates of vulgarity online compared to outer circle and L2-varieties. The results also show that inner circle varieties have lower type-token rations which indicates that inner circle variety speakers are more creative and rely on a larger repertoire of vulgar elements compared with speakers from other circle varieties. Finally, the results show that different regions exhibit preferences for vulgarity within different semantic fields: while speakers in the United States prefer vulgarity associated with sexual activity and speakers from Great Britain preferring vulgarity associated with sexual organs, speakers from Hong Kong, Singapore, and Ghana prefer vulgarity based in sex work and promiscuity, and Indian speakers exhibiting a tendency to use more racial slurs than speakers from other varieties. The findings are interpreted to show that cultural differences are reflected in region-specific preferences for vulgarity and that the creativity observed in inner circle varieties is linked to norm-setting compared to norm-reception associated with outer circle varieties.

**Keywords**: Swearing, Vulgarity, bad language, online discourse GloWbE, corpus linguistics

**Highlights**

* Inner circle varieties exhibit higher frequencies of vulgar elements compared with outer circle varieties in online communication
* Speakers in the United States use the highest frequencies of vulgarity online across 20 a sample of 20 English-speaking regions.
* Inner circle variety speakers have lower type-token ratios of vulgar elements indicating that they are making use of a wider range of vulgar forms as well as more non-standard forms and variants of standard forms compared with outer circle variety speakers
* Speakers from different regions show significant preferences for vulgar items originating from different semantic fields

# Introduction

Vulgar language generally refers to words or expressions that are considered rude, offensive, or inappropriate in certain social contexts at a given time (Hughes, 2015, p. x). The usual suspects that challenge social norms in this way include overlapping categories such as blasphemy, curses, ethnic-racial slurs, insults, name-calling, obscenity, profanity, scatology, slang, swearing, tabooed words, verbal aggression, and more — essentially, any form of speech capable of violating conventional standards of politeness. Given the central role vulgarity plays in shaping social interactions and boundaries, it is unsurprising that it has been both a topic of public debate and the subject of substantial scholarly attention.

Despite this interest, however, there remains a notable gap in large-scale studies comparing the frequency, commonness, and creativity of vulgarity across English-speaking regions. The present study aims to address this gap by analysing these aspects across 20 different regions, drawing from Kachru’s[[1]](#footnote-1) “inner circle” (e.g., the U.S., U.K., Australia) and “outer circle” (e.g., India, Nigeria, Singapore) varieties of English, based on the Global Web-Based English Corpus (GloWbE). While Kachru’s labels have been critiqued (including by Kachru himself), we argue that these distinctions are useful for understanding the sociolinguistic realities of vulgarity in both native and non-native varieties of English.

Vulgarity isn’t an inherent quality of a word but is determined of course by where and how it is used. A word deemed crude or coarse in one scenario (e.g. a work meeting or a formal gathering) may be normal or even endearing in another (e.g. a casual chat or a less formal online space). And given there is such complexity and variety of opinions and attitudes, we are unlikely to ever find uniformity of judgement between speakers of even very similar social backgrounds.

However, it would ignore reality to pretend that ordinary people do not perceive certain expressions (such as *shit*) as being somehow intrinsically vulgar. Milwood-Hargrave (2000), for example, is a British survey where participants were asked to respond to the perceived ‘strength’ of twenty-eight swearwords. No context was provided; yet, participants were clear about the severity of these words and the researchers were able to put together a broad topography of swearwords across all groups. Modern dictionaries also treat expressions as if they were inherently offensive when they brand particular entries with labels such as “coarse”, “impolite”, “insulting”, “obscene”, “offensive”, “profane”, “slang”, “taboo”, “vulgar” etc. For example, *shit* is labelled by the *Merriam- Webster* as “vulgar”, the *Macquarie Dictionary* as “colloquial (taboo)”, the *Oxford English Dictionary* as “coarse slang”, *Wiktionary* as “colloquial, vulgar”, the *American Heritage Dictionary* as “vulgar slang” and the *Collins English Dictionary* “informal, rude”; Wikipedia describes it as “an English-language profanity”. And yet, these dictionary descriptions provide no basis for these categorical value judgments, nor do other sources such as media reports, or even legal materials. Similarly, those who compile dictionaries of euphemism and dysphemism (e.g. Ayto, 1993, and Green, 1996), base their collections on the prejudiced social attitudes to the situation in which they believe given expressions to be uttered. This sort of ‘sensitive handling’ employs a type of good etiquette gauge by which is determined the value of an expression without reference to the context of use (something Keith Allan dubbed “the middle-class politeness criterion”; cf. Allan and Burridge, 2006, p. 33; Allan, 2015a). This is also the default measure we have to assume for the decontextualized expressions we examine here.

Of course, applying such a default measure globally becomes more complex. Politeness and vulgarity are culturally specific, and the criterion may not align with diverse, localized uses of English — what might be considered vulgar or offensive in one context may not carry the same weight in another. Though barely a taboo word in Australian English, *bloody* still raises eyebrows in other parts of the English-speaking world, especially when it appears in the public arena(e.g. Tourism Australia’s 2006 international advertising slogan *So where the bloody hell are you?* was banned from British TV and p-parts of North America). Even terms, such as *cunt,* widely recognized as offensive, may not appear on the radar for some nations, or for certain groups within those nations. In Australia, *cunt* is now viewed by younger speakers to be a significant part of Australian culture and identity, even by those who choose not to use the word (“to have an issue with the word *cunt* is seen as having an issue with the concept of Australianness”; see Hughes this volume). Given the diversity of Englishes around the world, it is conceivable that some expressions might not even be widely understood, and local languages may also have their own set of offensive terms that take precedence.

To identify vulgar language across the English-speaking nations examined here, we have relied on the List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words (Emerick, 2024), a vulgarity filter used by many globalized tech companies to monitor and moderate content. Any such standardized list will be a blunt tool admittedly, but it is nonetheless useful for gauging obscenity rates across the different regions — while on the one hand, it oversimplifies the spectrum of vulgarity and regional sensitivities (e.g. nuances of slang or context-dependent terms), on the other hand, it provides a consistent, albeit limited, baseline for analysis. At the very least, this will reveal patterns of language use that are broadly shared within the English-speaking world, and this is useful for identifying trends in obscenity across English-speaking regions, and for understanding language varieties that might be in the process of transitioning from one category to the other (e.g. the change from ESL to ENL in Singapore).

The following section discusses previous research relevant for the present study while section 3 provides information about the data that formed the basis of the current analysis as well as details concerning data processing and the quantitative analysis. Section 4 presents the findings of the analysis and section 5 discusses these findings in light of the existing literature. Finally, section 6 offers an outlook on future directions and potential applications of the analysis performed here.

# Previous Research

Rather than being a simple, easily definable phenomenon, vulgarity proves to be a complex and multifaceted linguistic phenomenon. Even the use of swear words, probably the most usual way people engage in vulgar language, is characterised by “inherent variability and […] subjectivity” (Beers Fägersten, 2012). Nonetheless, vulgar language has been the subject of analysis across various academic disciplines, each bringing its own perspective and terminology to the study.

The terminology used to describe vulgar language highlights the different dimensions of the phenomenon that each discipline prioritizes. The term swearing in corpus linguistics, for instance, conveys a focus on the linguistic form itself, whereas terms like offense and face-threatening in pragmatics emphasize the interpersonal and communicative impact of vulgarity. In contrast, the use of terms like slang, taboo, or profanity points to a broader social and cultural context, considering vulgarity as part of the larger landscape of non-standard language that plays a role in identity construction and social dynamics. These divergent terminologies and perspectives underscore the richness and complexity of vulgarity as a subject of study, revealing how it operates across different layers of language and society.

Various studies have approached vulgarity in language from a phenomenological perspective, trying to understand its functions in discourse (Jay, 2009; Allan & Burridge, 2010, McEnery, et al., 2023) and how it differs from other forms of expression (see, e.g., Allan, 2018; Beers Fägersten & Stapleton, 2017; Bergen, 2016). Another well-established strand of research focuses on psychological aspects of swearing and its relationship with emotion (e.g., Dewaele, 2004) and its physiological effects (Stephens, Atkins, & Kingston, 2009; Stephens et al., 2018; Stephens & Robertson, 2020). Also, research has analysed the perception of swearing (Johnson & Lewis, 2010), how it affects listeners, and impacts their evaluations of users of profanity. Corpus-based research emphasizes the quantifiable aspects of swearing, identifying patterns and trends in how specific vulgar forms (swear words) are employed in different contexts (e.g. Schweinberger, 2018) or the development of swearing over time (e.g. Love, 2021; McEnery, 2004; Mohr, 2013).

On the other hand, scholars in pragmatics are more likely to frame their investigations around the concept of offense or face-threatening acts, exploring how vulgarity functions in communication (Jay & Janschewitz, 2008), particularly in terms of politeness, impoliteness, and the negotiation of social norms (Allan & Burridge, 1991; 2006; Culpeper, 2011; 2018). Here, the emphasis is on the pragmatic effects of using vulgar language, how it is reacted to in discourse, and the intentions behind it, rather than merely its presence. In contrast to larger scale corpus-based studies, this research typically analyses smaller samples in detail, allowing this research to investigate the context-dependence of offence, impoliteness, and vulgarity. Crucially, fine-grained analyses that focus on the discursive context highlight that whether words considered offensive or impolite depends on culture or setting-related factors: what is deemed offensive in one context might be less offensive or even benign in another (see, e.g. Allan, 2015b).

Sociolinguists approach vulgarity from a cultural and social perspective, often referring to the study of such language as the analysis of slang, taboo words, or non-standard language. This type of research is interested in how vulgarity is embedded in the social fabric, reflecting and reinforcing cultural identities, group memberships, and social hierarchies. For example, this form of research might explore how specific vulgar terms carry cultural significance or serve as markers of in-group solidarity and may examine the social stratification of vulgar language, analysing how factors such as class (McEnery, 2004), gender (Bednarek, 2015; Beers Fägersten, 2012), age (e.g., Stenström, 2017), and ethnicity (e.g., Troutman, 2006) influence the use and perception of vulgarity.

*2.1 Types and functions of vulgarity*

Being a multifaceted tool in human communication, vulgar language serves a variety of functions in communication, from emotional expression and social bonding to cognitive processing and identity formation. Recent research has also shown the positive effects of vulgar language on relieving stress, coping with pain and increasing strength and endurance.

A number of researchers have identified different types of vulgar language, in particular of swearing where most of the research as been done (e.g. Jay, 2000; Pinker, 2007; Adams, 2016). The following brief account has organized all these various uses into four different (but overlapping) functions (originally outlined in Allan & Burridge, 2010). Each function is illustrated by examples drawn from conversations in the Australian contribution to the International Corpus of English (ICE-AUS, Smith & Peters, 2023), and the Wellington Corpus of Spoken New Zealand English (WSC, Holmes, Vine, & Johnson, 1998).

*Expletive swearing:* Expletives are a type of interjection used in emotionally intense situations, like anger, frustration or sudden pain. The cathartic effect of swearing comes from breaking taboos, which helps the speaker cope with the emotional surge, especially in moments of a sudden realization, such as making a major mistake (the “onosecond”).

1. I ran off because it’s something like you know eeh eeh eeh eh eh eeh eeh eeh eh suddenly this string just went Boom I don’t know bang and I just went **Fuck** and ran off the stage. (ICE-AUS)

Sometimes the situation provoking the emotional outburst is pleasing, as in this second example.

1. **SHIT** that’s great. (WSC)

Clearly not everyone approves of this sort of auto-cathartic swearing, and society recognizes the dilemma by providing an out – speakers can choose between a swearword and one of the many conventionalised remodelled expressions such as *Oh sugar!* or *Oh shucks!*.

1. Oh **sugar**. We’ve burnt it. (ICE-AUS)

*Abusive swearing:* Malediction includes curses, name-calling, and derogatory comments meant to insult or offend someone; it is also cathartic when it involves highly emotional language produced in anger or in frustration. This abusive language can target people directly (e.g. “You rotten bugger!”) or be used to disparage third parties (e.g. “The bastard stole my pen”) or irritating situations (e.g. “Writing grant applications is a bugger”). Swearing expresses frustration, disapproval or an intent to degrade, and it includes curses, such as the imprecative in the following example.

1. **Fuck** you NAME! (ICE-AUS)

*Social swearing:* Swearing can be a symbol of cohesion within a group, particularly when it’s part of a shared casual style that distinguishes them from out-group individuals. This is vulgar language woven into friendly chatter — a kind of mock impoliteness that is common between speakers who are close and who interact frequently.

1. S1: pray to baby Jesus open up your heart let god’s love come pouring in let god’s love shine down on you like it has me and Miss Suzanne over here.

S2: oh **fuck off**. (ICE-AUS)

Sometimes banter becomes more like a ritual slanging match. In the following example (from urban Aboriginal Australia), you can see teenagers in front of an audience taunting each other with insults. These mimic true insults but are used among friends as a form of playful competition, and though the language mirrors offensive terms, it is not meant to wound but rather to strengthen bonds, with both parties exchanging the insults without animosity.

1. **A:** If I had a pussy like yours I’d take it to the cat’s home and have it put down …

**B:** If I had brains like yours I’d ask for a refund …

**A:** Well, if I had tits like yours I’d sell them off for basketballs … (Allen 1987: 66)

*Stylistic swearing:* Speakers may opt for a taboo term instead of a milder alternative to convey a certain stance. Whether it’s disapproval, astonishment, frustration or excitement, this use of swearing injects emotional intensity or flair, giving more impact to what is being communicated.

1. There’s no point in Eddie taking her out because she’s **bloody** too stuffed you know. She’s an old duck. She doesn’t want to **bloody** stuff around town all day. […] Yeah she went down there and **bloody** went all over the place. (ICE-AUS)

For decades, and across a wide range of studies, social swearing has emerged as by far the most usual type (e.g. Ross, 1960; Allan & Burridge, 2010; Musgrave & Burridge, 2014; Beers Fägersten & Stapleton, 2017). We might assume that this would also be case for everyday written communication online and be reflected in the wide range of genres offered on GloWbE (especially the informal social media posts, blogs, and forums). However, given our focus here is simply on the form and frequency of these so-called “bad” words, we have not been able to incorporate their functional aspects into the analysis.

*2.2 The vulgar lexicon and creativity*

As Darwin (1872, p. 93) observed, humans have a deep-seated craving for novelty and expressiveness, and they are constantly reimagining and adapting the aesthetics of their expressions. Vulgar language is the natural playground for unleashing this linguistic creativity.

Vulgar language taps into the taboos that arise out of social constraints on individuals’ behaviour where such actions can lead to discomfort, harm or injury (Allan & Burridge, 2006, p. 1) — it is societal fears that underpin vulgarity and give it its power and impact (the shock value, the emotional charge and the social fall-out when boundaries are breached). And they are what help to drive the creativity. Much of the taboo lexicon involves comedic and playful language use, allowing people to manage discomfort by downplaying what is hard for them to confront (Freud, 1927, p.163; see also “*shit happens*-humour” in Storch & Nassenstein, 2020).

Of course, it is also true that speakers will generally manipulate words and play with expressions for the sheer pleasure of it. Crystal’s early account of “ludic linguistics” demonstrated the ubiquity and artistry of language play among ordinary English language users: “when children arrive in school, their linguistic life has been one willingly given over to language play” (1988, p. 183). For a number of reasons, children have a particular fascination for vulgar language, but this fascination doesn’t stop at childhood. For children and adults alike, vulgarity is a tool for creative expression and playful experimentation with words and phrases, often involving literary devices traditionally associated with great literary works (e.g. startling imagery and vibrant sound associations), and in the online world, inventive spellings, intentional misspellings, often accompanied by visual play (emojis, symbols, alternative characters etc.). Creative possibilities increase when it comes to outer circle Englishes, when experimentation can involve a dynamic fusion of English blended with local idioms or slang.

A particularly creative type of verbal play, and one that has been part of inner circle varieties of English for centuries, involves unexpected mixes of terms to create humorous, shocking and memorable compounds and blends. Consider the flourishing of insults for stupidity. Since at least the 1500s, speakers have been creating compounds based on *ass/arse* (*arsehole, arsewipe, dumbarse, dumbass, donkeys arse* are a few; Burridge in press); *ass/arse* is a tabooed part of the anatomy and at the opposite end from the head, both features rendering it an appropriate seat of stupidity. Creative -*head* compounds for stupidity have been in the language a similarly long time — hundreds of examples exist, and many are still used today (e.g. *blockhead* from the early 1500s) — and, like the new wine that gets poured into an old bottle, it is the turn-over of words combining with *head* that maintains the contemporary appeal of these expressions. Earlier examples include animals (*asshead* 1400s), fish and insects (*cod’s head* and *dorhead* 1500s), vegetables (*cabbage head* 1600s), soft materials (*woollen head* 1700s), porridge-like substances (*pudding head* 1800s), and so on. More recent times, the words that combine with *head* involve bodily effluvia, tabooed body-parts and sexual past-times; for example, *poop-head, butthead, dickhead, fuckhead*. These most recent *-head* examples mirror evolving taboos of modern times and are also reflected in current shifts in swearing patterns; e.g. *(God) damn X* provides the model for more potent modern imprecatives such as *Shit on X, Fuck X, Bugger X*, and so on. The potency of physically and sexually based taboos is now waning, and recent appearances of expressions such as *vaxhead* might well be forecasting another change in idiom.

Certainly *-head* appears prominently among the lively “suffixes” in Morris’s (2022) recipes for forming effective derogatory noun-noun compounds in English: he describes it as an incredibly versatile suffix that can be preceded by just about any taboo word. Morris’s list of 140 affix-like forms (e.g. *butt-*, *fart-*, *poop-*, -head, -face, -brains) combine to form 4,800 possible pejorative compounds (e.g. *butthead, fuckwaffle, dicknose, assbag*),more than half of which appeared in Reddit comments from 2006 to the end of 2020. “The truly creative”, as Morris describes, “can go outside the box with oddball constructions like *fucktrumpet*, or *wankpuffin*”. Pinpointing the creativity around vulgar items is clearly an important part of profanity detection across the Englishes.

*2.3 Research questions*

Despite the increased interest in slang, swearing and terms of insult over the last few decades, research has focused overwhelmingly on inner circle Englishes, and there have been very few studies of vulgarity in non-Western varieties of English (e.g. Wen et al., 2021 on social swearing amongst Singaporean youth; Farquharson et al. 2020 on the semantic and syntactic aspects of swear words in Jamaican, both those with and without an English provenance). And while there have been cross-cultural analyses of impoliteness (e.g. Culpepper, 2011), there has been little comparative research across English varieties, even those within the inner circle. Notable exceptions are Goddard (2015), who identifies potential differences between Australia and the USA in the public use of swear words, and Dewaele (2015), who investigates differences between speakers of British American English regarding understanding, perceived offensiveness and the frequency of offensive words).

Clearly, we can’t assume that usage and attitudes will be the same across these very different speech communities, and what is needed is a comprehensive study that compares the frequency and usage, as well as the inventiveness of vulgar language across various English-speaking regions — taking in both traditional English-speaking countries and the outer circle of influence. In response to this, the present study thus aims to answer the following research questions:

1. How frequent is vulgar language in online discourse across the 20 different English-speaking regions represented in GloWbE?
2. What vulgar elements are used in the data and are there region-specific differences or preferences?
3. Is there are difference in the variability (creativity) of vulgar elements that are used?

# Data and Methodology

## Data

This study uses the texts of the downloadable version of the Global Web-Based English Corpus (GloWbE, Davies & Fuchs, 2015) as a basis for analysis. The downloadable version of GloWbE consists of 1.7 billion words across 20 regional components[[2]](#footnote-2) representing informal blog content (approximately 60 percent of the total data) and a wide range of other web-based materials, genres, and text types (here referred to as general web content which approximates to 40 percent of the total data)[[3]](#footnote-3). The regions represent 6 inner circle (Australia, Canada, Great Britain, Ireland, New Zealand, and United States) and 14 outer circle varieties of English or English-speaking regions (Bangladesh, Ghana, Hong Kong, India, Jamaica, Kenya, Malaysia, Nigeria, Pakistan, Philippines, Singapore, South Africa, Sri Lanka, and Tanzania). These regions where determined based on regional identifiers that are part of urls (e.g., au, ie, us, etc.). For copyright reasons, every 200 words, 10 words are replaced with a @ symbol. An overview of the distribution of files and tokens in the GloWbE data is provide in Table 2.

The data processing and extraction of vulgar elements was performed in R (version 4.3.2, R Core Team, 2023) using RStudio (version 2024.4.1.748, Posit team, 2024) and consisted of the following steps:

* The GloWbE text data was loaded into R and texts were split if words had been blacked.
* The resulting texts were then converted into a data frame with columns for the region, file name, text identifier, subtext number, discourse type (general web versus blog), and raw text.
* The raw text was then part-of-speech tagged and cleaned by converting the txt to lower case and removing non-word elements such as emojis and punctuation marks.
* In a next step, the data frame was converted to a long format where each row represented a single token (rather than a whole text).
* Then, using a list of regular expressions which was based on the English [*List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words*](https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/tree/master)(LDNOOBW, Emerick, 2024), it was determined for each token if it was a potentially vulgar item.

The LDNOOBW (Emerick, 2024) is an openly accessible resource used by platforms (professional, gaming or social) for content moderation, censorship and automated filtering systems — the goal being to prevent inappropriate and offensive language from appearing in shared spaces such as schools, workplaces and other public domains. As earlier flagged, what is considered vulgar will vary significantly by culture, region, community norms and also platforms, and there will always be risks of over-filtering (e.g. words on the list can be non-offensive in certain contexts, such as educational or medical). Hence, this is not an official or universally standardized list, but a collaborative and constantly evolving document, which can vary across platforms and across time. Guided by algorithms, user feedback and community recommendations, individual platforms and organizations update and adapt the list of “bad” words and phrases, adjusting it according to their audiences and specific use cases.  Shutterstock, for example, in creating their own version of the list, make judgment calls based on the question: “What wouldn’t we want to *suggest* that people look at?”. They conclude, “what goes in these lists is subjective […] so in the end we just have to make our best guess”. We modified the LDNOOBW to incorporate findings from previous studies on vulgar language. Specifically, we added terms identified as vulgar in earlier research that were not originally included in the LDNOOBW (e.g., *idiot*, *screw*, *wanker*). At the same time, we removed medical terms referring to sexual organs (e.g., *vagina*, *penis*, *vulva*) and non-offensive terms such as *incest*, *escort*, *babeland*, and *BBW*. Once the LDNOOBW was cleaned and expanded, we converted the list into a series of regular expressions to capture variations of the identified vulgar terms.

In this study, our focus is on the form and frequency of these so-called “bad” words, without looking at their functional aspects. Hence, our list takes in both literal and non-literal uses (e.g. *She fucked him* versus *You’re fucked*). The inclusion of literal uses appears to be the predominant practice, on the grounds that both literal and non-literal usesare linked to taboo and both will have emotional impact. Even though some may not view it as swearing, most will judge referential *shit* ‘piece of excrement’ (*it was caked in shit*) as vulgar — certainly more evocative and perhaps more shocking than expressions such as *excrement* or *manure*. For this reason, we also include words that may be used offensively but wouldn’t be considered swear words (e.g. *bonk, pig*). (Compare Ljung, 2011, who excludes literal uses of swearwords with McEnery’s, 2005, inclusive approach.)

Using a list of regular expressions, we were able to detect spelling variations, as well as prefixed, suffixed, and compound forms of vulgar words. Regular expressions (commonly abbreviated as regex or regexp) are powerful tools for matching specific sequences of characters in text. They enable flexible and efficient text processing by defining rules that can identify particular word patterns, including variations in spelling and formatting. For example, a regular expression can be used to locate all forms of the word *fuck* by specifying possible letter combinations and character types.

In the expression \\bf[ou\\W]{1,}[c]{0,}[k\\W]{1,}(s|a|er|ers|ed|ing){0,}\\b, the computer looks for:

* A word boundary (\\b), followed by the letter *f*.
* The next part of the word must contain *o*, *u*, or any non-alphanumeric character ([ou\\W]).
* This is followed by an optional *c* and either a *k* or another non-alphanumeric character.
* The final part is optional and can include *s*, *a*, *er*, *ers*, *ed*, or *ing*.

This approach allows for capturing a wide range of word forms efficiently, including common spelling and morphological variations. As the detection was based on a list of regular expressions, the results were noisy, meaning that the list of potentially vulgar elements contained many false positives. The results amounted to 1,800 potentially vulgar forms which were thus manually crosschecked and non-vulgar elements were removed. False positives included many non-English tokens such as *titisee* (the name of a German lake), *hora* (a Bulgarian dance and Spanish word for hour or time), *fukai* (Japanese for deep), *gungho* (a term meaning overly energetic that originated during the Sino-Japanese War) as well as regular, non-vulgar English words, for example, *album*, *appearances*, or *area*. For elements where a form-based evaluation was not possible due to ambiguity or context dependence (for example terms like *pig*, *pisspot*, *tart*, or *prick*), selected samples were drawn and inspected. If the term was used frequently in a way that was deemed vulgar or insulting, the term was retained but removed if the use in the sample was dominantly non-vulgar. After the manual crosscheck, 555 vulgar types were retained in the data. Another issue were variety specific uses of potentially vulgar words such as *jerk* (a common dish in Jamaica) or *hooker* (a popular vacation destination in New Zealand as well as a term for a ruby move that was particularly common in the Australian data). Although we could have excluded these terms, we decided to retain them and contextualising their frequency of use.

To be able to get a more detailed understanding of the use of vulgarity, the forms of vulgarity detected in the GloWbE were then annotated, that is additional information was added regarding the lemma of each vulgar token and the basis of the form. The annotation classified all types into lemmas to harmonise spelling variation, creative variants, or compounds. The 555 vulgar types where classified into 75 vulgar lemmas (*arse, ass, asshole, ballsack, bastard, bellend, bint, bitch, bloody, bollocks, boner, bonk, boob, bugger, bukkake, butt, chink, cock, coon, cornhole, cracker, crap, cum, cunt, damn, darkie, dick, dildo, dyke, fagit, fanny, feck, frigg, fuck, git, hag, hoe, hooker, idiot, jab, jap, jerk, jizz, juggs, knob, lesbo, libtard, masturbate, minger, moron, muff, nigger, nonce, nympho, online, paki, pecker, penis, pig, piss, poofter, prick, pussy, shag, shit, sissy, skank, slag, slut, spastic, tits, tosser, twat, wank, whore*). For example, the types of the lemma *fuck* included the 81 types *fuck, fucking, fucked, eff, fuckin, effing, fucks, fk, fac, effie, motherfucker, fcuk, clusterfuck, mofo, fuk, effin, fck, motherfuckers, fuckedup, fucken, effed, fuckup, phuc, effi, fke, fcking, fkn, fking, fukien, fks, fak, phukan, fuckups, fack, phuck, fukin, fuckoff, facc, fuckyou, effs, fuckall, fuking, fka, fukk, fcks, fkin, f, fucka, mofos, mindfuck, fukang, fuka, fuckn, fckn, effies, fuks, fuke, fki, effyou, effe, phucking, fukn, fked, fkd, effedup, headfuck, fuked, fuckton, fucki, fkz, fckin, fcked, effen, phucas, muhfuck, fuuuuuck, fukkk, fuki, fuecks, fuckiing,* and *effend*.

The resulting annotated data set is summarised in Table 1 amounts to approximately 1.7 billion tokens across 1,74 million files representing blog and general web data from 20 English-speaking regions. Of the 1.7 billion tokens 400,661 tokens (0.023 percent of all tokens) were identified as vulgar.

Table 1: Overview of files, text fragments, tokens, vulgar types, and vulgar tokens across regional components and discourse types in GloWbE.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Region | Discourse Type | Files | Text Fragments | Tokens | Vulgar Types | Vulgar Tokens |
| Australia (au) | General | 81,677 | 93,366 | 97,761,223 | 171 | 24,188 |
| Bangladesh (bd) | Blog | 14,212 | 15,171 | 10,081,315 | 74 | 999 |
| Bangladesh (bd) | General | 30,807 | 32,975 | 26,892,210 | 97 | 2,129 |
| Canada (ca) | Blog | 54,053 | 61,216 | 40,854,594 | 145 | 9,509 |
| Canada (ca) | General | 81,642 | 90,982 | 85,038,924 | 158 | 12,842 |
| Ghana (gh) | Blog | 15,162 | 16,387 | 10,369,906 | 82 | 1044 |
| Ghana (gh) | General | 32,189 | 34,918 | 25,917,297 | 109 | 2,406 |
| Great Britain (gb) | Blog | 149,407 | 171,608 | 122,564,578 | 206 | 36,458 |
| Great Britain (gb) | General | 232,418 | 266,543 | 238,349,216 | 219 | 62,303 |
| Hong Kong (hk) | Blog | 16,040 | 17,697 | 11,704,117 | 94 | 2,295 |
| Hong Kong (hk) | General | 27,897 | 30,349 | 26,146,584 | 104 | 3,092 |
| India (in) | Blog | 37,154 | 40,084 | 26,499,984 | 110 | 3,704 |
| India (in) | General | 76,618 | 82,290 | 63,775,220 | 144 | 6,612 |
| Ireland (ie) | Blog | 26,712 | 30,306 | 19,019,874 | 125 | 4,920 |
| Ireland (ie) | General | 75,433 | 85,165 | 75,285,227 | 176 | 15,766 |
| Jamaica (jm) | Blog | 15,820 | 17,792 | 10,388,260 | 105 | 3,770 |
| Jamaica (jm) | General | 30,924 | 34,196 | 26,610,239 | 113 | 5,094 |
| Kenya (ke) | Blog | 14,794 | 16,297 | 11,686,318 | 96 | 1,969 |
| Kenya (ke) | General | 31,165 | 33,511 | 26,771,355 | 99 | 2,470 |
| Malaysia (my) | Blog | 16,287 | 18,651 | 12,444,023 | 107 | 3,208 |
| Malaysia (my) | General | 28,738 | 32,231 | 26,485,418 | 118 | 4,785 |
| New Zealand (nz) | Blog | 27,816 | 32,054 | 21,083,227 | 139 | 6,833 |
| New Zealand (nz) | General | 54,861 | 61,787 | 54,890,902 | 157 | 10,771 |
| Nigeria (ng) | Blog | 13,956 | 15,375 | 11,233,221 | 79 | 1,538 |
| Nigeria (ng) | General | 23,329 | 26,736 | 28,762,438 | 127 | 4,816 |
| Pakistan (pk) | Blog | 16,917 | 18,338 | 12,484,910 | 84 | 1,382 |
| Pakistan (pk) | General | 25,852 | 29,239 | 35,704,698 | 101 | 3,795 |
| Philippines (ph) | Blog | 17,949 | 19,782 | 12,574,833 | 102 | 1,913 |
| Philippines (ph) | General | 28,391 | 31,676 | 27,868,201 | 114 | 4,596 |
| Singapore (sg) | Blog | 17,127 | 19,555 | 12,804,269 | 99 | 3,083 |
| Singapore (sg) | General | 28,329 | 32,202 | 27,324,427 | 126 | 6,574 |
| South Africa (za) | Blog | 16,995 | 19,002 | 12,753,616 | 96 | 2,540 |
| South Africa (za) | General | 25,312 | 27,991 | 31,725,376 | 101 | 3,790 |
| Sri Lanka (lk) | Blog | 13,079 | 14,465 | 11,959,198 | 89 | 2,044 |
| Sri Lanka (lk) | General | 25,312 | 27,990 | 31,725,376 | 101 | 3,797 |
| Tanzania (tz) | Blog | 13,820 | 14,804 | 9,581,162 | 87 | 1,195 |
| Tanzania (tz) | General | 27,533 | 29,250 | 23,340,175 | 96 | 1,847 |
| United States (us) | Blog | 106,389 | 125,776 | 124,180,889 | 189 | 54,279 |
| United States (us) | General | 168,757 | 197,330 | 237,094,745 | 204 | 76,305 |
| Total |  | 1,740,873 | 1,965,087 | 1,721,737,545 | 4,743 | 400,661 |

We now turn to the statistical methods that we used to analyse the data.

## Statistical Methods

This study makes use of simple frequency tabulation paired with corrected type-token ratios (CTTR) and configural frequency analysis (CFA) to ascertain distributional differences in the use of vulgarity across English-speaking regions. The following provides details on the methods and how they are calculated to better understand how these methods work and what their results show.

Frequency Tabulation

Frequency tabulation involved extracting the raw frequencies of vulgar elements for the regional components, sometimes subdivided by discourse types. To then normalise these raw frequencies, the relative frequencies (either percentages or per-million-word frequencies) were calculated by extracting the frequencies of all elements in the respective subsections of the data and then dividing the raw frequencies of the vulgar elements by the size of the subsection multiplied by 100 (for percentages) or by one million (for per-million-word frequencies). The results represented the normalised frequencies used in the visualisations.

Corrected Type-Token Ratio (CTTR)

The Corrected Type Token Ratio (CTTR) is a measure of lexical diversity that adjusts for text length to provide a more standardized metric of vocabulary richness, particularly useful when comparing texts of different lengths (Malvern et al., 2004). Unlike the simple Type Token Ratio (TTR), which divides the number of unique words (types) by the total number of words (tokens), CTTR corrects for text length by normalizing this ratio as shown below.

The CTTR is calculated as the number of types divided by the square root of twice the number of tokens:

By normalizing the ratio against the square root of twice the number of tokens, CTTR controls for variations in text length, making it a better comparison across regions. In this study, CTTR is used to evaluate the variability in the use of vulgar words across different regions. Regions with higher CTTRs indicate greater diversity in the use of vulgar types, while lower CTTRs suggest less diversity.

Configural Frequency Analysis (CFA)

Configural Frequency Analysis (CFA) is a statistical method used to detect patterns of association or disassociation within categorical data (Bortz et al., 2008; Lienert & Krauth, 1975). The CFA can be considered a multivariate extension to Pearson’s χ2-test (chi-square test). In contrast to Pearson’s χ2-test which only informs that somewhere in a table with more than two rows and/or two columns there is at least one cell where the expected value differs significantly for the observed value, a CFA identifies specific configurations (combinations of variables or categories) that occur more or less frequently than expected by chance, highlighting significant preferences or aversions in the data. In this study, CFA is used to identify whether speakers in different regions have a statistically significant preference or aversion to specific vulgar forms. This method allows for a more detailed analysis of vulgarity usage patterns that goes beyond raw frequency counts or what Pearson’s χ2-tests would be able to show.

# Findings

The following section presents the findings of the analysis. We start by reviewing the relative frequency (percentage) of vulgar tokens among all tokens across regions represented in the GloWbE data (see Figure 1).

|  |
| --- |
| Figure 1: Percentage of vulgar elements (vulgar tokens) in GloWbE across English-speaking regions. |

Figure 1 shows that vulgarity is most common in the United Sates, Great Britain, and Australia. Overall, the bar plot indicates that vulgarity in English using English expressions is more common in Kachru’s (1988, 1991) inner circle varieties or endonormative, differentiating varieties (Schneider, 2007), compared with outer circle, nativizing, or L2 varieties.

This pattern is reproduced in Figure 2 which offers additional detail by differentiating between discourse type. In most varieties, blog data contains higher frequencies of vulgar elements compared to other online genres and text types (here referred to as general web). The only exemptions are:

* Australia, for which the GloWbE does not contain blog data and where we therefore must rely solely on general web data;
* Singapore and Pakistan, where there is no difference between blog and general web data;
* Nigeria and the Philippines, where we find higher frequencies of vulgar elements in the general web content compared to the blog data.

|  |
| --- |
| Figure 2: Percentage of vulgar elements (vulgar tokens) in GloWbE across English-speaking regions and discourse types. |

Despite the overall relative sparsity of vulgar elements (again, they only amount to 0.023 percent of all tokens), the use of vulgarity is notably widespread: Figure 3 shows that between 6.2 percent of texts in general web data from Tanzania and 18.2 percent of US blog texts contain at least one token that can be considered vulgar. This finding is quite notable as it suggests that while people use vulgarity very sparingly, the use of vulgar elements in language is highly common with speakers being aware of the discursive functions that the use of vulgarity fulfills.

|  |
| --- |
| Figure 3: Percentage of files containing vulgar elements in GloWbE across English-speaking regions and discourse types. |

We now turn to results regarding the variability in the use of vulgar language. This variability can reflect linguistic creativity and it shows that speakers have access to a wide repertoire of forms deemed vulgar for stylistic reasons, for emphasis, or to fulfill other functions in discourse. To ascertain if there are regional differences in creativity associated with variability in using different types of vulgar language, Figure 4 displays the corrected type-token ratio (CTTR) across English-speaking regions and discourse types in the GloWbE data. The lower the CTTR value, the higher the variability and diversity of forms used.

|  |
| --- |
| Figure 4: Corrected Type-Token Ratio of Vulgar Types and Tokens in GloWbE across English-speaking regions and discourse types. |

Figure 4 shows that regions associated with inner circle varieties of English have substantively lower CTTR values compared to outer circle varieties. This strongly suggests that in online discourse speakers of inner circle varieties use a higher diversity of vulgar forms compared to outer circle variety speakers.

To get a better understanding of how this variability is linked to the use of different forms of vulgarity, Figure 5 shows the relative frequency of vulgar forms that have been used at least 10 times per million words across the regions represented by the GloWbE data.

|  |
| --- |
| Figure 5: Relative frequency (per million words) of vulgar lemmas in GloWbE across English-speaking regions (with lemmas having a relative frequency lower than 10 per million being collapsed into a single bin category, other). |

Figure 5 reveals some interesting differences in the patterning of vulgar forms across English-speaking regions. There are three particularly noteworthy results emerging from Figure 5:

* Speakers from Great Britain have the highest rate of low frequency vulgar forms suggesting that, compared to speakers from other regions, they make use of the widest repertoire of vulgar forms in online language.
* Speakers from the United States show by far the highest rates for *fuck*, *damn*, and *shit* with relatively fewer low frequency forms.

Figure 6 reports significant deviances between observed frequencies and the frequencies that would be expected if the lemma was distributed evenly across regions. The measure displayed is the phi-values, an effect size measure. If the phi value is positive, this shows a lemma used more often than would be expected (i.e., the lemma is preferred in that region) while blue indicates that the lemma occurs less often than would be expected (indicating that the lemma is avoided in the lemma in that region).

|  |
| --- |
| Figure 6: Association plot showing preferred and dispreferred vulgar lemmas across regions based on the phi-value (red = preferred, blue = dispreferred, grey = no trend, white = absent). Colouring only shown if the phi-value is greater than 5 or smaller than -5. |

To statistically evaluate the visualisation shown in Figure 5, we now turn to the results of the CFA which was used to determine if speakers from specific regions show significant preferences for certain vulgar forms. The results of the CFA are summarised in Table 2. Only significant associations between vulgar lemmas and region are shown the phi-value, a measure of association strength, is greater than 0.05.

Table 2: Results of the CFA analysing significant preferences (types) for and avoidance (antitypes) of vulgar lemmas across English-speaking regions.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Region | Vulgar lemma | Observed frequency | Expected frequency | Q-value | X2-value | Bonferroni-corrected  p-value | Type | Phi-value |
| United States | fuck | 25419 | 20674.56 | 0.01251 | 1088.77 | < .001\*\*\* | type | 0.052 |
| United States | asshole | 2995 | 1400.66 | 0.004 | 1814.79 | < .001\*\*\* | type | 0.067 |
| Great Britain | arse | 3238 | 1474.81 | 0.00443 | 2107.97 | < .001\*\*\* | type | 0.073 |
| Great Britain | bollocks | 1051 | 357.96 | 0.00173 | 1341.76 | < .001\*\*\* | type | 0.058 |
| Ireland | feck | 448 | 51.63 | 0.00099 | 3042.73 | < .001\*\*\* | type | 0.087 |
| New Zealand | hooker | 416 | 93.95 | 0.00081 | 1103.96 | < .001\*\*\* | type | 0.053 |
| Pakistan | bint | 571 | 15.83 | 0.00139 | 19469.15 | < .001\*\*\* | type | 0.221 |
| Jamaica | jerk | 1395 | 148.83 | 0.00312 | 10434.32 | < .001\*\*\* | type | 0.162 |
| India | jizz | 276 | 10.26 | 0.00066 | 6880.46 | < .001\*\*\* | type | 0.131 |
| Hong Kong | hoe | 1063 | 158.58 | 0.00226 | 5158.26 | < .001\*\*\* | type | 0.114 |
| Singapore | hoe | 1233 | 284.4 | 0.00237 | 3164.08 | < .001\*\*\* | type | 0.089 |
| Ghana | hoe | 575 | 101.04 | 0.00119 | 2223.14 | < .001\*\*\* | type | 0.075 |
| Sri Lanka | bloody | 1485 | 529.89 | 0.00239 | 1721.53 | < .001\*\*\* | type | 0.066 |
| Bangladesh | paki | 96 | 5.32 | 0.00023 | 1544.19 | < .001\*\*\* | type | 0.062 |
| Malaysia | hoe | 829 | 234.79 | 0.00149 | 1503.87 | < .001\*\*\* | type | 0.061 |
| Philippines | nigger | 327 | 52.06 | 0.00069 | 1451.84 | < .001\*\*\* | type | 0.06 |
| United States | arse | 409 | 1950.94 | 0.00388 | 1218.68 | < .001\*\*\* | antitype | -0.055 |
| United States | bloody | 5338 | 11889.8 | 0.01689 | 3610.33 | < .001\*\*\* | antitype | -0.095 |
| Great Britain | ass | 2442 | 4973.71 | 0.00641 | 1288.69 | < .001\*\*\* | antitype | -0.057 |

Table 2 shows that there are significant preferences for (shown as types, upper part of the table) and reservations against (shown as antitypes in the lower part of the table) using certain vulgar lemmas. The rather general vulgar lemmas, *fuck* and *asshole* are significantly associated (preferred) in the United States when choosing to use vulgar language. In contrast, vulgar lemmas associated with British and L2-English, *arse* and *bloody*, are significantly dispreferred by speakers of American English. Indeed, in Great Britain speakers show a significant preference for using variants of the lemmas *arse* and *bollocks* and underuse forms of *ass* compared to the use of these lemmas in other regions. Speakers in Ireland exhibit a preference for variants of the lemma *feck* and New Zealanders overuse variants of *hooker* (though this preference is the likely fall-out of the fact that *hooker* is not only a position in rugby league football, but also a placename in various New Zealand regions; e.g. Hooker Valley and Hooker Glacier). In Hong Kong, Singapore, Ghana, and Malaysia, variants of the lemma *hoe* are more popular in online discourse than in other regions while Jamaican overuse variants of *jerk* and Indians *jizz*. The use of variants of *bint* is significantly associated with Pakistan while forms of *paki* are overused in Bangladesh. Finally, speakers in Sri Lanka overuse *bloody* and speakers in the Philippines overuse forms of *nigger*.

After reviewing the results, we now turn to their interpretation in light of the existing literature and (socio-)linguistic theorising.

# Discussion

The present study set out to explore the use of vulgarity in online language across 20 English-speaking regions. It aimed to improve retrieval of vulgar language items by combining widely used word lists with regular expressions. By combining word lists with regular expressions in this way, the study was able to identify vulgar elements that would not have been detected by either word lists alone or by focusing on selected lemmas associated with vulgar language use. This was showcased by the various variants of the lemma *fuck* (see Section 3).

The most notable and surprising finding the study unearthed relates to the difference between inner circle varieties on the one hand and outer circle varieties on the other: inner circle varieties not only exhibit a higher rate of vulgar language (see Figures 1, 2, and 3) but also higher rates of users of vulgar language, as well as substantively more variability in vulgar language use (see Figures 4, 5, and 6). We interpret this finding to reflect a more general tendency in inner circle varieties for colloquialisation, “a trend towards the informal and colloquial in written communication” (Mair, 1998: 153), which has been shown to be less prevalent in outer circle (see Mair, 2024; Yao & Collins, 2018b). In this context, the present study represents one of the first cross-varietal case-studies that investigate how region-specific social norms affect stylistic as well as socio-cultural norms.

The relationship between vulgarity and informality is demonstrated in the findings for the online blogs. As Figure 2 shows, blog data for most varieties have higher frequencies of vulgar elements compared to other online genres (with the exception of Singapore and Pakistan with no difference, and Nigeria and the Philippines where blog data actually showed lower rates of vulgarity). Given that blogs are used as platforms for personal experiences (to express opinions or offer insights), professional advice (to share expertise), or as tools for marketing or personal branding, their style is conversational and interactional — readers can interact with blog content through comments and social media, and this fosters community engagement. It is significant therefore that the Australian English findings could include no blog data — higher vulgarity rates would be expected were this not the case. The informal character of Australian culture, and its greater willingness to embrace colloquial styles is now well documented. Peters & Burridge (2012) suggest in their study of areal features of Australian and New Zealand English that both varieties go well beyond the kinds of vernacular and informal grammar and lexis noted for varieties elsewhere; Collins and Yao’s (2018a) diachronic account of the role of colloquialization in the endonormativisation of Australian English grammar indicates that the penchant for informality is stronger in Australian English compared to British and American English; more recent support comes from Burke & Burridge (2021).

In the case of outer circle, nativizing or L2 varieties, however, there is a danger that we have under-estimated informality by not being able to include regional markers of casual and colloquial communication. Drawing on Yao & Collins’ (2018b) account of Philippines English, Mair (2024) points out, “there are two ways of ‘doing colloquial’ in these New Englishes: (i) the global way, comparable to the stylistic continua of L1 varieties, and (ii) the local way, by using resources from the contact languages” (p. 202). It is important to note, therefore, that the present study focuses exclusively on the use of vulgar terms *in English*, thus neglecting the use of non-English vocabulary which is more likely to surface and be utilised in these outer circle varieties. A fact supported by the tendency for higher use of code-switching in colloquial discourse in ESL varieties (Montes Alcalá, 2016: 44; Mair, 2024: 202-209). As examples in Carson & Jiang (2021) show, Hokkien expletives in Singaporean English can be used alongside English expressions for added emphasis (the *Coxford Singlish Dictionary* captures the creativity in cross-cultural word play of *Singlish*; Goh & Who, 2002). Clearly, this incorporation of the local is an important aspect of the innovation of outer circle L2 languages, and an increasingly important one if English is in the process of shifting from more formal and polite discourse to become more integral to everyday speech.

*5.1 Creativity and Schneider’s dynamic model of World Englishes*

The inner circle shows a clear split from the outer circle in the degree of variability (a proxy for creativity) as indicated by the lower corrected type-token ratios as well as the relatively higher frequency of *other* items among inner circle varieties in Figure 5. We would like to pose that this finding can be interpreted in light of Schneider’s dynamic model of postcolonial varieties of English (2007). According to this model, non-native varieties of English undergo different developmental stages associated with different features and tendencies for norm-reception versus norm-giving. In this model, varieties of English can be regarded as being on a cline ranging from native, L1-varieties characterised by internal stylistic and socio-linguistic stratification as well as norm exportation to non-native, EFL varieties that exhibit little internal differentiation and norm reception. During dialect maturation, varieties undergo different stages which reflect different milestones along the developmental pathway. The fact that our findings show outer circle varieties in general exhibiting less creativity around swearing (i.e. lower rates of variability) is likely to reflect their position in the early stages of Schneider’s model — stages that are more prone to following external norms regarding appropriate language use and thus lower rates of linguistic variability and creativity.

Jamaica is interesting in this regard. We expect that its high levels of vulgarity might be associated with creativity (though this is not reflected in the data here). Jamaican Creole (Patois) is the primary spoken language for most people from childhood and remains the dominant language in informal, everyday communication for the vast majority of Jamaicans; Jamaican English, on the other hand is used in formal contexts (e.g. education, government, media). It is a speech community whose complexity, and potential for vulgarity, could not be captured here. Farquharson et al.’s (2020) discussion of Jamaican swearing practices points to the very sizeable swear word lexicon that includes words and expressions derived from English; e.g. *dyam* (< English *damn*), *fok* (< English *fuck*), *kis mi aas* (< *English kiss my arse*). He also describes another set of swear words that carry more local and cultural weight — “evaluated by native speakers of Jamaican Creole to be more authentically Jamaican than those already mentioned” (p. 151). These are compounds formed by concatenating a body part with *klaat* (< English *cloth*); for example, *bomboklaat* (< *bombo* ‘vagina, vulva’, originally a sanitary towel used by menstruating women), which was then the inspiration for other expressions such as *blodklaat*, *pusiklaat* (< English *pussy*, *raasklaat* (*raas* < English *arse*). All show a number of different spellings (e.g. *bumboclaat, bumboclot, bumbaklaat, bumbaklot, bomboclat*). The word *-klaat* works as a kind of template and has almost suffix-like status. This is a typical aspect of the creativity around successful slang and vulgar expressions — sequences of words or linguistic routines provide a kind of life support framework to maintain the vitality of an idiom, sometimes over centuries, as shown in the *-arse* and *-head* compounds described earlier (see also Burridge in press on the longevity of slang terms).

In terms of Schneider’s model, Indian English and Singapore provide an interesting contrast — both have passed through the earlier stages (foundation, exonormative stabilization, nativization) and now enjoy broad acceptance and stability within the respective countries. Both show very different rates of vulgarity, however — rates for India are relatively low, while those for Singapore are startlingly high (it follows Australia in fourth position, and is ahead of New Zealand, Ireland and Canada).

In India, English is becoming a primary language in certain urban areas of India, but it is unlikely to become the dominant L1 across the country in the near future for a number of reasons (e.g. its large and widely dispersed population, its linguistic diversity and cultural attachment to regional languages such as Hindi, Tamil, Bengali). Singapore, on the other hand, is in the process of transitioning from ESL to ENL — increasingly English is spoken as a first language by younger generations; it is the main medium of instruction in schools and is widely spoken in households and promoted by language policies.

The higher rates of vulgarity in Singapore is the predicated fall-out of the expansion of Singapore English into informal contexts, where swearwords, colloquial expressions are freely used — the transition to L1 status means that English is no longer confined to official or formal domains but is fully integrated into everyday life. Increasingly Singaporeans view English, not just as a learned second language, but as a language of identity, belonging and self-expression, and jokey publications such as the *Coxford Singlish Dictionary* illustrate how speakers are making English their own by creatively blending it with influences from local languages like Malay, Tamil, and Chinese dialects.

The Jamaican and Singlish examples illustrate that a creative hallmark of outer circle Englishes is their use of innovative compounding, merging English with local languages or slang. This incorporation of the local is an important aspect of the innovation of newer varieties of English and is something that remains to be investigated.

*5.3 Schneider’s manifestations of cultures in corpus texts*

Schneider (2018, 2021) investigates whether cultural differences can be systematically reflected in language forms within modern electronic corpora, and whether corpus-linguistic methodology can reveal the extent and nature of cultural influences present in these corpora. He postulates three types of “nexus” between language and culture, the second of which, cultural dimensions, is particularly relevant here.

Drawing on Geert Hofstede’s model of cultural dimensions (2001 [1980]), this nexus focuses on the attitudinal and behaviouralparameters on which cultures around the world can vary. Schneider associates these sociological dimensions with characteristic “indicator terms”, words or phrases, whose relative frequency provides a measure of the strength and salience of each dimension in a given variety (e.g. the use of the first person plural as an indicator of collectivism versus individualism, or address terms such as *Sir/Madam* that mark high (or low) status in verbal encounters). It seems to us that patterns of vulgarity would also be among the words and phrases that potentially project the attitudes and interactive behaviours of a speech community — in particular, individualism (vs collectivism) and power distance (solidarity, relationship and politeness). Table 3 provides the scores for both cultural dimensions pertaining to the countries examined here[[4]](#footnote-4).

Table 3: Hofstede scores for the cultural dimensions of individualism and power distance.

|  |  |  |
| --- | --- | --- |
| Country | Individualism Index | Power distance index |
| 1. Australia (au) | 90 | 36 |
| 2. Bangladesh (bd) | 20 | 80 |
| 3. Canada (ca) | 80 | 39 |
| 4. Ghana (gh) | 20 | 77 |
| 5. Great Britain (gb) | 89 | 35 |
| 6. Hong Kong (hk) | 25 | 68 |
| 7. India (in) | 48 | 77 |
| 8. Ireland (ie) | 70 | 28 |
| 9. Jamaica (jm) | 39 | 45 |
| 10. Kenya (ke) | 27 | 64 |
| 11. Malaysia (my) | 26 | 104 |
| 12. New Zealand (nz) | 79 | 22 |
| 13. Nigeria (ng) | 20 | 77 |
| 14. Pakistan (pk) | 14 | 55 |
| 15. Philippines (ph) | 32 | 94 |
| 16. Singapore (sg) | 20 | 74 |
| 17. South Africa (za) | 65 | 49 |
| 18. Sri Lanka (lk) | 35 | 80 |
| 19. Tanzania (tz) | 27 | 64 |
| 20. United States (us) | 91 | 40 |

High individualism scores reflect a culture where interpersonal ties are loose, and individual rights, personal achievement and self-reliance are highly valued; low individual scores reflect societies that tend to prioritise the group’s well-being over their individual interests. The power distance index is defined as the extent to which less powerful members of organizations and institutions (such as the family) accept inequality in the exercise of authority, as well as in wealth, status and privilege (Hofstede, 2001: xix) — high scorers in power distance endorse obedience and are likely to accept inequality as part of the natural or societal structure, and low scorers endorse egalitarianism and informality. This dimension tends to decrease as individualism increases.

Cultures high in individualism lean towards personal freedom, autonomy and equality of opportunities, and informality is promoted in social and workplace interactions. In highly individualistic Australia, for example, displays of respect are rejected (unless it is for the ordinary Australian) and achievement and ambition downplayed; the reciprocal use of first names is usual, (regardless of social distance and difference in relative power). In contrast, cultures with high collectivism and high power distance are often more formal and emphasize hierarchy, where authority and social structures are respected more strictly.

Inner circle countries represent a western culture and a shared individualistic orientation (with US, Great Britain and Australia showing the highest individualism scores: 91, 89 and 90, respectively). This contrasts with the collectivist community orientation of Asian and African cultures, whose low individualism scores reflect the values placed on group loyalty and extended familial or community networks (e.g. Pakistan, Singapore, Nigeria with scores of 14, 20 and 20, respectively). It is not difficult to see how high vulgarity rates might correlate with cultures that value individualism, egalitarianism and informality — for cultures that favour personal autonomy and freedom of expression, and even the rejection of authority and social norms, vulgarity is made to measure. There are two notable exceptions, however, where vulgarity rates do not correlate with the extent to which the culture leans more (or less) towards individualism and power distance — Singapore and Jamaica.

Singapore (individualism score 20; power distance 74) represents a Chinese-dominated, Buddhist-influenced cultures, but one which has undergone a considerable degree of modernisation and westernisation, and where the majority of children are growing up speaking English as an L1 (Schneider 2021 describes its “dual Asian-cum-Western cultural orientation”, p. 41). Certainly, there is plenty of evidence that young modern Singaporeans hold personal beliefs that challenge restrictions on self-expression. In their study of youth swearing in Singapore, Wen et al. (2021) describe how their participants (aged between 18 and 25) engage in social swearing as “an act of rebellion against Singapore’s collectivistic expectation to inhibit personal expressions” (p. 145); for example, one female observed: “Being able to curse, swear and say anything that I like, it’s like a breath of fresh air – a giant middlefinger to the repressive social code that dictates what can or cannot be said. It feels good”. These authors present a number of responses with similar sentiments and also many examples where participants use vulgar language as a means of challenging or subverting established norms. This dismissal of conventional authority is undoubtedly heightened by the Singapore government’s strict stance against swearing and offensive language in public areas (parks, streets, shopping centres, government buildings and local media)[[5]](#footnote-5).

As already discussed, our study also shows Jamaica as having high rates of vulgar language, a finding supported by the fact that Jamaican swear words have become popular beyond the borders of the Caribbean island, now even used by non-Jamaicans (see Farquharson et al. 2020). In terms of Hofstede’s cultural dimension values, Jamaica (individualism score 39; power distance 45) combines a lowish individualism score and a moderate score for power distance; this suggests a society that values both communal ties and personal expression, allowing individuals to maintain a sense of identity while still being part of a larger community. These scores highlight the complexities of Jamaican culture, linguistically complicated by the existence of a creole continuum spanning from Jamaican Creole (a strong marker of cultural identity) to Jamaican English, with many varieties in between — most Jamaicans navigate this continuum, shifting their language depending on context.

Of course, countries are by no means uniform, and Hofstede’s cultural dimensions have attracted criticism (e.g. scores failing to reflect regional, ethnic or social variation and to account for cultural change). Nonetheless, they provide a useful foundational framework for understanding cultural differences, and how these differences might manifest themselves in the language forms (in our case, vulgar forms) in modern corpora.

*5.3 Use versus perception of vulgarity*

As described, the patterning of vulgar language emerging from the present study shows marked differences between inner and outer circle varieties, with inner circle varieties exhibiting higher frequencies, more common use of vulgarity, and more variability in the expression of vulgarity compared with outer circle varieties. However, these results contradict the self-perceptions of two nations in relation to swearing.

Australians might well be disheartened when they discover that they are not the top users of profanity among English-speaking countries. Their deep national attachment to the vernacular dates back to the original mix of slang, dialect and underworld jargon that gave rise to Australian English — fuelled by anti-authoritarian sentiment, the colloquial part of the language expanded to become the feature that best distinguished the established citizen (or *old chum*) from the stranger (or *new chum*). Australia’s love of swear words in particular is very evident in the public life of these words.

Earlier we mentioned Tourism Australia’s global campaign with the tagline, “So where the bloody hell are you?”. The ad showcased Australians and iconic locations, but its casual use of the phrase “bloody hell” led to controversy. The ad was banned in Canada[[6]](#footnote-6), censored in Singapore (the phrase “bloody hell” was removed from the advertisement), and banned in the UK from television and billboards. The UK ban was surprising and certainly inconsistent with the British use of *bloody* and swear words generally in popular culture (born out by our findings here). It was also not in keeping with general attitudes towards offensive language; Millwood-Hargrave (2000), for example, researched multiple UK media bodies ad found that only 3% of respondents considered *bloody* offensive. Corpus evidence (Allan & Burridge 2019) also suggested that Australians lag well behind Londoners in actual use of *bloody* — in the Corpus of London Teenage Speech there are 291 instances of bloody in 107,429 sentences (27 per 10,000); in the conversational data of the ICE-AUS it occurs 46 times in 22,933 sentences (20 per 10,000). British English speakers use *bloody* freely, but they are not used to seeing it on public display.

This very public airing of swear words it is something that is often remarked upon by overseas visitors to Australia; for example, the very visible TV and billboard slogan ‘If You Drink, Then Drive, You’re a Bloody Idiot’ is part of a long-running and highly successful public safety campaign (dating from the 1980s). It is aimed to resonate with Australian general public but for outsiders, the use of *bloody* in a government-backed campaign comes across as surprising or even unprofessional — colloquial language, especially swear words, would not be used in official messaging in other countries.

As already flagged by Goddard (2015), the showcasing of swearwords is what potentially distinguishes Australian swearing from American English. And yet, our findings show the US as having the highest rates for vulgar language usage in GlobWE (we argued earlier this was the fall-out of the strong individualistic culture in the US). Their rejection of public displays of swearing, however, is in keeping with the broader societal expectation of civility and restraint in the US, particularly in public spaces — the religious influence would also support a cultural aversion to swearing as a breach of moral decency (recall that the US recorded the highest use of the only religious-based swearword we included *damn*). Future research that takes in context and prevailing attitudes might well reveal that profanity in the public sphere is indeed much more accepted in Australia than in the USA — Australians might still live up to their popular image of having an unusually rich and creative “bad” language, and Americans can retain their conservative attitude toward public morality while at the same time as being the most potty-mouthed of the inner circle nations.

# Outlook

Ideally, research into vulgarity across the Englishes would interrogate findings from a functional perspective, taking into account contextual factors such as the relative perceived strength, pragmatic function and conversational context of the vulgar words. Word embeddings (a way of placing words with similar meanings close together in multi-dimensional spaces) offers a more nuanced way of representing and analysing word meanings based on their contexts. These would capture the semantic relationships between words based on how they are used and would also allow to differentiate between vulgar uses and non-vulgar uses (e.g. ball in the context of sports versus used in a sexual context) as well as to acknowledge that words carry different connotations within and across cultural spaces. This would help overcome the limitation of our study is that words are not understood within their respective cultural, social and historical contexts — while we examine vulgarity in non-Western varieties of English, the perspective remains a Western one. Further investigation, for example, would reveal whether region-specific preferences shape specific forms of vulgarity; for example, findings already reveal preferences among inner circle dialectal markers, such as the predilection for *arse* and *bollocks* in Great Britain and for *feck* in Ireland. These items have surfaced as being typically British and Irish in variation-focused studies before (Kirk, 2023). Similarly, the avoidance of regional items linked to other varieties (e.g. the avoidance of British features such as *arse* and *bloody* by American English speakers or the avoidance of *ass* as an American English feature by British English speakers; see Table 2). The preference for sex work/promiscuity linked forms such as *hoe* in several outer circle varieties, however, suggest there true cultural reflexes might be confirmed by the statistical analysis of preferred semantic fields (though the complexity of the semantics of vulgar words, especially ,across cultures makes fine-grained categorisation difficult; e.g. *bitch* can be a general term of abuse for women (and occasionally men) and unpleasant situations or things; a term of praise for something or someone wonderful; a word for sexual activity; to name a few). Interestingly, racial slurs only surface as a preferred form of swearing and vulgarity among Indian speakers and when referring to their geographical neighbour (Pakistan).

**Author Declaration: Use of AI-assisted Technologies**

In the preparation of this manuscript, the authors used ChatGPT, an AI-assisted tool developed by OpenAI, to improve the style and readability of the text. The content and interpretations of the work are solely my own, and the AI was used exclusively for enhancing clarity and coherence. No AI technology was used in the development of the research methodology, data analysis, or conclusions. This statement is provided in compliance with the publisher’s guidelines on transparency in the use of AI-assisted technologies.

**References**

Adams, M. (2016). *In praise of profanity.* Oxford University Press.

Allan, K. (2015a). A Benchmark for Politeness, in A. Capone & J. L. Mey (eds) *Interdisciplinary Studies in Pragmatics, Culture and Society* (Vol. 4 of the series Perspectives in Pragmatics, Philosophy & Psychology); pp. 397-420.

Allan, K., (2015b). When is a slur not a slur? The use of nigger in “Pulp Fiction”. *Language Science 52,* 187–199.

Allan, K., (2018). Taboo words and language: An overview. In: Allan, K. (Ed.), *The Oxford Handbook of Taboo Words and Language*. Oxford University Press, Oxford, pp. 1–27.

Allan, K. & Burridge, K. (1991). *Euphemism and Dysphemism: language used as shield and weapon.* Oxford University Press.

Allan, K. & Burridge, K. (2006). *Forbidden Words: Taboo and the censoring of language.* Cambridge University Press.

Allan, K. & Burridge, K. 2010. Swearing and taboo language in Australian English. In Peters, P., Collins, P. & Smith, A. (Eds.), *Comparative Grammatical Studies in Australian and New Zealand English*. John Benjamins, pp. 295-347.

Allen, W. F. (1987). *Teenage speech: The social dialects of Melbourne teenagers*. B.A. Honours Thesis. Linguistics Department, La Trobe University, Melbourne.

Bednarek, M., 2015. “Wicked” women in contemporary pop culture: “Bad” language and gender in Weeds, Nurse Jackie, and Saving Grace. *Text & Talk 35* (4), 431–451.

Beers Fägersten, K., 2012. *Who’s Swearing Now? The social aspects of conversational swearing*. Cambridge Scholars Publishers.

Beers Fägersten, K., Stapleton, K., 2017. Introduction: Swearing research as variations on a theme. In: Beers Fägersten, K., Stapleton, K. (Eds.), *Advances in Swearing Research: New languages and new contexts*. John Benjamins, pp. 1–15.

Bergen, B., 2016. *What the f: What swearing reveals about our language, our brains, and ourselves*. Basic Books, New York.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik*. Springer-Verlag.

Burridge, K. (forthcoming). “Survival of the fittest” — the evolution of slanguage. *Australian Journal of Linguistics.*

Burridge, K. & Benczes, R. (2019). Taboo as a driver of language change. In Allan, K. (Ed.), *The Oxford Handbook of Taboo Words and Language*. Oxford: Oxford University Press, pp. 180-99.

Burke, I. & Burridge, K. (2021). Privileging informality: cultural influences on the structural patterning of Australian English.In: Peters, P. & Burridge, K. (Eds.), *Exploring the Ecology of World Englishes in the Twenty-First Century: Language, Society and Culture*. Edinburgh University Press, pp. 15-45.

Carson, L., Jiang, N. (2021). Offensive Words in Chinese Dialects. In: An Anatomy of Chinese Offensive Words. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-63475-9\_5

Collins, P & Yao, X (2018). Colloquialisation and the evolution of Australian English: A cross-varietal and cross-generic study of Australian, British, and American English from 1931 to 2006, *English World-Wide 39*(3), 253-277.

Crystal, D. 1988. *Language Play.* Penguin Books.

Culpeper, J., 2011. *Impoliteness: Using language to cause offence*. Cambridge University Press.

Culpeper, J., 2018. Taboo language and impoliteness. In: Allan, K. (Ed.), *The Oxford Handbook of Taboo Words and Language*. Oxford University Press, Oxford, pp. 28–40.

Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, *36*(1), 1-28.

Dewaele, J.-M., 2004. The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of Multilingual and Multicultural Development 25* (2/3), 204–219.

Dewaele, J.-M., (2015). British ‘bollocks’ versus American ‘jerk’: Do native British English speakers swear more—or differently—compared to American English speakers? *Applied Linguistics Review 6,* 309–340.

Emerick, J. (2024). *List of dirty, naughty, obscene, and otherwise bad words* [GitHub repository]. GitHub. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Farquharson, J. T, Forrester, C. & Hollington, A. (2020). The linguistics of Jamaican swearing: forms, background and adaptations. In Nassenstein, N. & Storch, A. (Eds.). Swearing and cursing: Contexts and practices in critical linguistic perspective. [Language and Social Life 22] Berlin: De Gruyter Mouton.

Goh, C., & Woh, Y.-Y. (2002.) Coxford Singlish Dictionary. Singapore:Angsana. <http://colingoh.com/project/the-coxford-singlish-dictionary/>. Accessed 21st October 2024.

Goddard, C. (2015). “Swear words” and “curse words” in Australian (and American) English. At the crossroads of pragmatics, semantics and sociolinguistics. *Intercultural Pragmatics 12*(2), 189–218

Holmes, J., Vine, B. & Johnson, G. (1998). The Wellington Corpus of Spoken New Zealand English: A Users' Guide. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Hughes, G. (2015). *An Encyclopedia of Swearing. The Social History of Oaths, Profanity, Foul Language, and Ethnic Slurs in the English-speaking World*. Routledge.

Jay, T. (2000). *Why we curse: A neuro-psycho-social theory of speech*. Benjamins.

Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives on Psychological Science 4* (3), 153–161.

Jay, T. Janschewitz, K., (2008). The pragmatics of swearing. *Journal of Politeness Research 4*, 267–288.

Johnson, D.I., Lewis, N., (2010). Perceptions of swearing in the work setting: An expectancy violations theory perspective. *Communication Reports 23*, 106–118.

Kachru, B. B. (1988). The sacred cows of English. *English Today*, 4(4), 3-8.

Kachru, B. B. (1990). World Englishes and applied linguistics. *World Englishes*, *9*(1), 3-20.

Kassem, H. (2023). *FUCK THIS SHIT: Swearing in different varieties* (University of Oslo Master's thesis). <https://www.duo.uio.no/handle/10852/106572> (accessed 19th October 2024).

Kirk, J. (2023). The Vocabulary of Irish English. In R. Hickey (Ed.), *The Oxford Handbook of Irish English* (pp. 179–203). Oxford University Press.

Lienert, G. A., & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. *Educational and Psychological Measurement*, *35*(2), 231-238.

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

Love, R. (2021). Swearing in informal spoken English: 1990s–2010s. *Text & Talk*, *41*(5-6), 739-762.

Mair, C. (1998). Corpora and the study of the major varieties of English: Issues and results. In H. Lindquist, S. Glintford, M. Levin, & M. Estling (Eds.), *The major varieties of English: Papers from MAVEN 97, Växjö, 20–22 November 1997* (pp. 139–157). Växjö University.

Mair, C. (2024). Colloquialisation: Twenty-five years on. In Hiltunen, T., Vartiainen, T. & Räikkönen, J. (Eds.) *Corpus-Pragmatic Studies of Democratization in Public Discourses: New perspectives, methods and materials* (Special issue of the *Journal of Historical Pragmatics*, *25*(2)), 193-214.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development* (pp. 16-30). Palgrave Macmillan UK.

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

McEnery, A., Baker, J.P. & Hardie, A. (1999) Assessing claims about language use with corpus data – swearing and abuse. In Kirk, J.M. (ed.) *Corpora Galore: Papers from ICAME 1998*. Amsterdam: Rodopi, 45-55. <https://www.lancaster.ac.uk/staff/hardiea/pdf/mceneryetal-CorporaGalore.pdf> (accessed 9th September 2024).

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

McEnery, T., Brookes, G., Hanks, E., Gerigk, K., & Egbert, J. (2023). Swearing, discourse and function in conversational British English. *Journal of Pragmatics*, *213*, 36-48.

Mejia, G. E. A. ., & Ngo, C. G. A. . (2024). Profanity in Social Media: An Analysis of Pragmatic Functions and Politeness Maxims Violation. *Journal Corner of Education, Linguistics, and Literature*, *4*(1), 92–117.

Millwood-Hargrave, A. (2000). *Delete Expletives?* [research undertaken jointly by the Advertising Standards Authority, British Broadcasting Corporation, Broadcasting Standards Commission and the Independent Television Commission] Broadcasting Standards Commission]. Broadcasting Standards Commission, 2000.

Mohr, M. (2013). *Holy sh\* t: A brief history of swearing*. Oxford University Press.

Montes-Alcalá, C. (2016). iSwitch: Spanish-English mixing in computer-mediated communication. *Journal of Language Contact, 9*(1), 23–48.

Morris, C. (2022). Compound pejoratives on Reddit – from *buttface* to *wankpuffin.* <https://colinmorris.github.io/blog/compound-curse-words>. Accessed 20th October 2024.

Musgrave, S. & Burridge, K. (2014). Bastards and buggers – Historical snapshots of Australian English swearing patterns. In K. Burridge & R. Benczes (Eds.) *Wrestling with words and meanings: Essays in honour of Keith Allan.* Monash University Publishing, pp. 3-32.

Peters, P. & Burridge, K. (2012). Areal linguistics in the South Pacific. In Hickey, R. (Ed.) *Areal Features of the Anglophone World*. Berlin: Mouton de Gruyter, pp.233-260.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Posit team (2024). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA. <http://www.posit.co/>.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Ross, H. (1960). Patterns of swearing. *Discovery, 21*, 479-81.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the world*. Cambridge University Press.

Schneider, E.W. (2018). Reflections of culture in corpus texts. *ICAME Journal, 42.* DOI: 10.1515/icame-2018-0006.

Schneider, E.W. (2021). Platform Paper: Reflections of Cultures in Corpus Texts: Focus on the Indo-Pacific Region.In Peters, P. & Burridge, K. (Eds.), *Exploring the Ecology of World Englishes in the Twenty-First Century: Language, Society and Culture*. Edinburgh University Press, pp. 15-45.

Schweinberger, M. (2018). Swearing in Irish English–A corpus-based quantitative analysis of the sociolinguistics of swearing. *Lingua, 209*, 1-20.

Smith, A. (Owner), & Peters, P. (Contributor). (2023, December 21). International Corpus of English (ICE) (Version 1). Macquarie University. https://doi.org/10.25949/24769173.v1

Stapleton, K. (2010). Swearing. In M. A. Locher & S. L. Graham (Eds.), *Interpersonal pragmatics* (Vol. 6, pp. 289-306). De Gruyter Mouton.

Stapleton, K., Fägersten, K. B., Stephens, R., & Loveday, C. (2022). The power of swearing: What we know and what we don’t. *Lingua*, *277*, 103406.

Stenström, A.-B., 2017. Swearing in English and Spanish teenager talk. In: Beers Fägersten, K., Stapleton, K. (Eds.), Advances in Swearing Research: New languages and new contexts. John Benjamins, Amsterdam, pp. 157–181.

Stephens, R., Atkins, J., Kingston, A., 2009. Swearing as a response to pain. *Neurological Report 20* (12), 1056–1060.

Stephens R., Spierer, D. K., & Katehis, E. (2018). Effect of swearing on strength and power performance. *Psychology of Sport and Exercise,* 35, 111–117.

Stephens, R., Robertson, O., 2020. Swearing as a response to pain: Assessing hypoalgesic effects of novel “swear” words. *Frontiers in Psychology 11*, 723. <https://doi.org/10.3389/fpsyg.2020.00723>.

Troutman, D., 2006. “They say it's a man's world, but you can't prove that by me”: African-American comediennes' construction of voice in the public space. In: Baxter, J. (Ed.), *Speaking Out: The female voice in public contexts*. Palgrave Macmillan, Basingstoke, pp. 217–239.

*List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words.* url <https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en>

Wen, S. T., Binte Hassan, N. M., De Jun, L., & Kasara, N. A. (2021).  The Influence of Social Swearing on In-group Dynamics of Singaporean *Youth Pioneer Road 1*, 141-148.

Yao, X., & Collins, P. (2018). Exploring Grammatical Colloquialisation in Non-native

English: A Case Study of Philippine English. *English Language and Linguistics 22*(3), 457–482.

1. See Kachru (1988: 5). [↑](#footnote-ref-1)
2. In contrast, the online version of GloWbE comprises approximately 1.9 billion tokens (see Davies & Fuchs, 2015: 6). [↑](#footnote-ref-2)
3. The only exception is Australia for which only general web data, but no blog data is available. [↑](#footnote-ref-3)
4. For a useful summary, see <http://clearlycultural.com/geert-hofstede-cultural-dimensions/individualism/>; [accessed 22 October 2024]. [↑](#footnote-ref-4)
5. See *Understanding the Legality of Public Swearing in Singapore*; <https://reyabogado.com/us/is-it-illegal-to-swear-in-public-in-singapore/> [accessed 22 October 2024]. [↑](#footnote-ref-5)
6. The conflation of *bloody* and *hell* appears to have been what upset North Americans, not surprisingly given the greater importance they attach to the Christian religion; the concern was undoubtedly intensified by the popular etymology that derives the expression either from *By our lady,* an oath calling on the assistance of the Virgin Mary, or from *[God’s] blood*. There is no evidence for either of these histories; what is more, *bloody* is not an expletive like these two expressions, but rather an epithet. Yet, for some people the imagined blasphemous and profane implications have been enough to condemn the word (which in the 1700s was little more than an intensifier akin to *frightfully* orvastly; Allan & Burridge 2019). [↑](#footnote-ref-6)